Statistics for Data Science

Henri Elad Altman

Freie Universität Berlin

January 30, 2023

Table of contents

Introduction

Statistics have existed for several centuries. Originally designating methods to organise and interprete data, modern statistics have evolved to become a very wide and thriving branch of mathematics, with plentiful applications. The goals of statistics nowadays vary: make predictions, provide a classification, derive an estimation, etc, are all tackled by statistical methods. The Leitmotiv behind these different problems is usually the same: assume that there is some process generating data; given the observed data, what can we say about the process that generated these? How can we control the uncertainty in our results?

1 Probability theory

Probability is a mathematical language to model our problems, but it also offers a rich toolbox. Several theorems of probability (the Law of Large Numbers, the Central Limit Theorem, Hoeffding's inequality, ...) play a crucial role in statistics.

There are several ways to interpret the notion of probability.

2 Frequentist inference

In frequentist inference we think of a probablity as an approximate empirical mean observed when running some random experiment a large number N of times. Assume that we are measuring a random quantity X, and let x_i, $1 \leq i \leq N$ be the observed results. Then the probability $\mathbb{P}(X \in E)$ of an outcome E for this experiment is approximately the value, when N is very large, of the ratio of the number of experiments with outcome E with the total number N of experiments. Using probabilistic notations,

$$\mathbb{P}(X \in E) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1} 1_E(x_i)$$

The above equality is justified by the Law of Large Numbers (LLN), one of the cornerstones of the theory of probability.

3 Bayesian inference

In some cases the frequentist intrepretation of a probability is not meaningful. One example is in weather forecast: for instance the probability of the event "it will rain tomorrow" clearly cannot be thought of as the limit of the empirical mean of some experiment repeated several times. An alternative, equally valid interpretation is in terms of degree of belief: the higher the probability of an event, the more likely this event is to happen. This interpretation lies at the core of the Bayesian approach.

4 A brief history of statistics

As mentioned above, modern statistics realies on probability theory. Early on mathematicians used probabilistic concepts to make predictions on random events. In the 17th Century, gambing games attracted the interest of prominent mathematicians such as Blaise Pascal (1623-1662). The development of random contracts in Europe further promoted the development of probabilistic techniques. But only in 1933 did modern probability appear, with the publication of *Foundations of the Theory of Probability*, by the Soviet mathematician Andrey Kolmogorv (1903-1987). This article laid the cornerstones of modern probability which, from then on, became recognised as a rigorous branch of mathematics.

Statistics, a few landmarks: Bayes. Fisher. William Sealy Gosset (Student test).

5 How to use these lecture notes?

These lecture notes contain the core lecture material. They also contain supplementary material (remarks, examples, exercises) that you can use to check your familiarity with the course content and further your understanding. Remarks indicated with a * are not (always) intended to be discussed in class: they are indications for students wishing to explore more technical details than can be addressed in the framework of this course. Students who wish to consolidate their mathematical background may use the notes below:

https://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/computational_cogni_neurosc/teaching/Statistics_for_Data_Science_19_201/Mathematical_preliminaries.pdf

Part I

Probability basics

Chapter 1
Probability spaces

To expound the theory of statistics, it is essential to provide a robust, unequivocal mathematical framework on which our methods will be based. This framework is provided by Kolmogorv's axioms, which we present in this chapter, and which provides a mathematical definition for the notion of probability of an event in a given random experiment. Random experiments can be as diverse as throwing a dice, flipping a coin 3 times, or tomorrow's weather in Berlin. Events can be thought of as situations encountered in these experiments, such as "the dice returns 6", "the coin returns Heads 3 times" or "it will rain tomorrow in Dahlem". To any such event we wish to ascribe a probability, which will be a number between 0 and 1 satisfying certain rules.

In the following, for any non-empty set Ω, we denote by $\mathcal{P}(\Omega)$ the collection of all subsets of Ω (called the *power set* of Ω):

$$\mathcal{P}(\Omega) = \{A \colon A \subset \Omega\}.$$

For any finite set A we shall denote by $|A|$ the cardinality of A, i.e. the number of elements contained in A.

1.1 Sample space and events

Assume we perform a random experiment, e.g. throwing a dice. In order to model this experiment mathematically, we first choose a **sample space** Ω: this is a non-empty set encoding all possible outcomes of the experiment. E.g., when throwing a dice, a natural choice is $\Omega = \{1, \dots, 6\}$. Events are then defined as follows:

Definition 1.1. *An event is a subset of the sample space* Ω.

Practically speaking, and event represents a sub-collection of outcomes of the experiment we are interested in.

> **Example 1.2.** We throw a die. To model this experiment we choose $\Omega = \{1, \dots, 6\}$as our sample space. An event is any subset A of $\{1, \dots, 6\}$: for instance, $A = \{1, 3, 5\}$ represents the event that the result of the throw is an odd number.

Given two events A and B, we may consider their union $A \cup B$, which represents the event that either A *or* B (or both) occur. Likewise the intersection $A \cap B$ (a.k.a. joint occurrence of A and B) represents the event that both A *and* B occur simultaneously. If $A \cap B = \emptyset$ we say that A and B are incompatible.

Example 1.3. We throw a coin three times. To model this experiment, we consider $\Omega = \{H, T\}^3$, i.e. the set of all vectors with 3 entries, with each entry taking value H or T. Here H stands for "Heads" and T for "Tails" (of course, choosing e.g. $\Omega = \{0, 1\}^3$, with 0 and 1 representing Heads and Tails respectively, would be eqaully valid). An event is any subset of $\{H, T\}^3$. For instance we may consider the events

$$A = \{(H, H, T), (H, T, H), (T, H, H)\} \quad \text{"get Heads exactly twice"}$$

$$B = \{(H, H, H), (T, T, T)\} \quad \text{"get 3 times the same result"}.$$

Note that $A \cap B = \emptyset$, so A and B are incompatible. Consider now the event

$$C = \{\omega \in \Omega : \exists i = 1, 2, 3, \omega_i = H\} \quad \text{"get Heads at least once"}.$$

Then the joint occurrence of B and C is

$$B \cap C = \{(H, H, H)\} \quad \text{"get Heads 3 times"}$$

1.2 * A little detour through σ − algebras

Actually, the real definition of an event in the most general setting is slightly more complicated than in Definition 1.1, and requires the notion of σ-algebra. However, for simplicity, we skip this notion in class and, in this course, will stick to Definition 1.1.

Definition 1.4. * *Let Ω be a non-empty set. A collection $\mathcal{A}$ of subsets of Ω is called a sigma-algebra if the following properties hold:*

- $\emptyset \in \mathcal{A}$ *and* $\Omega \in \mathcal{A}$,
- *for all $A \in \mathcal{A}$, we also have $A^c \in \mathcal{A}$,*
- *if we have a sequence $(A_n)_{n \geq 1}$ with $A_n \in \mathcal{A}$ for all $n \geq 1$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.*

Elements of $\mathcal{A}$ are called ***measurable sets*** *or* ***events****, and $(\Omega, \mathcal{A})$ is said to be a* ***measurable space****.*

Remark 1.5. * Let $\mathcal{A}$ be a σ-algebra over Ω. Then it follows from the definition that:

- If $A_1, \ldots, A_n$ is a finite collection of elements of $\mathcal{A}$, then $\cup_{i=1}^{n} A_i \in \mathcal{A}$
- by de Morgan's identities, if $A_n \in \mathcal{A}$ for all $n \geq 1$, then $\cap_{n=1}^{\infty} A_n \in \mathcal{F}$.

- If $A, B \in \mathcal{A}$ then so does $A \setminus B$.

Let us give a few examples of σ-algebras. We start with a very general couple of examples.

Example 1.6. *: Let Ω be a non-empty set.

1. The smallest σ-algebra over $\mathcal{A}$ is $\{\emptyset, \mathcal{A}\}$, the *coarse* σ-algebra
2. The largest σ-algebra over $\mathcal{A}$ is $\mathcal{P}(\mathcal{A})$, the *discrete* σ-algebra

The second case above is used very often in the case where $\mathcal{A}$ is finite (or countably infinite): in such a case we will most often work with the discrete σ-algbera.

What if we want to work with an **uncountable** set, e.g. the set $\mathbb{R}$ of real numbers? In that case, the choice of a σ-algbera is more delicate.

Definition 1.7. ** Let $n \geq 1$. The Borel σ-algbera on $\mathbb{R}^n$ is the σ-algebra generated by open subsets of $\mathbb{R}^n$. It is denoted by $\mathcal{B}(\mathbb{R}^n)$. An element of $\mathcal{B}(\mathbb{R}^n)$ is called a Borel (measurable) subset.*

We refer to Chapter 3 of [4] for more details. Practically, $\mathcal{B}(\mathbb{R}^n)$ contains all subsets of $\mathbb{R}^n$ one usually works with: intervals, open/closed sets, singletons, are all Borel measurable. For the applications we have in mind here, Borel measurability issues will never happen.

1.3 Probability measure

Given a random experiment, and a sample space Ω encoding all possible outcomes, we wish to assign to each event of Ω a number known as its *probability*. Let $\mathcal{A}$ denote the collection of all events of Ω.

Definition 1.8. *A probability measure $\mathbb{P}$ on a set Ω is a map $\mathbb{P}: \mathcal{A} \to [0, 1]$ with the following properties:*

- $\mathbb{P}(\Omega) = 1$.
- *(σ-additivity) If $\{A_n\}_{n>0}$ is a countable collection of events that are **pairwise disjoint**, i.e. $A_n \cap A_m = \emptyset$ for all $n \neq m$, then one has*

$$\mathbb{P}(\bigcup_{n>0} A_n) = \sum_{n>0} \mathbb{P}(A_n).$$

The pair $(\Omega, \mathbb{P})$ is called a probability space.

Remark 1.9. * Using the language of measure theory, a probability measure is a non-negative measure with total mass equal to 1.

1.4 Examples

Definition 1.10. *If Ω is a finite, non-empty set, the uniform probability measure $\mathbb{P}$ on Ω is the probability measure defined by*

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

for all event A.

Exercise 1.1. Check that $\mathbb{P}$ is indeed a probability measure.

The uniform probability measure is often used to model random experiments where the different possible outcomes happen equally often, or are deemed equally likely to happen.

Example 1.11. We throw a fair die. As outcome space we set $\Omega = \{1,\dots,6\}$, and since the die is fair it is reasonable to consider the uniform probability measure $\mathbb{P}$ on it. With this probability space, for all $i = 1,\dots,6$ the event "the outcome is i" is represented by the event $\{i\}$, and its probability is $\mathbb{P}(\{i\}) = \frac{1}{6}$: this probability does not depend on i. As an example of event, consider $A = \{1,3,5\}$, which represents the event that the result of the throw is an odd number: that event has probability

$$\mathbb{P}(A) = \frac{3}{6} = \frac{1}{2}.$$

Example 1.12. We throw a fair coin three consecutive times. As outcome space we set $\Omega = \{H,T\}^3$, interpreting H as heads and T as tails. For instance, the element $\omega = (H,H,T)$ represents the outcome "Heads, Heads, Tails". Since the coin is fair it is reasonable to consider the uniform probability measure $\mathbb{P}$ on Ω. Under this measure, the event $A = \{(H,H,H), (T,T,T)\}$, which represents the event that the three tosses yield the same outcome, has probability

$$\frac{|A|}{|\Omega|} = \frac{2}{2^3} = \frac{1}{4}.$$

Note that the uniform probability $\mathbb{P}$ on a finite sample space Ω assigns an event a probability proportional to its cardinality (i.e. its number of elements): this reveals a link between probability and cardinality. In fact, many properties of cardinality of a set are also true for the probability of an event, even in the non-uniform case.

1.5 Case of a finite sample space

Assume that Ω is a finite set. There exists a convenient systematic way to define a probability measure on the finite set Ω:

Proposition 1.13. *Let $p:\Omega\to\mathbb{R}$ be a function such that*

1. *p is non-negative, i.e. $p(\omega)\geq 0$ for all $\omega\in\Omega$,*
2. *p is normalized, i.e. $\sum_{\omega\in\Omega}p(\omega)=1$*

Then there exists a unique probability measure $\mathbb{P}$ on Ω such that $\mathbb{P}(\{\omega\})=p(\omega)$ for all $\omega\in\Omega$. That probability measure is given by

$$\mathbb{P}(A)=\sum_{\omega\in A}p(\omega),\qquad A\subset\Omega.$$

Remark 1.14. a function $p:\Omega\to\mathbb{R}$ as above is called a **probability mass function**.

Remark 1.15. If one defines p by

$$p(\omega)=\frac{1}{|\Omega|},\qquad \omega\in\Omega$$

the associated prob. measure $\mathbb{P}$ is the uniform prob. measure on Ω.

1.6 Further properties

Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space.

Proposition 1.16. *The following properties hold:*

1. $\mathbb{P}(\emptyset)=0$.
2. *(additivity) for any finite collection of **pairiwse disjoint** events $A_1,\dots,A_N$, we have*
$$\mathbb{P}(\bigcup_{i=1}^{N}A_i)=\sum_{i=1}^{N}\mathbb{P}(A_i).$$
3. *If A,B are two events and $A\subset B$, then $\mathbb{P}(B\setminus A)=\mathbb{P}(B)-\mathbb{P}(A)$,*
4. *If A,B are two events and $A\subset B$, then $\mathbb{P}(A)\leq\mathbb{P}(B)$,*
5. *for any event A, we have $\mathbb{P}(A^c)=1-\mathbb{P}(A)$.*

Proof. 1. We apply σ-additivity by choosing, for $n\geq 1$, $A_n=\emptyset$. The events A_n are pairwise disjoint so by σ-addivity

$$\mathbb{P}\left(\bigcup_{n\geq 1}\emptyset\right)=\sum_{n\geq 1}\mathbb{P}(\emptyset),$$

i.e. $\mathbb{P}(\emptyset)=\sum_{n\geq 1}\mathbb{P}(\emptyset)$. If we had $\mathbb{P}(\emptyset)\in(0,1]$, we would necessarily have $\sum_{n\geq 1}\mathbb{P}(\emptyset)>\mathbb{P}(\emptyset)$, hence $\mathbb{P}(\emptyset)$ has to be 0.

2. Let $A_1, \ldots, A_N$ be pairwise disjoint events. We complete these into an infinite sequence $(A_n)_{n\geq 1}$ by setting $A_n := \emptyset$ for $n > N$. The sequence thus obtained is made of pairwise disjoint events, hence by σ-additivity
$$\mathbb{P}(\cup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n).$$
Now noting that $\cup_{n\geq 1} A_n = \cup_{n=1}^{N} A_n$ and that $\mathbb{P}(A_n) = \mathbb{P}(\emptyset) = 0$ for $n > N$, the requested property follows.
3. Note that $B = A \cup (B \setminus A)$, where A and $B \setminus A$ are disjoint. Hence, by additivity, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$, which yields the result.
4. By the previous point, we have $\mathbb{P}(A) = \mathbb{P}(B) - \mathbb{P}(B \setminus A) \leq \mathbb{P}(B)$.
5. It suffices to apply point 3. with $B = \Omega$. □

Note that the seond property above holds for events that are **pairwise disjoint**. What can we say if the events are not necessarily disjoint?

Proposition 1.17. *[sub-additivity] for any sequence of events $(A_n)_{n\geq 1}$, we have*
$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_i) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_i).$$

Proof. For $n \geq 1$, we define an event B_n as follows. We set $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup \ldots A_{n-1})$ if $n > 1$. Then the B_n are pairwise disjoint and $\cup_n B_n = \cup_n A_n$, hence
$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \mathbb{P}(\bigcup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mathbb{P}(B_n).$$
Since $B_n \subset A_n$ for all n, we have $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$, and the claim follows. □

We finally state an equality in the case of 2 events that are not necessarily disjoint:

Lemma 1.18. *For any two events A and B (not necessarily disjoint), we have*
$$\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B),$$
or equivalently
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B).$$

Proof. Applying additivity with $n = 3$ and the disjoint events $A_1 = A \cap B^c$, $A_2 = A \cap B$, and $A_3 = A^c \cap B$, and noting that $A_1 \cup A_2 \cup A_3 = A \cup B$, we get
$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B).$$

Thus

$$\begin{aligned}\mathbb{P}(A\cup B)+\mathbb{P}(A\cap B) =& \mathbb{P}(A\cap B^c)+2\,\mathbb{P}(A\cap B)+\mathbb{P}(A^c\cap B)\\ =& \underbrace{(\mathbb{P}(A\cap B^c)+\mathbb{P}(A\cap B))}_{I}\\ &+\underbrace{(\mathbb{P}(A\cap B)+\mathbb{P}(A^c\cap B))}_{II}\end{aligned}$$

Now since the events $A\cap B^c$ and $A\cap B$ are disjoint and their union is A, the quantity I equals $\mathbb{P}(A)$, and similarly the quantity II equals $\mathbb{P}(B)$. Hence

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

1.7 Independence and conditional probability

Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space.

Definition 1.19. *Two events A and B are said to be independent if*

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

Remark 1.20. Note that the above definition involves the probability measure $\mathbb{P}$. In particular sets A and B which are independent under the probability measure $\mathbb{P}$ may no longer be independent when we change the underlying probability measure.

That two events are independent is sometimes an obvious consequence of the assumptions, but sometimes it has to be checked by a computation.

Example 1.21. We throw a fair coin twice. To model this experiment, we consider the probability space $(\Omega,\mathbb{P})$ where $\Omega=\{H,T\}^2$ and $\mathbb{P}$ is the uniform probability measure on Ω. Let

$$A=\{(H,H),(H,T)\} \qquad \text{"1st toss gives Heads"}$$

$$B=\{(H,H),(T,H)\} \qquad \text{"2nd toss gives Heads"}$$

Then

$$\mathbb{P}(A)=\mathbb{P}(B)=\frac{1}{2},$$

while

$$\mathbb{P}(A\cap B)=\mathbb{P}(\{(H,H)\})=\frac{1}{4}.$$

Thus $\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B)$ so A and B are independent.

Example 1.22. We toss a fair die. This is modelled by $\Omega = \{1, \dots, 6\}$ endowed with the uniform probability $\mathbb{P}$. We consider the events

$$A = \{2, 4, 6\} \qquad \textbf{(the outcome is an even number)}$$

and

$$B = \{1, 2, 3, 4\} \qquad \textbf{(the outcome is smaller or equal to 4)}.$$

These events are independent. Indeed, $\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(B) = \frac{2}{3}$, and $\mathbb{P}(A \cap B) = \mathbb{P}(\{2, 4\}) = \frac{1}{3}$, so that

$$\mathbb{P}(A \cap B) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3} = \mathbb{P}(A)\ \mathbb{P}(B).$$

The above definition of independence generalises to an arbitrary, finite number of events, but we stick to the case of two events for simplicity.

Definition 1.23. *Let A and B be two events. **We assume that $\mathbb{P}(B) > 0$**. Then the conditional probability of A given B is*

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \in [0, 1].$$

$\mathbb{P}(A|B)$ represents the probability of A once we know that B occurs. From a frequentist perspective $\mathbb{P}(A|B)$ can be thought of as the limiting fraction of times A occurs among those for which B occurs. In the Bayesian setting, it represents our degree of belief that A will occur once we have the information that B occurs.

Remark 1.24. Given an event B such that $\mathbb{P}(B) > 0$, the map $A \to \mathbb{P}(A|B)$ defines a probability measure on Ω. That probability measure is supported on B, i.e. $\mathbb{P}(B|B) = 1$.

Warning: The quantities $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ are **NOT** the same! Actually the second quantity is not even well-defined if A has probability 0.

The notion of independence can be written in terms of conditional probability.

Lemma 1.25. *Assume $\mathbb{P}(B) > 0$. Then:*

- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\, \mathbb{P}(B)$
- *the events A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.*

Proof. The first point follows at once from the definition of $\mathbb{P}(A|B)$. For the second point, note that the condition for the independence of A and B, namely

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

can be written, upon dividing both sides by $\mathbb{P}(B)$, as

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

as requested. □

Heuristically, independence of A and B means that the a priori knowledge that B occurs does not change the probability that A occurs.

We will now state a couple of results that are often used in Bayesian inference.

Theorem 1.26. *[Law of total probability] Let $A_1,\dots,A_k$ be events that form a partition of Ω i.e. the A_i are pairwise disjoint and $\cup_{i=1}^k A_i = \Omega$. Then, for any event B, we have*

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\ \mathbb{P}(A_i).$$

Proof. We have

$$B = B \cap \Omega = B \cap (\cup_{i=1}^k A_i) = \cup_{i=1}^k B \cap A_i,$$

where we used the fact that $\cup_{i=1}^k A_i = \Omega$ in the third equality. Now, since the events A_i are pairwise disjoint, so are the events $B \cap A_i$, and hence by additivity we get

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}\left(B \cap A_i\right).$$

The result now follows upon applying the first point of Lemma 1.25 to the events $B \cap A_i$. □

Theorem 1.27. *[Bayes' Theorem] Let $A_1, \dots, A_k$ be events that form a partition of Ω i.e. the A_i are pairwise disjoint and $\cup_{i=1}^k A_i = \Omega$. We assume that $\mathbb{P}(A_i) > 0$ for all i. Let be B an event such that $\mathbb{P}(B) > 0$. Then, for all $i = 1, \dots, k$, we have*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\ \mathbb{P}(A_i)}{\sum_{j=1}^k \mathbb{P}(B|A_j)\ \mathbb{P}(A_j)}.$$

Proof. We have

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}\left(B \cap A_i\right)}{\mathbb{P}(B)}.$$

By applying the first point of Lemma 1.25 to the numerator, and Theorem 1.26 to the denominator, we get the result. □

Example 1.28. To see what the above theorem says in a simple case, assume that $k = 2$: we are thus given events A_1 and A_2 that partition Ω, as well as an event B, and we assume that all three events have non-zero probability. Then, choosing $i = 1$, Bayes Theorem says that

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(B|A_1)\ \mathbb{P}(A_1)}{\mathbb{P}(B|A_1)\ \mathbb{P}(A_1) + \mathbb{P}(B|A_2)\ \mathbb{P}(A_2)}.$$

Likewise, for $i=2$, we get

$$\mathbb{P}(A_2|B)=\frac{\mathbb{P}(B|A_2)\ \mathbb{P}(A_2)}{\mathbb{P}(B|A_1)\ \mathbb{P}(A_1)+\mathbb{P}(B|A_2)\ \mathbb{P}(A_2)}.$$

Note that in particular we have $\mathbb{P}(A_1|B)+\mathbb{P}(A_2|B)=1$, which makes sense since, by Remark 1.24 above, we have

$$\mathbb{P}(A_1|B)+\mathbb{P}(A_2|B)=\mathbb{P}\left(A_1\cup A_2|B\right)=\mathbb{P}(\Omega|B)=1.$$

1.8 Exercises

Exercise 1.2. An urn contains 40 balls enumerated from 1 to 40. In a lottery, 6 balls are drawn without replacement from the urn. Tickets bearing the correct sequence of numbers, up to permutation of the numbers, win a T-shirt, while the ticket with the correct ordered sequence wins a car. Compute the probability of winning a T-shirt, resp. a car.

Exercise 1.3. Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space.

- If $(A_n)_{n\geq 1}$ is an increasing sequence of events, i.e. $A_n\subset A_{n+1}$ for all $n\geq 1$, show that
$$\mathbb{P}\left(\bigcup_{n\geq 1}A_n\right)=\lim_{n\to\infty}\uparrow\mathbb{P}(A_n),$$
i.e. that the sequence $(\mathbb{P}(A_n))_{n\geq 1}$ is non-decreasing and converges from below to $\mathbb{P}\left(\cup_{n\geq 1}A_n\right)$.
- If $(A_n)_{n\geq 1}$ is an decreasing sequence of events, i.e. $A_{n+1}\subset A_n$ for all $n\geq 1$, show that
$$\mathbb{P}\left(\bigcap_{n\geq 1}A_n\right)=\lim_{n\to\infty}\downarrow\mathbb{P}(A_n),$$
i.e. that the sequence $(\mathbb{P}(A_n))_{n\geq 1}$ is non-increasing and converges from above to $\mathbb{P}\left(\cap_{n\geq 1}A_n\right)$.

Exercise 1.4. A class in primary school is composed of 25 pupils, all born outside of a leap year. We are interested in the probability that two or more children in the class have the same birthday.

1. Model this experiment with an appropriate sample space Ω and probability measure $\mathbb{P}$. **Hint: you may set Ω** to be the set of all maps from $\{1,2,\ldots,m\}$ to $\{1,\ldots,N\}$, for well-chosen numbers m and N.
2. We recall that the number of maps $f\colon\{1,2,\ldots,m\}\to\{1,\ldots,N\}$ such that $f(i)\neq f(j)$ for all $i\neq j$ is given by
$$\frac{N!}{(N-m)!}.$$
Represent the event that two or more children in the class have the same birthday by an appropriate $A\subset\Omega$, and compute its probability. Check, using Python, that this probability is greater than $1/2$.

Exercise 1.5. The medical test for a disease D has outcomes $+$ (positive) and $-$ (negative). We assume that

- the probability for an individual to have the disease is 0.01,

- the probability for an individual who has the disease to be tested positive is 0.9,
- the probability for an individual who does not have the disease to be tested negative is 0.9,

Compute the probability that an individual who has been tested positive does indeed have the disease. Comment on the quality of the test.

1.9 Appendix: a few useful formulas

We recall a few formulas that come in handy to compute probabilities.

1. If $N \geqslant 1$ and $0 \leqslant m \leqslant N$, the number of ways of picking m numbers $f(1), \dots, f(m)$ from $1, \dots, N$ is given by
$$N^m.$$
Mathematically, this corresponds to the number of maps f from $\{1, \dots, m\}$ to $\{1, \dots, N\}$.
2. If $N \geqslant 1$ and $0 \leqslant m \leqslant N$, the number of ways of picking m *distinct* numbers $f(1), \dots, f(m)$ from $1, \dots, N$ is given by
$$\frac{N!}{(N-m)!} = N(N-1)\dots(N-m+1).$$
Mathematically, this corresponds to the number of maps f from $\{1, \dots, m\}$ to $\{1, \dots, N\}$ that are *injective*, i.e. that satisfy
$$i \neq j \Longrightarrow f(i) \neq f(j)$$
3. If $N \geqslant 1$ and $0 \leqslant m \leqslant N$, the number of ways of picking a subset of m *distinct* elements from $1, \dots, N$ is given by
$$\binom{N}{m} := \frac{N!}{(N-m)!m!}.$$
Mathematically, this corresponds to the number of subsets of the set $\{1, \dots, N\}$ which have cardinality m.

Remark 1.29. Note the difference between the 2nd and the 3rd point above. In the 2nd point we think of the m elements $f(1), \dots, f(m)$ as being drawn one after the other, and we keep track of this labelling. While in the 3rd point we are drawing m elements simultaneously, without labelling them.

Chapter 2
Random variables

2.1 Definition

Let $(\Omega, \mathbb{P})$ be a probability space, and $\mathcal{X}$ be a set.

Definition 2.1. *A random variable (r.v.) X with values in E is a map $X\colon \Omega \to E$.*

Given a random variable $X\colon \Omega \to E$, for any $B \subset E$, we will use the notation $\{X \in B\}$ to denote the preimage of B under X:

$$\{X \in B\} := X^{-1}(B) = \{\omega \in \Omega\colon X(\omega) \in B\}.$$

We will henceforth call E the *outcome* or *target* space.

Remark 2.2. * In the standard theory of probability, a random variable is defined as a measurable map from $(\Omega, \mathcal{A})$ to $(\mathcal{X}, \mathcal{F})$, where $\mathcal{A}$ and $\mathcal{F}$ represent $\sigma-$algebras on Ω and $\mathcal{X}$, respectively. We omit this measurability condition in this course.

When the target space is a subset of $\mathbb{R}$, we say that X is a real-valued random variable. In the statistics literature, "random variable" often refers to real-valued random variables. When the target space is a subset of $\mathbb{R}^d$, $d \geq 2$, we call X a "random vector", or a "multivariate random variable".

Example 2.3. Let's throw two fair dice consecutively. We are interested in the sum of the pips. One can model this experiment by a probability space given by $\Omega = \{1, \ldots, 6\}^2$ endowed with the uniform probability $\mathbb{P}$:

$$\forall (i, j) \in \{1, \ldots, 6\}^2, \qquad \mathbb{P}(\{(i, j)\}) = \frac{1}{36}.$$

The quantity of interest then corresponds to the random variable $X\colon \omega \to \mathbb{R}$ given by

$$X((i, j)) = i + j, \qquad (i, j) \in \{1, \ldots, 6\}^2.$$

Remark 2.4. If X is a random variable on a probability space $(\Omega, \mathbb{P})$, given $\omega \in \Omega$, we call $X(\omega)$ a realisation of that random variable. Physically, one can think of a realization of a random variable as some measurement, or observation performed on a system.

The following is a key notion of probability and statistics.

Definition 2.5. *Let $(\Omega, \mathbb{P})$ be a probability space, E a set, and let $X\colon \Omega \to E$ be a random variable. The probability distribution, or law, of X is the collection of all probabilities*

$$P_X(B) := \mathbb{P}(\{X \in B\}),$$

for all subset B of the target space E.

Proposition 2.6. ** The map P_X defined on all subsets of E by*

$$P_X(B) := \mathbb{P}(\{X \in B\}), \qquad B \subset E,$$

defines a probability measure on E.

Definition 2.7. *Two random variables X and Y with the same target space E are said to be equal in distribution or equal in law if they have the same probability distribution, i.e. if*

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

for all event B in the target space.

Usually, in probability and statistics, we are ultimately interested in the laws of random variables, rather than random variables per se.

Example 2.8. Two players play Heads and Tails on a fair coin. The coin is thrown 10 times, the gain of player 1 (resp. player 2) is the total number of Heads (resp. the total number of tails). This situation is modeled by introducing $\Omega = \{H, T\}^{10}$ endowed with the uniform distribution, and defining random variables X and Y by

$$X(\omega) = \#\{i = 1 \dots 10 \colon \omega_i = H\}, \qquad Y(\omega) = \#\{i = 1 \dots 10 \colon \omega_i = T\},$$

for all $\omega \in \{H, T\}^{10}$. Then $X + Y = 10$. Clearly X and Y are not equal, however they have equal distribution, indeed, for all k,

$$\mathbb{P}(X = k) = \frac{1}{2^{10}}\binom{10}{k} = \frac{1}{2^{10}}\binom{10}{10-k} = \mathbb{P}(X = 10 - k) = \mathbb{P}(Y = k).$$

This implies that X and Y are equal in distribution as we shall see below.

There exist several functions that one can use to characterize the law of a random variable.

2.2 CDF, PMF, and PDF of a random variable

Definition 2.9. *If the target space E is countable, the random variable X is said to be discrete.*

If X is discrete, then for all subset $B \subset E$ of the target space we can write

$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x), \tag{2.1}$$

where $p_X(x) := \mathbb{P}(X = x)$, for all $x \in E$. We call p_X the **probability mass function** (PMF) of X.

Remark 2.10. The PMF p_X of a discrete r.v. X is :

- non-negative: $p_X(x) \geq 0$ for all $x \in E$,
- normalised: $\sum_{x \in E} p_X(x) = 1$

In particular, as a consequence of these two properties, it follows that $p_X(x) \leq 1$ for all $x \in E$.

Proposition 2.11. *The law of a discrete random variable X with target space E is uniquely determined by its PMF.*

Proof. Equation (2.1) shows that, for all event B of the target space,

$$P_X(B) := \mathbb{P}(X \in B) = \sum_{x \in B} p_X(x).$$

Hence we are able to completely retrieve the law of X from its PMF. Hence the claim. □

Definition 2.12. *A function $f : \mathbb{R} \to \mathbb{R}$ is called a **probability density function** (PDF) if the following conditions hold:*

- *$f(x) \geq 0$ for all $x \in \mathbb{R}$,*
- *$\int_{\mathbb{R}} f(x)\, dx = 1$.*

A real-valued random variable is said to be continuous if there exists a PDF $f_X : \mathbb{R} \to \mathbb{R}$ such that, for all $a \leq b$, we have

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx. \tag{2.2}$$

*We then call f_X the **probability density function** (PDF) of X.*

Remark 2.13. * Strictly speaking, the PDF of a given random variable X is not unique: two functions f_X and $\tilde{f}_X$ may both satisfy (2.2) for all subset A as soons as $f_X = \tilde{f}_X$ almost-everywhere (not necessarily everywhere). However, in most practical cases there is a unique "nice" (e.g. continuous) version of a PDF for X. We therefore intentionally ignore this subtlety in the sequel.

Proposition 2.14. *The law of a continuous random variable X with target space E is uniquely determined by its PDF.*

Proof idea Equation (2.2) shows that, for all subset A of $\mathbb{R}$ which is a closed interval, we have

$$P_X(A) := \mathbb{P}(X \in A) = \int_A f_X(x)\, d\, x.$$

By measure-theoretical arguments (see Theorem 5.7 in [4]), we can show that the above equality remains true for any measurable subset A, so we are able to completely retrieve the law of X from its PDF. Hence the claim.

Remark 2.15. One can think of (2.2) as a continuous version of (2.1). However be careful that PMF and PDF are two, quite different, types of functions. For instance, for a continuous r.v. with PDF f_X, the probability $\mathbb{P}(X = x)$ that X equals some real number x is generally not equal to $f_X(x)$: actually, for X a continuous r.v., we have

$$\mathbb{P}(X = x) = \int_x^x f_X(y)\, d\, y = 0.$$

An informal, but morally correct way of thinking of the PDF is

$$f_X(x) \approx \frac{\mathbb{P}(X \in [x - \epsilon, x + \epsilon])}{2\,\epsilon}$$

where $\epsilon > 0$ is very small. PMF and PDF are thus two quite different objects, with different properties, and which have their use in two mutually exclusive contexts: discrete r.v. for the first, continuous r.v. for the second.

2.3 Important examples of discrete random variables

Definition 2.16. *Let $p \in (0, 1)$. Let X be a random variable with values in $\{0, 1\}$ and with PMF given by*

$$Ber(0; p) = 1 - p, \qquad Ber(1; p) = p.$$

We then say that X is a Bernoulli random variable with parameter p, and we write

$$X \overset{(d)}{=} Ber(p) \qquad or \qquad X \sim Ber(p).$$

A Bernoulli r.v. with parameter p represents the result of throwing a coin that falls on Heads and Tails with probability p and $1-p$, respectively ($p=1/2$ if the coin is fair). The next example corresponds to the count of the number of Heads after n consecutive tosses of such a coin.

Definition 2.17. *Let $p \in (0,1)$ and $n \geq 1$ an integer. Let X be a random variable with values in $\{0, n\}$ and with PMF given by*

$$Bin(k; n, p) = \binom{n}{k} p^k \, (1-p)^{n-k}, \qquad k = 0, \ldots, n.$$

We then say that X is a binomial random variable with parameters n and p, and we write

$$X \overset{(d)}{=} Bin(n, p) \qquad or \qquad X \sim Bin(n, p).$$

Let us throw a coin with probability of hitting Heads equal to p, repeatedly. When does the coin hit Heads for the first time?

Definition 2.18. *Let $p \in (0,1)$. Let X be a random variable with values in $\mathbb{N}$ and with PMF given by*

$$Geo(k; p) = (1-p)^{k-1} \, p, \qquad k \geq 1.$$

We then say that X is a geometric random variable with parameter p, and we write

$$X \overset{(d)}{=} Geo(p) \qquad or \qquad X \sim Geo(p).$$

Another important class of discrete probability distribution is the Poisson[2.1] distribution:

Definition 2.19. *Let $\lambda > 0$. Let X be a random variable with values in $\mathbb{N} \cup \{0\}$ and with PMF given by*

$$\mathcal{P}(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k \geq 0.$$

We then say that X is a Poisson random variable with parameter λ, and we write

$$X \overset{(d)}{=} \mathcal{P}(\lambda) \qquad or \qquad X \sim \mathcal{P}(\lambda).$$

Poisson random variables can be used to model the count of rare events such as phone calls in a large city or nuclei deacaying in a radioactive sample.

2.1. Named after the French mathematician Siméon-Denis Poisson (1781-1840).

2.4 Important examples of continuous real-valued random variables

Definition 2.20. *Let $\lambda > 0$. Let X be a real-valued continuous random variable with PDF $f_X(x) = \mathcal{E}(x;\lambda)$ given by*

$$\mathcal{E}(x;\lambda) = \lambda\, e^{-\lambda x}\, 1_{\{x \geq 0\}}, \qquad x \in \mathbb{R}.$$

We then say that X is an exponential random variable with parameter λ, , and we write

$$X \overset{(d)}{=} \mathcal{E}(\lambda) \qquad or \qquad X \sim \mathcal{E}(\lambda).$$

Definition 2.21. *Let $\mu \in \mathbb{R}$ and $\sigma > 0$. Let X be a real-valued continuous random variable with PDF given by*

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}.$$

We then say that X is a Gaussian random variable with parameters μ and σ^2, and we write

$$X \overset{(d)}{=} \mathcal{N}(\mu,\sigma^2) \qquad or \qquad X \sim \mathcal{N}(\mu,\sigma^2).$$

μ is called the mean (or center) and σ is called the standard deviation (or spread) of X.

2.5 Cumulative distribution function (CDF)

Definition 2.22. *The **cumulative distribution function** (CDF) of a **real-valued** random variable X is the function $F_X \colon \mathbb{R} \to [0,1]$ given by*

$$F_X(a) := \mathbb{P}\,(X \leq a), \qquad a \in \mathbb{R}.$$

Note that the CDF is defined for any r.v. taking values in $\mathbb{R}$, whether discrete or continuous.

Proposition 2.23. *Let $F_X \colon \mathbb{R} \to [0,1]$ be the CDF of a real-valued r.v. X. Then:*

- *F_X is non-decreasing: if $a \leq b$ then $F_X(a) \leq F_X(b)$.*
- *F_X is right-continuous: for all $a \in \mathbb{R}$,*

$$F_X\,(a+) := \lim_{\substack{b \to a \\ b > a}} F_X(b) = F_X(a)$$

- $F_X\,(-\infty) := \lim_{a \to -\infty} F_X(a) = 0, \qquad F_X\,(+\infty) := \lim_{a \to +\infty} F_X(a) = 1.$

One can read off relevent information on the distribution of X from its CDF.

Lemma 2.24. *Let $F_X: \mathbb{R} \to [0,1]$ be the CDF of a real-valued r.v. X. Then:*

- *For any real numbers $a < b$,*
$$\mathbb{P}(X \in (a,b]) = F_X(b) - F_X(a),$$
- *For any $a \in \mathbb{R}$,*
$$\mathbb{P}(X > a) = 1 - F_X(a),$$
- *For any $x \in \mathbb{R}$,*
$$\mathbb{P}(X = x) = F_X(x) - F_X(x-),$$

where $F_X(x-) := \lim_{\substack{y \to x \\ y < x}} F_X(y)$.

Remark 2.25. In particular, if X is a continuous r.v., we have $F_X(x) = F_X(x-)$ for all $x \in \mathbb{R}$: no jumps occur. For a discrete r.v. the situation is very different: F_X is then a pure-jump function, meaning that it increases purely through jumps.

2.6 Relation between CDF and PMF/PDF

Proposition 2.26. *[Discrete case]Let X be a discrete r.v. taking values in a countable subset E of $\mathbb{R}$. Denoting the PMF of X by p_X, and its CDF by F_X, we have*

$$\forall a \in \mathbb{R}, \qquad F_X(a) = \sum_{\substack{x \in \mathcal{X} \\ x \le a}} p_X(x),$$

$$\forall x \in \mathcal{X}, \qquad p_X(x) = F_X(x) - F_X(x-).$$

Proof. The first relation follows upon applying equality (2.1) with $B = \{x \in E : x \le a\}$. For the second relation, note that

$$\{X = x\} = \bigcap_{n \ge 1} \{X \in (x - 1/n, x]\}.$$

Since the sets $\{X \in (x-1/n, x]\}$ form a decreasing sequence of events, by Exercise 1.3 above we have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{n \ge 1} \{X \in (x-1/n, x]\}\right) &= \lim_{n \to \infty} \mathbb{P}(X \in (x-1/n, x]) \\ &= \lim_{n \to \infty} F_X(x) - F_X(x - 1/n) \\ &= F_X(x) - F_X(x-),\end{aligned}$$

whence the claim. □

Proposition 2.27. *[Continuous case] Let X be a continuous real-valued r.v. Denoting the PDF of X by f_X, and its CDF by F_X, we have*

$$\forall a \in \mathbb{R}, \qquad F_X(a) = \int_{-\infty}^{a} f_X(y)\ dy.$$

*In addition, **if F_X is differentiable at x**, we have*

$$\forall x \in E, \qquad f_X(x) = F_X'(x).$$

Proof. For the first statement, note that, for all $u < a$, we have

$$F_X(a) - F_X(u) = \mathbb{P}\left(X \in (u, a]\right) = \mathbb{P}\left(X \in [u, a]\right) = \int_{u}^{a} f_X(y)\ dy$$

where in the second equality we used the fact that $\mathbb{P}(X = u) = 0$ since X is a continuous r.v. Sending $u \to \infty$, and recalling that $F_X(-\infty) = 0$, we obtain $F_X(a) = \int_{-\infty}^{a} f_X(y)\ dy$ as requested. We admit the second statement. □

Proposition 2.28. *The probability distribution of a real-valued r.v. is uniquely determined by its CDF*

Proof. We give a proof in the discrete case. Let X and Y be two discrete real-valued r.v. sharing the same CDF:

$$\forall a \in \mathbb{R}, \qquad F_X(a) = F_Y(a).$$

By Proposition 2.26 above, it follows that, for all $x \in E$,

$$p_X(x) = F_X(x) - F_X(x-) = F_Y(x) - F_Y(x-) = p_Y(x).$$

Thus X and Y have the same PMF, whence it follows that $X \overset{(d)}{=} Y$. □

2.7 Quantile function

Definition 2.29. *[Quantile function] Let X be a real-valued r.v. with CDF F. We define the quantile function $F^{-1}\colon (0,1) \to \mathbb{R}$ of X as*

$$F^{-1}(q) := \inf\{x \in \mathbb{R}, F(x) > q\}, \qquad q \in \mathbb{R}.$$

Beware, F^{-1} is not always the proper inverse of the CDF F, as F may not be invertible. For instance, the CDF of a Bernoulli random variable of paramter $1/2$ is given by

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1/2 & \text{if} \quad 0 \le x < 1 \\ 1 & \text{if} \quad x \ge 1 \end{cases}.$$

Such a function is not invertible in the usual sense as it is not injective. However the quantile function F^{-1} as defined above does still make sense.

Exercise 2.1. Compute F^{-1} in that case.

Proposition 2.30. *Let X be a real-valued random variable with CDF F_X. Then*

1. *For all $q \in (0,1)$, $F_X(F_X^{-1}(q)) \geq q$.*
2. *If X is a continuous r.v. then, for all $q \in (0,1)$, $F_X(F_X^{-1}(q)) = q$.*

Proof. 1. Let $q \in (0,1)$. By definition of $F_X^{-1}(q)$, there is a sequence of real numbers a_n such that $F_X(a_n) > q$ and converging to $F_X^{-1}(q)$ from above. By right-continuity of F_X, we then have

$$F_X(F_X^{-1}(q)) = \lim_{n\to\infty} F_X(a_n) \geq q.$$

2. From the previous point we have $F_X(F_X^{-1}(q)) \geq q$. Assume by contradiction that $F_X(F_X^{-1}(q)) > q$. Since F_X is CDF of a continuous r.v., it is continuous. By continuity of F_X at the point $F_X^{-1}(q)$ there exists some $a < F_X^{-1}(q)$ such that $F_X(a) > q$, but this contradicts the definition of $F_X^{-1}(q)$. Hence $F_X(F_X^{-1}(q)) = q$ as requested. □

Beware, even in the continuous setting, we do not necessarily have $F^{-1}(F(q)) = q$ for all $q \in (0,1)$.

Example 2.31. [CDF of a normal random variable]The CDF of a normal random variable X is often denoted by Φ.

$$\Phi(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{t^2}{2}\right) d\,t, \qquad x \in \mathbb{R}.$$

Typical values to remember:

$$\Phi(1.645) = \mathbb{P}(X \leq 1.645) \approx 0.95$$

$$\Phi(1.960) = \mathbb{P}(X \leq 1.960) \approx 0.975$$

In this case the CDF Φ is injective and the quantile function, denoted by Φ^{-1}, coincides with its inverse. The above equalities can be re-expressed as

$$\Phi^{-1}(0.95) \approx 1.645,$$

$$\Phi^{-1}(0.975) \approx 1.960.$$

2.8 Exercises

Exercise 2.2. John claims: "If X is a continuous r.v. with PDF f_X, then for all $x \in \mathbb{R}$,

$$\mathbb{P}(X = x) = \int_x^x f_X(y)\, d\,y = 0.$$

Therefore, for any event A,

$$\mathbb{P}(X \in A) = \mathbb{P}\left(\bigcup_{x \in A} \{X = x\} \right) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} 0 = 0,$$

where we used the fact that the sets $\{X = x\}$ are disjoint for different values of x". What do you think of John's claim? Carefully justify your answer.

Exercise 2.3. Let $(p_n)_{n \geq 1}$ be a sequence of real numbers in $(0, 1)$ and $\lambda > 0$. We assume that $\lim_n n\, p_n = \lambda$. Show that, for all $k \geq 0$,

$$\lim_{n \to \infty} \mathrm{Bin}(k; n, p_n) = \mathcal{P}(k; \lambda).$$

How do you interpret this result?

Exercise 2.4. Compute the CDF of:

- a geometric r.v.
- an exponential r.v.

Exercise 2.5. [Median of a r.v.] For a real-valued r.v. X, we say that $\mathfrak{m} \in \mathbb{R}$ is a median for X if $\mathbb{P}(X \geq \mathfrak{m}) \geq 1/2$ and $\mathbb{P}(X \leq \mathfrak{m}) \geq 1/2$.

1. If X is a continuous r.v. with CDF F, show that a median is provided by $F^{-1}(1/2)$.
2. Compute a median for X when X is a uniform random variable. Do the same with a Gaussian random variable and an exponential random variable.

Chapter 3
Joint distributions

3.1 Definition: joint distribution, marginal distributions

Let $(\Omega, \mathbb{P})$ be a probability space, and let X, Y be two random variables with target spaces E and F. One can see (X, Y) as a random variable taking values in $E \times F$.

Definition 3.1. *The joint distribution of (X, Y) is the collection of probabilities*

$$P_{X,Y}(C) := \mathbb{P}\left((X, Y) \in C\right),$$

for all subset C of the target set $E \times F$.

By contrast to the above definition, we shall refer to

$$P_X(A) := \mathbb{P}\left(X \in A\right), \qquad A \subset E$$

and

$$P_Y(B) := \mathbb{P}\left(Y \in B\right), \qquad B \subset F$$

as the marginal distributions of (X, Y). Note the relations

$$P_X(A) = P_{X,Y}\left(A \times F\right)$$

and

$$P_Y(A) = P_{X,Y}\left(E \times B\right),$$

which allow to retrieve the marginal distributions from the joint distribution.

Warning: The joint distribution of (X, Y) is in general **NOT** determined by its marginal distributions.

3.2 Discrete case

Assume that X and Y are discrete, i.e. E and F are countable. Then the product space $E \times F$ is countable as well, and (X, Y) may thus be studied as a discrete r.v.

Definition 3.2. *The joint PMF of* (X,Y) *is the function* $p_{X,Y}\colon E\times F\to[0,1]$ *defined, for all* $(x,y)\in E\times F$*, by*

$$p_{X,Y}(x,y):=\mathbb{P}\left((X,Y)=(x,y)\right)=\mathbb{P}(X=x,Y=y).$$

Note that, as in the univariate case, the joint PMF of (X,Y) is:

1. non-negative: $p_{X,Y}(x,y)\geq 0$ for all $(x,y)\in E\times F$
2. normalised: $\sum_{x,y}p_{X,Y}(x,y)=1$.

Lemma 3.3. *If* (X,Y) *admits a joint PMF* $p_{X,Y}$*, then the PMFs of* X *and* Y *are respectively given by*

$$p_X(x)=\sum_{y\in F}p_{X,Y}(x,y),\qquad x\in E,$$

and

$$p_Y(y)=\sum_{x\in E}p_{X,Y}(x,y),\qquad y\in F.$$

We call p_X *and* p_Y *the marginal PMFs of* (X,Y)

Example 3.4. Let (X,Y) be a bi-variate r.v. taking values in $\{1,2\}\times\{1,2,3\}$ and with joint PMF p given as below

$p(x,y)$	$y=1$	$y=2$	$y=3$
$x=1$	0.1	0.3	0.2
$x=2$	0.2	0.2	0

The values of the marginal PMF $p_X(x)$ for $x=1,2$ are obtained by summing up the probabilities in each of the corresponding rows. Thus

$$p_X(1)=0.1+0.3+0.2=0.6,$$

$$p_X(2)=0.2+0.2+0=0.4.$$

Likewise, the values of the marginal PMF $p_Y(y)$ for $y=1,2,3$ are obtained by summing up the probabilities in each of the corresponding columns:

$$p_Y(1)=0.1+0.2=0.3,\qquad p_Y(2)=0.3+0.2=0.5,\qquad p_Y(3)=0.2+0=0.2.$$

3.3 Continuous case

A bi-variate PDF is defined in a way very similar to the univariate case.

Definition 3.5. *A function $f:\mathbb{R}^2\to\mathbb{R}$ is called a* ***probability density function*** *(PDF) if the following conditions hold:*

- $f(x,y)\geq 0$ *for all* $(x,y)\in\mathbb{R}^2$,
- $\int_{\infty}^{+\infty}\int_{\infty}^{+\infty} f(x,y)\,dy\,dx=1$.

Two real-valued random variables X and Y are said to admit a continuous joint distribution (or to admit a joint density) if there exists a PDF $f_{X,Y}$: $\mathbb{R}^2\to\mathbb{R}$ such that, for all subset A of $\mathbb{R}^2$, we have

$$\mathbb{P}\left((X,Y)\in A\right)=\int_A f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy}\,. \tag{3.1}$$

We then call f_X the ***probability density function*** *(PDF) of X.*

Lemma 3.6. *If (X,Y) admits a joint density $f_{X,Y}$, then X and Y are continuous r.v. with pdf respectively given by*

$$f_X(x)=\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy,\qquad x\in\mathbb{R},$$

and

$$f_Y(y)=\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx,\qquad y\in\mathbb{R}.$$

We call f_X and f_Y the marginal PDFs of (X,Y).

Warning: If X and Y are both continuous random variables, this does **NOT** necessarily imply that (X,Y) has a continuous joint distribution. For instance, if $X=Y\sim N(0,1)$, then X and Y are clearly both continuous random variables, however, (X,Y) does not admit a joint density (exercise: justify why).

Example 3.7. Let $a,b,c,d\in\mathbb{R}^2$ such that $a<b$ and $c<d$. Then the function $f:\mathbb{R}^2\to\mathbb{R}$ given by

$$f(z)=\frac{1}{(b-a)\,(d-c)}\,1_{[a,b]\times[c,d]}(z),\qquad z\in\mathbb{R}^2,$$

is a PDF. It corresponds to the *uniform distribution* on the rectangle $[a, b]\times[c,d]$. The marginal distributions are univariate uniform distributions on the intervals $[a,b]$ and $[c,d]$, respectively;

$$X\sim\mathcal{U}(a,b),\qquad Y\sim\mathcal{U}(c,d)$$

Example 3.8. [bivariate Gaussian distribution] Let $\mu\in\mathbb{R}^2$ and let $K\in\mathbb{R}^{2\times 2}$ be a symmetric, positive definite 2-dimensional square matrix. The function $f:\mathbb{R}^2\to\mathbb{R}$ given by

$$f(z)=\frac{1}{2\,\pi\,\sqrt{\det(K)}}\exp\left(-\frac{1}{2}\,(z-\mu)^T K^{-1}\,(z-\mu)\right),\qquad z\in\mathbb{R}^2,$$

is a PDF. A random vector $Z=(X,Y)$ with PDF p is said to have Gaussian distribution with mean μ and covariance matrix K (see below for justification of this treminology). Denoting

$$\mu=\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix},\qquad K=\begin{pmatrix}\sigma_X^2 & \sigma_{XY}^2\\ \sigma_{XY}^2 & \sigma_Y^2\end{pmatrix},$$

then the marginal PDFs are given by

$$f_X(x)=\frac{1}{\sqrt{2\,\pi\,\sigma_X^2}}\exp\left(-\frac{(x-\mu_X)^2}{2\,\sigma_X^2}\right)$$

and

$$f_Y(y)=\frac{1}{\sqrt{2\,\pi\,\sigma_Y^2}}\exp\left(-\frac{(y-\mu_Y)^2}{2\,\sigma_Y^2}\right).$$

Thus $X\sim\mathcal{N}(\mu_X,\sigma_X^2)$ and $Y\sim\mathcal{N}(\mu_Y,\sigma_Y^2)$. A very special case corresponds to $\mu=0$ and $K=I_2$, that is $\mu_X=\mu_Y=0$, $\sigma_{XY}^2=0$ and $\sigma_X^2=\sigma_Y^2=1$. Then

$$f(z)=\frac{1}{2\,\pi}\exp\left(-\frac{1}{2}\|z\|^2\right),\qquad z\in\mathbb{R}^2,$$

where, for $z=(x,y)$, $\|z\|:=\sqrt{x^2+y^2}$ denotes the Euclidean norm of z.

3.4 Independence

Let $(\Omega,\mathbb{P})$ be a probability space, and let again X,Y be two random variables with target spaces E and F.

Definition 3.9. *X and Y are said to be **independent** if, for any subsets A of E and B of F, the events $\{X\in A\}$ and $\{Y\in B\}$ are independent, i.e.*

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

Theorem 3.10. *[Independence: discrete case] Assume that (X,Y) is a discrete random variable, with joint PMF $p_{X,Y}$ and marginal PMFs p_X and p_Y. Then X and Y are independent if and only if*

$$p_{X,Y}\,(x,y)=p_X(x)\;p_Y(y),\qquad (x,y)\in E\times F.$$

Proof. First assume that X and Y are independent. Then for all $x\in E$ and $y\in F$,

$$p_{X,Y}\,(x,y)=\mathbb{P}(X=x,Y=y)=\mathbb{P}\,(X=x)\;\mathbb{P}\,(Y=y)=p_X(x)\;p_Y(y),$$

where we used independence of X and Y in the second equality. Conversely, assume that $p_{X,Y}\,(x,y)=p_X(x)\;p_Y(y)$ for all $(x,y)\in E\times F$. Then for all subsets A of E and B of F respectively, we have

$$\mathbb{P}(X\in A,Y\in B)=\sum_{(x,y)\in A\times B}p_{X,Y}\,(x,y)=\sum_{(x,y)\in A\times B}p_X(x)\;p_Y(y).$$

The last sum factorizes and takes the form

$$\sum_{x\in A} p_X(x) \sum_{y\in B} p_Y(y) = \mathbb{P}\,(X\in A)\ \mathbb{P}\,(X\in B).$$

Thus we have shown $\mathbb{P}(X\in A, Y\in B) = \mathbb{P}\,(X\in A)\ \mathbb{P}\,(X\in B)$, so A and B are indeed independent. □

A similar result can be shown in the continuous case. We will not give a proof.

Theorem 3.11. *[Independence: continuous case] Assume that (X,Y) is a continuous random variable. with joint PDF $f_{X,Y}$ and marginal PDFs f_X and f_Y. Then X and Y are independent if and only if*

$$f_{X,Y}(x,y) = f_X(x)\ f_Y(y), \qquad (x,y)\in E\times F.$$

3.5 Conditional distribution

Definition 3.12. *Let (X,Y) be a discrete random variable with joint PMF $p_{X,Y}$ and marginal PMFs p_X and p_Y. The conditional PMF $p_{X|Y}$ of X given Y is defined by*

$$p_{X|Y}(x|y) := \frac{p_{X,Y}\,(x,y)}{p_Y(y)},$$

for all $x\in E$ and $y\in F$ such that $p_Y(y)>0$.

Definition 3.13. *Let (X,Y) be a continuous random variable in $\mathbb{R}^2$ with joint PDF $f_{X,Y}$ and marginal PDFs f_X and f_Y. The conditional PDF $f_{X|Y}$ of X given Y is defined by*

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

for all $x\in\mathbb{R}$ and $y\in\mathbb{R}$ such that $f_Y(y)>0$.

3.6 Multivariate distributions

Let $n\geq 1$ and let $X_1,\ldots,X_n$ be n random variables defined on the same probability space $(\Omega,\mathbb{P})$, and with realization spaces $E_1,\ldots,E_n$. One can see the map

$$X := (X_1,\ldots,X_n)\colon \omega\mapsto (X_1(\omega),\ldots,X_n(\omega))$$

as a single, multi-variate, random variable with values in the product space $E_1\times\ldots\times E_n$.

Remark 3.14. When $X_1, \ldots, X_n$ all take values in $\mathbb{R}$, the random variable $(X_1, \ldots, X_n)$ takes values in $\mathbb{R}^n$: it is then common to refer to $(X_1, \ldots, X_n)$ as an n-dimensional random **vector**.

The definitions and properties stated above for bi-variate random variables can be naturally generalised to the multi-variate case.

Definition 3.15. *By the **joint distribution** of $X_1, \ldots, X_n$ we mean the specification of the probabilities*

$$P_{X_1,\ldots,X_n}(C) := \mathbb{P}\left((X_1, \ldots, X_n) \in C\right),$$

for all $C \in E_1 \times \ldots \times E_n$. In contrast the respective probability distributions $P_{X_1}, \ldots, P_{X_n}$ of the random variables $X_1, \ldots, X_n$ are referred to as the marginal probabilities.

Remark 3.16. For all $i = 1, \ldots, n$, the marginal distribution P_{X_i} is related to the joint distribution $P_{X_1,\ldots,X_n}$ by

$$P_{X_i}(A) = P_{X_1,\ldots,X_n}(E_1 \times \ldots, E_{i-1} \times A \times E_{i+1} \times \ldots \times E_n).$$

for all event $A \subset E_i$.

3.6.1 Discrete case

Definition 3.17. *Assume that the realization spaces $E_1, \ldots, E_n$ are countable, so that $X_1, \ldots, X_n$ are discrete random variables. Their joint PMF $p_{X_1,\ldots X_n}: E_1 \times \ldots \times E_n \to [0,1]$ is defined as*

$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) := \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n), \qquad (x_1,\ldots,x_n) \in E_1 \times \ldots \times E_n$.

Proposition 3.18. *For all event $C \subset E_1 \times \ldots \times E_n$, we have*

$$P_{X_1,\ldots,X_n}(C) = \sum_{(x_1,\ldots,x_n)\in C} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

Proof. Follows by σ-additivity upon noting the equality of events

$$\{(X_1, \ldots, X_n) \in C\} = \bigcup_{(x_1,\ldots,x_n)\in C} \{X_1 = x_1, \ldots, X_n = x_n\}. \qquad \square$$

One can recover the individual PMFs of the random variables $X_1, \ldots, X_n$ from the joint PMF $p_{X_1,\ldots,X_n}$

$$p_{X_i}(x) = \sum_{\substack{x_1\in E_1 \\ \ldots \\ x_{i-1}\in E_{i-1} \\ x_{i+1}\in E_{i+1} \\ \ldots \\ x_n \in E_n}} p_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n).$$

Remark 3.19. More generally, for any subset of indices $I \subset \{1, \dots, k\}$ we can recover the joint PMF of the random variables X_i, $i \in I$ from the joint PMF of $X_1, \dots, X_n$, by summing up $p_{X_1,\dots,X_n}$ over all possible values in the coordinates $j \notin I$. For instance, if $n = 4$, we can recover the joint PMF of X_2, X_3 as follows:

$$p_{X_2,X_3}(x, y) = \sum_{\substack{x_1 \in E_1 \\ x_4 \in E_4}} p_{X_1,X_2,X_3,X_4}(x_1, x, y, x_4).$$

3.6.2 Continuous case

A multi-variate PDF is defined in a way very similar to the univariate or bivariate case.

Definition 3.20. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a **probability density function** (PDF) if the following conditions hold:*

- $f(x_1, \dots, x_n) \geq 0$ *for all* $(x_1, \dots, x_n) \in \mathbb{R}^n$,
- $\int_{\infty}^{+\infty} \dots \int_{\infty}^{+\infty} f(x_1, \dots, x_n)\, d\,x_n \ \dots,\, d\,x_1 = 1$.

n real-valued random variables $X_1, \dots, X_n$ are said to admit a continuous joint distribution (or to admit a joint density) if there exists a PDF $f_{X_1,\dots,X_n} : \mathbb{R}^n \to \mathbb{R}$ such that, for all subset A of $\mathbb{R}^n$, we have

$$\mathbb{P}\left((X_1, \dots, X_n) \in A\right) = \int_A f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, d\,x_1 \ \dots\ d\,x_n. \tag{3.2}$$

*We then call $f_{X_1,\dots,X_n}$ the **probability density function** (PDF) of X.*

Lemma 3.21. *If $X_1, \dots, X_n$ admit a joint density $f_{X_1,\dots,X_n}$, then $X_1, \dots, X_n$ are continuous r.v. with PDF given by*

$$f_{X_i}(x) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f_{X_1,\dots,X_n}(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n)\, d\,x_n \dots d\,x_{i+1}\, d\,x_{i-1}, \dots, d\,x_1, \qquad x \in \mathbb{R}.$$

We call f_{X_i}, $i = 1, \dots, n$, the marginal PDFs of $X_1, \dots, X_n$.

Thus the value at x of the marginal PDF in the i-th coordinate is given by the joint PDF where we insert x as the i-th coordinate and we integrate over all possible values in the other coordinates.

Remark 3.22. More generally, for any subset of indices $I \subset \{1, \dots, k\}$ we can recover the joint PDF of the random variables X_i, $i \in I$ from the joint PDF of $X_1, \dots, X_n$, by integrating $f_{X_1,\dots,X_n}$ over all possible values in the coordinates $j \notin I$. For instance, if $n = 4$, we can recover the joint PDF of X_2, X_3 as follows:

$$f_{X_2,X_3}(x, y) = \int_{x_1=-\infty}^{+\infty} \int_{x_4=-\infty}^{+\infty} f_{X_1,X_2,X_3,X_4}(x_1, x, y, x_4)\, d\,x_4\, d\,x_1.$$

3.6.3 Independence

Definition 3.23. *The random variables $X_1, \ldots, X_n$ are said to be **independent** if, for any subsets $A_1 \subset E_1, A_2 \subset E_2, \ldots, A_n \subset E_n$, we have*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \ \ldots \ \mathbb{P}(X_n \in A_n).$$

Theorem 3.24. *[Independence: discrete case] Assume that $X_1, \ldots X_n$ are discrete random variables, with joint PMF $p_{X_1,\ldots,X_n}$ and marginal PMFs $p_{X_1}, \ldots, p_{X_n}$. Then $X_1, \ldots X_n$ are independent if and only if*

$$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n), \qquad (x_1,\ldots,x_n) \in E_1 \times \ldots \times E_n.$$

Theorem 3.25. *[Independence: continuous case] Assume that $X_1, \ldots X_n$ are discrete random variables admitting a joint density $f_{X_1,\ldots,X_n}$ and marginal PDFs $f_{X_1}, \ldots, f_{X_n}$. Then $X_1, \ldots X_n$ are independent if and only if*

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n), \qquad (x_1,\ldots,x_n) \in \mathbb{R}^n.$$

Chapter 4

Transformations of random variables

4.1 Probability integral transform

Let X be a continuous real-valued random variable with CDF F_X and quantile function F_X^{-1}.

Theorem 4.1. *1. The random variable $U = F_X(X)$ is distributed uniformly in $[0,1]$, i.e. $U \sim \mathcal{U}(0,1)$.*

2. If U is a uniform random variable in $[0,1]$, then the random variable $F_X^{-1}(U)$ has the same distribution as X, i.e. $F_X^{-1}(U) \overset{(d)}{=} X$.

The above theorem is very useful for simulations: it allows one to simulate any real-valued random variable X using a uniform random variable on $[0,1]$. To do so, one "only" needs to have an expression for the CDF of X. More generally, we can simulate n i.i.d. random variables $X_1, \dots, X_n$ with common CDF F as follows: we simulate n i.i.d. uniform r.v.'s on $[0,1]$, $U_1, \dots, U_n$, and construct the sequence $F^{-1}(U_1), \dots, F^{-1}(U_n)$.

Example 4.2. Assume we want to simulate an exponential random variable X. That is $X \sim \mathcal{E}(\lambda)$ for some $\lambda > 0$, see Definition 2.20. The associated CDF F_X is given by

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\, dx = 1_{[0,+\infty)}(a) \int_0^a \lambda\, e^{-\lambda x}\, dx = 1_{[0,+\infty)}(a)\, (1 - e^{-\lambda a}).$$

It follows that

$$F_X^{-1}(q) = -\frac{1}{\lambda} \log(1-q), \qquad q \in (0,1).$$

Hence, to simulate $X \sim \mathcal{E}(\lambda)$, it suffices to generate a random variable $U \sim \mathcal{U}(0,1)$, and compute $X = -\frac{1}{\lambda} \log(1-U)$.

Exercise 4.1. By noting that if $U \sim \mathcal{U}(0,1)$, then $V := 1 - U$ also satifies $V \sim \mathcal{U}(0,1)$, show that, in the above, one can alternatively compute $X = -\frac{1}{\lambda} \log(U)$ to obtain the desired result.

4.2 Change of variable formulae

Suppose we know the distribution of a random variable X, and $Y = r(X)$, where f is some map. How can one compute the distribution of Y?

4.2.1 Discrete Case

Assume that X and Y are discrete random variables with outcome spaces E, F and PMF p_X, p_Y, respectively. Assume that $y = r(X)$, where $r: E \to F$ is a map.

Proposition 4.3. *The PMF of Y is given by:*

$$p_Y(y) = \sum_{\substack{x \in E \\ r(x)=y}} p_X(x).$$

In words, the PMF of Y at a point y is computed by summing up the PMF of X over the preimage of $\{y\}$.

Proof. We have

$$p_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(r(X) = y).$$

Now the event $\{r(X) = y\}$ can be rewritten as the disjoint union of events

$$\bigcup_{\substack{x \in E \\ r(x)=y}} \{X = x\},$$

so by σ-additivity,

$$\mathbb{P}(r(X) = y) = \sum_{\substack{x \in E \\ r(x)=y}} \underbrace{\mathbb{P}(X = x)}_{p_X(x)}.$$ □

4.2.2 Continuous case: univariate framework

The continuous case is more involved. Let us assume that X is a continuous real-valued r.v. with PDF X, and that $r: \mathbb{R} \to \mathbb{R}$ is a function. Does the r.v. $Y := r(X)$ admit a PDF? If so, how can one compute it? We first make a remark:

Remark 4.4. Denoting by F_Y the CDF of Y, we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(X \in A_y), \qquad y \in \mathbb{R},$$

where $A_y = \{x \in \mathbb{R}: r(x) \leq y\}$

Here is a general method to derive the PDF of Y, which is based on first computing the CDF of Y.

Three steps for transformations (see Wasserman, Section 2.12):

1. For $y \in \mathbb{R}$, compute the set
$$A_y := \{x \in \mathbb{R} : r(x) \leq y\}$$
2. Compute the CDF F_Y of Y using the relation
$$F_Y(y) = \mathbb{P}(X \in A_y)$$
3. If F_Y is differentiable, then Y has a PDF f_Y given by $f_Y = F_Y'$.

Example 4.5. Assume that $X \sim \mathcal{U}(0,1)$. Let $r : \mathbb{R} \to \mathbb{R}$ be the map given by $r(x) = x^2$, for $x \in \mathbb{R}$. We compute the PDF of $r(X)$ using the above method. First note that, for $y \in \mathbb{R}$, we have
$$A_y = \{x \in \mathbb{R} : x^2 \leq y\} = \begin{cases} \emptyset & \text{if} \quad y < 0 \\ [-\sqrt{y}, \sqrt{y}], & \text{if} \quad y \geqslant 0 \end{cases}.$$
Now, if $y \geq 0$, we have
$$\mathbb{P}(X \in [-\sqrt{y}, \sqrt{y}]) = \int_{-\sqrt{y}}^{\sqrt{y}} 1_{[0,1]}(x)\, dx = \begin{cases} \sqrt{y} & \text{if} \quad y \in [0,1] \\ 1 & \text{if} \quad y > 1. \end{cases}$$
It follows that
$$F_Y(y) = \begin{cases} 0 & \text{if} \quad y < 0 \\ \sqrt{y} & \text{if} \quad y \in [0,1] \\ 1 & \text{if} \quad y > 1. \end{cases}$$
By differentiation we obtain
$$f_Y(y) = 1_{[0,1]}(y)\, \frac{1}{2\sqrt{y}}, \qquad y \in \mathbb{R}.$$

In the special case where $r : \mathbb{R} \to \mathbb{R}$ is a strictly increasing (or strictly decreasing) function, one has a general formula:

Theorem 4.6. *Assume that $r : \mathbb{R} \to \mathbb{R}$ is a C^1-differentiable and strictly increasing (or strictly decreasing) function. Then Y admits a PDF f_Y given by*
$$f_Y(y) = \frac{1}{|r'(r^{-1}(y))|}\, f_X(r^{-1}(y)), \qquad y \in \mathbb{R}. \tag{4.1}$$

4.2.3 Continuous case: multivariate setting

The above methods generalise to higher dimensions. For instance, assume that we have two realvalued random variables X, Y, a function $r : \mathbb{R}^2 \to \mathbb{R}$, and we wish to derive the PDF of the real-valued random variable $Z = r(X, Y)$. For example, Z could be $X + Y$, $X - Y$, $\min(X, Y)$, etc.

Three steps for transformations (see Wasserman, Section 2.12):

1. For $z \in \mathbb{R}$, compute the set
$$A_z := \{(x, y) \in \mathbb{R}^2 : r(x, y) \leq z\}$$
2. Compute the CDF F_Z of Z using the relation
$$F_Z(z) = \mathbb{P}\left((X, Y) \in A_z\right)$$
3. If F_Z is differentiable, then Z has a PDF f_Z given by $f_Z = F_Z'$.

Example 4.7. Assume that $X, Y \sim \mathcal{U}(0,1)$ (i.e. X, Y are indepdent uniform random variables on $[0,1]$). Let $Z = \max(X, Y)$. We can compute the PDF f_Z of Z using the above methods. Since Z takes values in $[0,1]$, we will have $f_Z(z) = 0$ for $z \notin [0,1]$. Let now $z \in [0,1]$. First note that

$$A_z = \{(x, y) \in \mathbb{R}^2 : \max(x, y) \leq z\} = \{(x, y) \in \mathbb{R}^2 : x \leq z \text{ and } y \leq z\} = (-\infty, z]^2.$$

Hence,
$$F_Z(z) = \mathbb{P}\left((X, Y) \in A_z\right) = \mathbb{P}(X \leq z,\, Y \leq z).$$

Since $X \perp\!\!\!\perp Y$ and both r.v.'s are uniformly distributed in $(0,1)$, we get

$$F_Z(z) = \mathbb{P}(X \leq z)\ \mathbb{P}(Y \leq z) = \left(\int_{-\infty}^{z} 1_{[0,1]}(t)\ dt\right)^2 = z^2.$$

Differentiating the above yields $f_Z(z) = 2\,z$, which holds for all $z \in [0,1]$. Finally:
$$f_Z(z) = 2\ 1_{[0,1]}(z)\, z, \qquad z \in \mathbb{R}.$$

A last change of variable tool we present pertains to the case where X is an n-dimensional random vector, $n \geq 1$, $r : \mathbb{R}^n \to \mathbb{R}^n$, and we aim at obtaining the PDF of the random vector $Y := r(X)$.

Theorem 4.8. *Assume that $r : \mathbb{R}^n \to \mathbb{R}^n$ is C^1-diffeormorphism, i.e., r is a bijective transformation of $\mathbb{R}^n$, and both r and r^{-1} are C^1-continuous. Then $Y = r(X)$ admits the PDF f_Y given by*

$$f_Y(y) = \frac{1}{|\det\left(J^r(r^{-1}(y))\right|}\, f_X(r^{-1}(y), \qquad y \in \mathbb{R}^n. \tag{4.2}$$

Example 4.9. Assume that r is an affine transformation:
$$r(x) = A\,x + b, \qquad x \in \mathbb{R}^n,$$
for some fixed vector $b \in \mathbb{R}^n$ and invertible matrix $A \in \mathbb{R}^{n \times n}$. Then the Jacobian matrix of r is given by
$$J^r(x) = A, \qquad x \in \mathbb{R}^n$$

and we have

$$r^{-1}(y) = A^{-1}(y-b), \qquad y \in \mathbb{R}^n$$

therefore (4.2) yields

$$f_Y(y) = \frac{1}{|\det(A)|} f_X(A^{-1}(y-b)), \qquad y \in \mathbb{R}^n.$$

We sum up the different change of variable methods explained above in the following chart:

Case	Technique
discrete case	sum up PMF over preimage
continuous case, $r: \mathbb{R} \to \mathbb{R}$	3-steps to transformation
continuous case, $r: \mathbb{R} \to \mathbb{R}$ s.t. $r' \neq 0$	univariate Jacobian formula
continuous case, $r: \mathbb{R}^d \to \mathbb{R}^d$ s.t. $\det(J^r) \neq 0$	multivariate Jacobian formula

4.3 Sums of independent random variables

.

Let X, Y be two **independent** random variables. How does one compute the probability distribution of $X+Y$.

4.3.1 Discrete case

Assume that X and Y are **independent**, discrete, real-valued random variables with PMFs p_X, p_Y, respectively.

Theorem 4.10. *The random variable $Z = X+Y$ has PMF given by*

$$p_Z(z) = \sum_{\substack{x,y \\ x+y=z}} p_X(x)\, p_Y(y) = \sum_x p_X(x)\, p_Y(z-x).$$

Example 4.11. Let $X \sim \mathcal{P}(\lambda)$ and $Y \sim \mathcal{P}(\mu)$ be two independent Poisson random variables with parameters $\lambda, \mu > 0$. Then $X+Y \sim \mathcal{P}(\lambda+\mu)$.

4.3.2 Continuous case

Theorem 4.12. *The random variable $Z = X+Y$ has PDF given by*

$$p_Z(z) = \int_{-\infty}^{+\infty} f_X(x)\, f_Y(z-x)\, dx, \qquad z \in \mathbb{R}. \tag{4.3}$$

Remark 4.13. * The integral appearing in (4.3) is often called convolution product of f_X and f_Y, and sometimes written $f_X * f_Y$.

4.4 Transformations of Gaussian random variables

In this section we present results that show how Gaussian random variables (and vectors) behave under certain specific transformations.

Proposition 4.14. *[Z-transform]The (1-dimensional) Gaussian distribution satisfies the following properties:*

1. *Let $\mu \in \mathbb{R}$ and $\sigma > 0$. If $X \sim \mathcal{N}(0,1)$, then $Y := \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$.*
2. *Conversely, if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $X := \frac{1}{\sigma}(Y - \mu) \sim \mathcal{N}(0,1)$.*

Proof. 1. Let

$$r(x) = \sigma x + \mu, \qquad x \in \mathbb{R}.$$

Then, in virtue of the Jacobian change of formula(4.1), $Y = r(X)$ admits the PDF

$$f_Y(y) = \frac{1}{\sigma} f_X\left(\frac{1}{\sigma}(y-\mu)\right).$$

But by assumption $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$, so the expression above yields

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right), \qquad y \in \mathbb{R}.$$

We recognise the PDF of a Gaussian with mean μ and σ^2: the claim follows.

2. It suffices to apply (4.1), now to the function

$$r^{-1}(y) = \sigma^{-1}(y - \mu), \qquad y \in \mathbb{R}. \qquad \square$$

Proposition 4.15. *If $X_1, \dots, X_n$ are independent Gaussian variables with parameters μ_i and σ_i^2, and if $X := \sum_{i=1}^n X_i$, then*

$$X \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

Proof. By induction it is sufficient to treat the case $n = 2$. In turn the formula for $n = 2$ can be proven using (4.3). We omit the details. $\square$

The above results generalise to the multivariate case. We first state:

Lemma 4.16. *Let $d \geq 1$, let $\mu \in \mathbb{R}^d$, let $K \in \mathbb{R}^{d \times d}$ be a symmetric, strictly positive definite matrix, and consider a d-dimensional Gaussian vector $X \sim \mathcal{N}(\mu, K)$. If $p \leq d$ and $A \in \mathbb{R}^{p \times d}$ is a $p \times d$ matrix with full rank, then $Y := AX$ is a p-dimesnional Gaussian vector, namely $Y \sim \mathcal{N}(A\mu, AKA^T)$.*

Remark 4.17. In the above, we can for example take $p = 1$ and, for all $i = 1, \dots, d$, we can choose $A_i \in \mathbb{R}^{1\times d} = \mathbb{R}^d$ to be the i-th vector of the canonical basis of $\mathbb{R}^d$. Then $Y = X_i$ and

$$A_i\,\mu = \mu_i, \qquad A\,K\,A^T = K_{i,i},$$

we thus retrieve the result on the marginal distribution of X_i.

We now state a multivariate version of the Z-transform. Recall that if $d \geq 1$ and $K \in \mathbb{R}^{d\times}$ is symmetric, strictly positive definite, then K admits a square-root, i.e. there exists a symmetric, strictly positive matrix A, sometimes denoted by $K^{1/2}$, such that $A^2 = K$.

Proposition 4.18. *Let $d \geq 1$. The multivariate d-dimensional Gaussian distribution satisfies the following properties:*

1. *Let $\mu \in \mathbb{R}^d$, let $K \in \mathbb{R}^{d\times d}$ be a symmetric, strictly positive definite matrix, and let A be a square-root of K. If $X \sim \mathcal{N}(0, I_d)$, then $Y := \mu + A\,X \sim \mathcal{N}(\mu, K)$.*
2. *Conversely, if $Y \sim \mathcal{N}(\mu, K)$, then $X := A^{-1}\,(Y - \mu) \sim \mathcal{N}(0, I_d)$.*

Proposition 4.19. *If $X_1, \dots, X_n$ are independent Gaussian vectors with parameters μ_i and K_i^2, and if $X := \sum_{i=1}^n X_i$, then*

$$X \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n K_i\right).$$

We end this chapter with a result on squares of Gaussian random variables.

Definition 4.20. *For $n \geq 1$, the $\chi^2(n)$ distribution is the continuous distribution with PDF*

$$\chi^2(x; n) = \mathbf{1}_{\{x>0\}}\,\frac{1}{\Gamma\left(\frac{n}{2}\right)\,2}$$

4.5 Exercises

Exercise 4.2. [The Box-Muller transform] Let $R \sim \mathrm{Exp}(1/2)$ and $\theta \sim \mathrm{U}([0, 2\,\pi])$. We assume that R and θ are independent. Show that $X := R\cos(\theta)$ and $Y := R\sin(\theta)$ are two i.i.d. standard normal random variables.

Chapter 5
Expectation and covariance

This Chapter follows Chapter 3 of Wasserman.

5.1 Expectation: Definition and examples

We wish to give a proper meaning to the notion of "mean" or "average" of a real-valued random variable X. If the random variable takes finitely many values $x_1, \dots, x_n$ with equal probability $\frac{1}{n}$, then it is natural to define the average of X as the arithmetic average $\frac{1}{n}\sum_{i=1}^n x_i$. More generally, if X takes the value x_i with some probability p_i for all i, then it is natural to define the average of X, as being the weighted average $\sum_{i=1}^n p_i\, x_i$. With this definition, values x_i that are more likely to be realised are assigned a stronger weight p_i, while x_i that are less likely to occur are assigned a smaller weight.

Definition 5.1. *Let X be a discrete, resp. continuous, real-valued random variable with PMF p_X, resp. with PDF f_X. We define the* ***expectation*** *(also called mean) of X as*

$$\mathbb{E}(X) = \int x \; d\, F_X(x) = \begin{cases} \sum_x x\, p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} x\, f_X(x)\; d\, x & \text{if } X \text{ is continuous}, \end{cases} \tag{5.1}$$

this quantity being well-defined provided the sum, resp. the integral above is absolutely convergent: we then say that X is ***integrable****.*

Remark 5.2. The integrability condition ensures that the sum $\sum_x x\, p_X(x)$, resp. the integral $\int_{-\infty}^{+\infty} x\, f_X(x)\; d\, x$, is well-defined. Note that this condition is automatically fulfilled in the case of a discrete r.v. with finite outcome space. Most real-valued r.v's we will encounter in the sequel will be integrable. However there does exist random variables that are not integrable, see Exercise 5.2 below.

The expectation of X can be interpeted as the value that X will take on average. Heuristically, if we observe realisations $x_1, \dots, x_n$ of X, then, for large n, the empirical mean should be close to $\mathbb{E}(X)$:

$$\frac{1}{n}\sum_{i=1}^n X_i \approx \mathbb{E}(X).$$

The Law of Large Numbers, to be stated in the next Chapter, provides a rigorous mathematical proof for this fact.

We now give examples of computations of $\mathbb{E}(X)$.

Example 5.3. Assume that X is deterministic, i.e. there exists $x \in \mathbb{R}$ such that $X = x$ a.s. Then $\mathbb{E}(X) = x$.

Example 5.4. Let X be a discrete r.v. whose target space is a **finite** set $\mathcal{X} \subset \mathbb{R}$. Then X is always integrable, and

$$\mathbb{E}(X) = \sum_{x \in \mathcal{X}} x \; p_X(x).$$

In particular, when X is uniformly distributed in $\mathcal{X}$, we have

$$\mathbb{E}(X) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} x,$$

so the expectation of X coincides with the algebraic average of the values $x \in \mathcal{X}$.

Example 5.5. Let $a < b$ and assume that $X \sim \mathcal{U}(a, b)$. Then

$$\mathbb{E}(X) = \int x \, f_X(x) \; d\,x = \int_a^b x \; \frac{1}{b-a} \, d\,x = \frac{b^2 - a^2}{2\,(b-a)} = \frac{a+b}{2}.$$

Example 5.6. Let $\mu \in \mathbb{R}$ and $\sigma > 0$, and consider $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x \; \frac{\exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right)}{\sqrt{2\,\pi\,\sigma^2}} \; d\,x.$$

Performing the change of variable $y = x - \mu$, the above integral can be rewritten

$$\int_{-\infty}^{+\infty} (y \; + \; \mu) \; \frac{\exp\left(-\frac{y^2}{2\,\sigma^2}\right)}{\sqrt{2\,\pi\,\sigma^2}} \; d \; y \; = \; \int_{-\infty}^{+\infty} \underbrace{y \, \frac{\exp\left(-\frac{y^2}{2\,\sigma^2}\right)}{\sqrt{2\,\pi\,\sigma^2}}}_{\text{odd function of } y} d \; y \; + \; \mu \int_{-\infty}^{+\infty} \underbrace{\frac{\exp\left(-\frac{y^2}{2\,\sigma^2}\right)}{\sqrt{2\,\pi\,\sigma^2}}}_{\mathcal{N}(y;0,\sigma^2)} d\,y$$

The first integral in the right-hand side vanishes (as the integrand is odd), while the integral equals 1 (as the integrand is the PDF of a Gaussian, hence normalised). So $\mathbb{E}(X) = 0 + \mu.\,1 = \mu$. This justifies calling the parameter μ the **mean** of the Gaussian r.v. X.

The following result unveils a tight link between expectation and probability:

Proposition 5.7. *Let A be a an event from some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and consider the random variable*

$$1_A : \omega \mapsto \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A. \end{cases}$$

It has expectation $\mathbb{E}(1_A) = \mathbb{P}(A)$.

Proof. The random variable $X = 1_A$ takes values in $\{0, 1\}$, its PMF satisfies

$$p_X(0) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A), \qquad p_X(1) = \mathbb{P}(A).$$

Therefore

$$\mathbb{E}(X) = 0 \; p_X(0) + 1 \; p_X(1) = \mathbb{P}(A). \qquad \square$$

5.1.1 Properties of expectation

We first ask the following question: if X is a, potentially multivariate, random variable, and r is a map, how to compute the expectation of $r(X)$?

Theorem 5.8. *[Rule of the lazy statistician (RLS)] Let X be a r.v. with outcome space E and $r : E \to \mathbb{R}$ a map. Then*

$$\mathbb{E}(r(X)) = \begin{cases} \sum_x r(x) \, p_X(x) & \text{if } X \text{ is discrete} \\ \int r(x) \, f_X(x) \, dx & \text{if } X \text{ is continuous}, \end{cases}$$

Thus, the knowledge of the distribution of X is sufficient in order to compute $\mathbb{E}(r(x))$ for any function $r : \mathbb{R} \to \mathbb{R}$, we do not need to re-compute the distribution of $r(X)$, hence the (informal) name of the above theorem.

Example 5.9. [Wasserman, Example 3.8] We break a stick of unit length at a random point distributed uniformly along the stick: the stick is thereby broken into two pieces, and we denote by Y the length of the larger piece. Let us compute $\mathbb{E}(Y)$. By assumption, the position of the breaking point is given by $X \sim \mathcal{U}(0, 1)$, and Y is then given by $Y = \max(X, 1 - X)$. By the rule of the lazy statistician, we thus get

$$\begin{aligned} \mathbb{E}(Y) = \int_{-\infty}^{+\infty} \max(x, 1-x) \; f_X(x) \, dx = \int_0^1 \max(x, 1-x) \; dx = \int_0^{1/2} (1 - x) \, dx + \int_{1/2}^1 x \, dx \\ = \left[\frac{-(1-x)^2}{2} \right]_0^{1/2} + \left[\frac{x^2}{2} \right]_{1/2}^1 = \\ \frac{1 - 1/4}{2} + \frac{1 - 1/4}{2} = \frac{3}{4}. \end{aligned}$$

Example 5.10. An important class of examples of maps r is given by

$$r(x) = |x|^k, \qquad x \in \mathbb{R},$$

for $k \geq 1$. The expectation

$$\mathbb{E}(|X|^k) \begin{cases} \sum_x |x|^k p_X(x) & \text{if X is discrete} \\ \int_{-\infty}^{+\infty} |x|^k f_X(x)\, dx & \text{if X is continuous,} \end{cases}$$

is called k-th moment of X. It is a well-defined, finite number provided the sum, resp. the integral above is convergent. We then say that X admits k moments. When $k = 2$, we say X is **square-integrable**.

Example 5.11. Let $a < b$ and assume that $X \sim \mathcal{U}(a, b)$. Then

$$\mathbb{E}(X^2) = \int x^2 f_X(x)\, dx = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}.$$

In particular, when $a = 0$ and $b = 1$, we get $\mathbb{E}(X^2) = 1/3$.

We now state further fundamental properties of expectations:

Proposition 5.12. *[Linear-affine transformation] Let X be a real-valued random variable, and $a, b \in \mathbb{R}$. Then*

$$\mathbb{E}(aX + b) = a\,\mathbb{E}(X) + b.$$

Proof. We assume the X is discrete, the continuous case being similar. By the RLS

$$\mathbb{E}(aX + b) = \sum_x (ax + b)\, p_X(x) = a \underbrace{\sum_x x\, p_X(x)}_{=\mathbb{E}(X)} + b \underbrace{\sum_x p_X(x)}_{=1}. \qquad \square$$

Theorem 5.13.

1. *If $X \geq 0$ (resp. $X \leq 0$) almost-surely, then $\mathbb{E}(X) \geq 0$ (resp. $\mathbb{E}(X) \leq 0$).*
2. *(Linearity) If $X_1, \ldots, X_n$ are real-valued random variables, and $\alpha_1, \ldots \alpha_n \in \mathbb{R}$, then*

$$\mathbb{E}\left(\sum_{i=1}^n \alpha_i X_i \right) = \sum_{i=1}^n \alpha_i\, \mathbb{E}(X_i).$$

3. *(Monotonicity) If $X \leq Y$ almost-surely, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.*

Proof.

1. This follows from (5.1).
2. Admitted.
3. Since $Y - X \geq 0$ a.s., we deduce from the first point above that $\mathbb{E}(Y - X) \geq 0$, which yileds the claim by linearity of $\mathbb{E}$. $\square$

Finally we state a result pertaining to the product of **independent** random variables.

Theorem 5.14. *Let $X_1, \ldots, X_n$ be **independent** real-valued random variables. Then*

$$\mathbb{E}[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} \mathbb{E}(X_i).$$

5.2 Variance: definition and examples

Definition 5.15. *Let X be a real-valued r.v. with mean μ. We define the **variance** of X, denoted by σ_X^2 or $\mathbb{V}(X)$, as*

$$\mathbb{V}(X) := \mathbb{E}((X-\mu)^2),$$

*this quantity being well-defined provided the random variable X is **square-integrable**. The **standard deviation** (or spread) of X, denoted by σ_X or $sd(X)$, is defined as*

$$sd(X) := \sqrt{\mathbb{V}(X)}.$$

Note that if X is discrete with PMF p_X, resp. continuous with PDF f_X, then

$$\mathbb{V}(X) = \begin{cases} \sum_x (x-\mu)^2 \; p_X(x) & \text{if X is discrete} \\ \int_{-\infty}^{+\infty} (x-\mu)^2 \; f_X(x) \; d\,x & \text{if X is continuous,} \end{cases}$$

Remark 5.16. Note that $\mathbb{V}(X)$ is always **non-negative**. While $\mathbb{E}(X)$ represents the average value of X, $\mathbb{V}(X)$ quantifies how far realisations of X can spread away from this average value.

Theorem 5.17. *[Variance translation theorem] The variance admits the following alternative expression:*

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

Proof. $\mathbb{V}(X) = \mathbb{E}\left(X^2 - 2\,\mu\,X + \mu^2\right) = \mathbb{E}(X^2) - 2\underbrace{\mu\,\mathbb{E}(X)}_{=\mu^2} + \mu^2 = \mathbb{E}(X^2) - \mu^2$. □

Example 5.18. Let $a < b$ and assume that $X \sim \mathcal{U}(a,b)$. We have shown above that

$$\mathbb{E}(X) = \frac{a+b}{2}, \qquad \mathbb{E}(X^2) = \frac{a^2 + a\,b + b^2}{3}.$$

Therefore

$$\mathbb{V}(X) = \frac{a^2 + a\,b + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12},$$

and $\mathrm{sd}(X) = \frac{b-a}{2\sqrt{3}}$. Thus, the larger the interval $[a,b]$ we are considering, the larger the standard deviation. In the special case, when $a=0$ and $b=1$, we get $\mathbb{V}(X) = 1/12$ and $\mathrm{sd}(X) = \frac{1}{2\sqrt{3}}$.

Example 5.19. Let $\mu \in \mathbb{R}$ and $\sigma > 0$, and consider $X \sim \mathcal{N}(\mu, \sigma^2)$. Recall that $\mathbb{E}(X) = \mu$. We now compute $\mathbb{V}(X)$:

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 \, \frac{\exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right)}{\sqrt{2\,\pi\,\sigma^2}} \, d\,x$$

Performing the change of variable $y = \frac{x-\mu}{\sigma}$, so that $d\,x = \sigma\; d\,y$, the above integral can be rewritten

$$\sigma^2 \int_{-\infty}^{+\infty} y^2 \, \frac{\exp\left(-\frac{y^2}{2}\right)}{\sqrt{2\,\pi}} \, d\,y.$$

Since the integrand is even the integral equals

$$2 \int_{0}^{+\infty} y^2 \, \frac{\exp\left(-\frac{y^2}{2}\right)}{\sqrt{2\,\pi}} \, d\,y$$

In turn, after the change of variable $z = \frac{y^2}{2}$, so that $d\,y = \frac{1}{\sqrt{2\,z}}\, d\,z$, we can rewrite that quantity as

$$2 \int_{0}^{+\infty} z^{1/2} \, \frac{\exp(-z)}{\sqrt{\pi}} \, d\,z = \frac{2\,\Gamma(3/2)}{\sqrt{\pi}},$$

where, for all $a > 0$, $\Gamma(a) := \int_{-\infty}^{+\infty} t^{a-1}\, e^{-t}\, d\,t$. But it is a fact that $\Gamma(3/2) = (1/2)\,\Gamma(1/2) = \frac{\sqrt{\pi}}{2}$. We thus obtain $\mathbb{V}(X) = \sigma^2$. This justifies calling the parameter σ^2 the variance of the Gaussian random variable X.

Theorem 5.20. *1. If X is a real-valued r.v. and $\alpha, \beta \in \mathbb{R}$, then*

$$\mathbb{V}(a\,X + b) = a^2\; \mathbb{V}(X).$$

*2. If $X_1, \ldots, X_n$ are n **independent** real-valued r.v.'s, and $a_1, \ldots, a_n \in \mathbb{R}$, then*

$$\mathbb{V}\left(\sum_{i=1}^{n} a_i\, X_i\right) = \sum_{i=1}^{n} a_i^2\; \mathbb{V}(X_i).$$

5.3 Covariance and correlation

Definition 5.21. *[See Definition 3.18, Wasserman] Let X, Y be two real-valued random variables admitting means μ_X and μ_Y and variances σ_X^2 and σ_Y^2. Then we define the* ***covariance*** *of X and Y as*

$$Cov(X,Y) := \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right).$$

We define the ***correlation*** *of X and Y as*

$$\rho_{X,Y} := \frac{Cov(X,Y)}{\sigma_X\ \sigma_Y}.$$

Remark 5.22. Using the Cauchy-Schwarz inequality (see p.286 in [3]), that the correlation satisfies

$$-1 \le \rho_{X,Y} \le 1, \tag{5.2}$$

this we call the *correlation inequality.*

Theorem 5.23. *We can rewrite the covariance of X and Y as*

$$Cov(X,Y) = \mathbb{E}(XY) - \mu_X\ \mu_Y.$$

When $\mathrm{Cov}(X,Y)=0$, or equivalently $\rho(X,Y)=0$ we say that X and Y are **uncorrelated**

Theorem 5.24. *If X and Y are independent, then X and Y are uncorrelated.*

Proof. Assume $X \perp\!\!\!\perp Y$. Then

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mu_X\ \mu_Y,$$

bu the first term in the right-hand side equals $\mu_X\ \mu_Y$ in virtue of (5.14), and the claim follows. □

Remark 5.25. BEWARE!!! The converse implication is not true in general.

Theorem 5.26. $\mathbb{V}(X+Y) = \mathbb{V}(X) + 2\,Cov(X,Y) + \mathbb{V}(Y)$

$$\mathbb{V}(X-Y) = \mathbb{V}(X) - 2\,Cov(X,Y) + \mathbb{V}(Y).$$

More gnerally,

Definition 5.27. *Let $d \ge 1$. Let X be a d-dimensional random vector. We define the mean $\mu \in \mathbb{R}^d$ and the covariance matrix $K \in \mathbb{R}^{d\times d}$ of X by*

$$\mu_i := \mathbb{E}(X_i), \qquad i = 1,\dots,d$$

and

$$K_{i,j} := Cov(X_i, X_j), \qquad 1 \le i,j \le d.$$

Example 5.28. Assume X is a d-dimensional Gaussian vector: $X \sim \mathcal{N}(\mu, K)$, where $\mu \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d\times d}$ is a strictly definite positive, symmetric matrix. Then the mean vector of X is given by μ, and its covariance matrix by K.

Corollary 5.29. *Let $X = (X_1, \ldots, X_d)$ as in the above example. Then $X_1, \ldots, X_d$ are independent if and only if the matrix K is diagonal, i.e. $K_{i,j} = 0$ for all $i \neq j$.*

Proof. We have already seen, in the previous chapter, that if K is diagonal, then $X_1, \ldots, X_D$. The converse follows from the fact that, if $X_1, \ldots, X_n$ are independent, then for all $i \neq j$, $X_i \perp\!\!\!\perp X_j$, so

$$K_{i,j} = \mathrm{Cov}(X_i, X_j) = 0. \qquad \square$$

5.4 Sample mean and sample variance

In practice we do not observe random variables: we observe realisations, or a sample, thereof. It is therefore useful to define notions of sample mean and sample variance, which are quantities that can be computed directly from the realisations at hand.

Let $X_1, \ldots, X_n$ be n real-valued random variables: anticipating on the chapter on frequentist inference, one may think of $X_1, \ldots, X_n$ as representing a sample from some random variable.

Definition 5.30. *The **sample mean** of $X_1, \ldots, X_n$ is defined as the arithmetic average:*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

*The **sample variance** of $X_1, \ldots, X_n$ is defined as*

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \tag{5.3}$$

*and the **sample standard deviation** by*

$$S_n := \sqrt{S_n^2}.$$

Remark 5.31. Note that the sample mean $\bar{X}_n$ and the sample variance S_n^2 are **random variables**. However, as we will see in the next Chapter, if $X_1, \ldots, X_n$ are i.i.d, provided some integrability assumptions are satisfied, we have for large n,

$$\bar{X}_n \approx \mathbb{E}(X_1), \qquad S_n^2 \approx \mathbb{V}(X_1).$$

Likewise, if $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n bidimensional random vectors, we can define their sample covariance and sample correlation.

Exercise 5.1. Write down the definition of the sample covariance C_n and the sample correlation R_n of $(X_1, Y_1), \ldots, (X_n, Y_n)$.

5.5 Exercises

Exercise 5.2. [Cauchy random variable]We consider the function $f: \mathbb{R} \to \mathbb{R}_+$ given by

$$f(x) = \frac{1}{\pi\,(1+x^2)}, \qquad x \in \mathbb{R}.$$

1. Show that f is a PDF.
2. Let X be a r.v. with PDF f. We say that X is a Cauchy random variable. Show that X is not integrable.

Chapter 6
Inequalities and limits

In this Chapter we first state important inequalities: one of themain aims is to provide bounds on the probabilities that a random variable X stays away from its mean by a certain distance $t>0$:

$$\mathbb{P}\left(|X-\mathbb{E}(X)|>t\right), \qquad t>0$$

In the second part we address two results that are cornerstones of statistical inference: the Law of Large Numbers (LLN) and Central Limit Theorem (CLT). Both of these are limit theorems concerning the empirical sum of a sequence of i.i.d. random variables $(X_i)_{i\geq 0}$. The first statement will tell us that the empirical mean converges to the expectation of the random variables

$$\bar{X}_n \underset{n\to\infty}{\longrightarrow} \mathbb{E}(X_1)$$

in a sense to be specified. The second statement will tell us tha this convergence holds with speed $1/\sqrt{n}$, in a sense to be specified below. At the end we will be able to obtain *approximate* bounds on the probability that the empirical sum remains away from its mean

$$\mathbb{P}\left(|\bar{X}-\mathbb{E}(\bar{X})|>t/\sqrt{n}\right), \qquad t>0$$

for n large.

6.1 Inequalities

6.1.1 Inequalities for expectations*

We start by stating an inequality

Theorem 6.1. **[Cauchy-Schwarz inequality] Let X and Y be two square-inetgrable real-valued random variables (a condition which, as we recall, ensures that the second moments $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ are well-defined, finite numbers). Then*

$$\mathbb{E}\left(XY\right) \leq \sqrt{\mathbb{E}(X^2)}\ \sqrt{\mathbb{E}(Y^2)}$$

6.1.2 Bounds on probabilities

Let X be a real-valued random variable. In statistics, it is often crucial to be able to bound from above the probability $\mathbb{P}(X > t)$, for $t > 0$. Clearly, if t is very large, $\mathbb{P}(X > t)$ should be small, but the problem is to make this statement quantitative. Here is a very general statement

Theorem 6.2. *[Markov's inequality] If X is an integrable, non-negative real-valued random variable and $t > 0$, then*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$

Proof. We assume that X is a continuous r.v., the discrete case is similar. We have

$$\mathbb{P}(X > t) = \int_t^{+\infty} f_X(x)\,\mathrm{d}x.$$

Now, for $x \geq t$, we have $1 \leqslant \frac{x}{t}$, so the above integral is bounded from above by

$$\int_t^{+\infty} \frac{x}{t} f_X(x)\,\mathrm{d}x = \frac{1}{t}\int_t^{+\infty} x\, f_X(x)\,\mathrm{d}x \leq \frac{1}{t}\int_0^{+\infty} x\, f_X(x)\,\mathrm{d}x.$$

Note now that $f_X(x) = 0$ for $x < 0$ since $X \geq 0$, whence

$$\int_0^{+\infty} x\, f_X(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x\, f_X(x)\,\mathrm{d}x = \mathbb{E}[X],$$

whence the claim. □

Assume now that X is a square-integrable real-valued random variable. The average value of X is given by $\mathbb{E}(X)$, and we can bound from above the probability that X drifts away from its average by a certain distance $t > 0$:

Theorem 6.3. *[Chebyshev's inequality] For all $t > 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| > t) \leq \frac{\mathbb{V}(X)}{t^2}. \tag{6.1}$$

Proof. By Markov's inequality, we have

$$\mathbb{P}(|X - \mathbb{E}(X)| > t) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^2)}{t^2} = \frac{\mathbb{V}(X)}{t^2}.$$

□

This confirms the interpretation of $\mathbb{V}(X)$ as a description of the amount by which X may deviate from its mean.

The Chebyshev inequality is very useful in situations where we have no priori information on the distribution of X, except from its mean and variance. On the other hand, it is quite a rough bound. When we do know the distribution of X, the probability $\mathbb{P}(|X-\mathbb{E}(X)|>t)$ can be computed, allowing thus to obtain a - often much better - bound. Let us consider for instance the case where X is a Gaussian random variable.

Example 6.4. Assume that $X\sim\mathcal{N}(0,1)$. Then for all $t>0$,

$$\mathbb{P}(|X|>t)=\int_{-\infty}^{-t}\frac{1}{\sqrt{2\,\pi}}\exp\left(-\frac{s^2}{2}\right)\,d\,s+\int_{t}^{+\infty}\frac{1}{\sqrt{2\,\pi}}\exp\left(-\frac{s^2}{2}\right)\,d\,s=$$
$$2\int_{t}^{+\infty}\frac{1}{\sqrt{2\,\pi}}\exp\left(-\frac{s^2}{2}\right)d\,s. \tag{6.2}$$

Although there is no closed expression for the above integral, one can bound it from above, using the following theorem

Theorem 6.5. *[Mill's inequality] If $X\sim\mathcal{N}(0,1)$, then for all $t>0$,*

$$\mathbb{P}(|X|>t)\leq\sqrt{\frac{2}{\pi}}\,\frac{\exp\left(-\frac{t^2}{2}\right)}{t}.$$

Proof. We bound from above the last integral in (6.2) above as follows

$$\int_{t}^{+\infty}\frac{1}{\sqrt{2\,\pi}}\,\exp\left(-\frac{s^2}{2}\right)\,d\,s\leq\int_{t}^{+\infty}\frac{s}{t}\,\frac{1}{\sqrt{2\,\pi}}\,\exp\left(-\frac{s^2}{2}\right)\,d\,s=$$
$$\frac{1}{t\,\sqrt{2\,\pi}}\int_{t}^{+\infty}s\exp\left(-\frac{s^2}{2}\right)d\,s.$$

For the last integral we have

$$\int_{t}^{+\infty}s\exp\left(-\frac{s^2}{2}\right)d\,s=\left[\exp\left(-\frac{s^2}{2}\right)\right]_{t}^{+\infty}=\exp\left(-\frac{t^2}{2}\right),$$

and the claim follows. □

The above result generalises to centered Gaussians with arbitrary covariance

Corollary 6.6. *Let $\mu\in\mathbb{R}$, $\sigma>0$ and let $X\sim\mathcal{N}(0,\sigma^2)$. Then for all $t>0$,*

$$\mathbb{P}(|X-\mathbb{E}(X)|>t)\leq\sqrt{\frac{2\,\sigma^2}{\pi}}\,\frac{\exp\left(-\frac{t^2}{2\,\sigma^2}\right)}{t}.$$

Proof. By the Z-transform, the random variable $Y:=\frac{1}{\sigma}(X-\mu)$ is a standard normal variable, so

$$\mathbb{P}(|X-\mu|>t)=\mathbb{P}(|Y|>\sigma^{-1}\,t),$$

and the result follows by Mill's inequality with t replaced by $\sigma^{-1}t$. □

6.2 Limit Theorems

We will now state two fundamental limit theorems for sums of i.i.d. random variables. To do so, we first need to clarify what we mean by convergence of a sequence of random variables.

6.2.1 Different notions of limits

Let X be a real-valued random variable, and $(X_n)_{n\geq 0}$ a sequence of real-valued random variables.

Definition 6.7. *We say that X_n converges to X in probability, and write $X_n \xrightarrow[n\to\infty]{(P)} X$, if, for any $\epsilon > 0$, there holds*

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

In other words, X_n converges to X in probability if the probability that X_n remains away from X by any (even very small) distance vanishes as n grows.

We will also encounter another, weaker form of convergence. Its formulation involves the CDFs F_{X_n} and F_X of the r.v.'s X_n and X, respectively:

Definition 6.8. *We say that X_n converges to X in law, or in distribution, and write $X_n \xrightarrow[n\to\infty]{(d)} X$, if, for any $x \in \mathbb{R}$ where F_X is continuous, there holds*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Remark 6.9. If X is a continuous random variable, then F_X is everywhere continuous, and the above condition may be restated as point-wise convergence of F_{X_n} to F_X. On the other hand, when X is dicrete, F_X will be discontinuous at every point x such that $\mathbb{P}(X = x) > 0$. The above definition says that, when checking whether X_n converges in distribution to X, we do not need look at these points of discontinuities.

That X_n converges in distribution to X thus means that

$$\mathbb{P}\left(X_n \leq x\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(X \leq x\right),$$

for all point x where F_X does not jump. It is only a statement about the probability distributions of X_n and X. In particular, it does not say at all that X_n is close to X when n is large.

Proposition 6.10.

1. *If X_n and X are square-integrable, and*
$$\mathbb{E}((X_n - X)^2) \underset{n\to\infty}{\longrightarrow} 0, \tag{6.3}$$
then $X_n \underset{n\to\infty}{\xrightarrow{(P)}} X$. The converse is false in general.
2. *If X_n converges to X in probability, then X_n also converges to X in law. The converse is false in general.*
3. *If X is constant, i.e. there exists some constant $a \in \mathbb{R}$ such that $X = a$ a.s., then*
$$X_n \underset{n\to\infty}{\xrightarrow{(P)}} X \iff X_n \underset{n\to\infty}{\xrightarrow{(d)}} X$$

The convergence (6.3) is called "convergence in quadratic mean", and written $X_n \underset{n\to\infty}{\xrightarrow{q.m.}} X$. By the above Proposition, convergence in quadratic mean is strictly stronger than convergence in probability, which itself is strictly stronger than convergence in distribution.

Proof. We prove the first claim. Assume that $X_n \underset{n\to\infty}{\xrightarrow{q.m.}} X$. Then, for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|X_n - X|^2 > \epsilon^2) \leq \frac{\mathbb{E}(|X_n - X|^2)}{\epsilon^2},$$

where the last inequality follows by Markov's inequality. By assumption, $\mathbb{E}(|X_n - X|^2) \underset{n\to\infty}{\longrightarrow} 0$. Hence, by the above inequality , we get $\mathbb{P}(|X_n - X| > \epsilon) \underset{n\to\infty}{\longrightarrow} 0$. This proves that $X_n \underset{n\to\infty}{\xrightarrow{P}} X$. That the converse implication is false in general is left as an exercise. □

6.2.2 The Law of Large Numbers (LLN)

Before stating the LLN, we need to state a technical, but intuitive, lemma.

Lemma 6.11. *Let X, Y be real-valued random variables such that $X \overset{(d)}{=} Y$. Then, for any real-valued map such that $f(X)$ is integrable, we have $\mathbb{E}(f(X)) = \mathbb{E}(f(Y))$.*

Proof. Assume that X and Y are both discrete. Then $p_X = p_Y$, so for all function f as in the statement, by the RLS, we have

$$\mathbb{E}(f(X)) = \sum_x p_X(x)\ f(x) = \sum_y p_Y(y)\ f(y) = \mathbb{E}(f(Y)).$$

The case where X, Y are continuous is proven in the same way, just replacing PMF by PDF and sums by integrals. □

The above result implies in particular that, if $X \overset{(d)}{=} Y$, we have

$$\mathbb{E}(X) = \mathbb{E}(Y), \qquad \mathbb{E}(X^2) = \mathbb{E}(Y^2), \qquad \mathbb{V}(X) = \mathbb{V}(Y),$$

provided these quantities are well-defined.

Let now $(X_n)_{n\geq 1}$ be a sequence of i.i.d. copies of a real-valued random variable X. By this we mean $(X_n)_{n\geq 1}$ is a sequence of i.i.d. real-valued random variable having the same law as X. For all $n\geq 1$, let $\bar{X}_n$ denote the sample mean of $X_1,\dots,X_n$:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

If the X_n are integrable, then by linearity of the expectation we have

$$\mathbb{E}(\bar{X}_n) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \mathbb{E}(X).$$

Heuristically, we actually expect $\bar{X}_n$ to converge to $\mathbb{E}(X)$ when $n\to\infty$. This heuristics is made rigorous by the following theorem.

Theorem 6.12. *[Weak Law of Large Numbers] If the X_n are integrable, then*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{(P)} \mathbb{E}(X).$$

Proof. For simplicity, we provide a proof in the special case where the X_n are also square-integrable. We then have

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

Now

$$\mathbb{V}(\bar{X}_n) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \frac{1}{n^2}\,\mathbb{V}(X_i),$$

where the second equality holds because the X_i are independent. Now, by Lemma 6.11, $\mathbb{V}(X_i) = \mathbb{V}(X)$ for all i, so we get

$$\mathbb{V}(\bar{X}_n) = \frac{\mathbb{V}(X)}{n} \xrightarrow[n\to\infty]{} 0.$$

Hence $\mathbb{E}((\bar{X}_n - \mathbb{E}(X))^2) \xrightarrow[n\to\infty]{} 0$, therefore $\bar{X}_n$ converges to $\mathbb{E}(X)$ in quadratic mean, hence also in probability. □

Remark 6.13. * Above we stated the weak LLN. Actually a stronger statement, known as the strong LLN, holds with the same assumptions. It says that $\bar{X}_n \xrightarrow[n\to\infty]{(P)} \mathbb{E}(X)$ almost-surely, i.e. that

$$\mathbb{P}\Big(\{\omega\in\Omega\colon \bar{X}_n(\omega) \xrightarrow[n\to\infty]{} \mathbb{E}(X)\}\Big) = 1.$$

However we will not need this stronger version of the LLN in the sequel.

Remark 6.14. * The LLN is a cornerstone of the theory of probability. Its ramifications in statistics are considerable. In particular, it provides a concrete way of obtaining the value of $\mathbb{E}(X)$ by sampling values of X a large number of times and taking the sample mean.

Example 6.15. Let $X_1, \dots, X_n \sim X \overset{(d)}{=} \mathrm{Ber}(p)$ for some $p \in (0,1)$. Then

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{(P)} \mathbb{E}(X) = p.$$

In other words, when we keep throwing a coin with parameter p a large number of times, the rate of success will converge in probability to p. If the coin is fair, i.e. $p = 1/2$, the rate of success approaches $1/2$ for n large.

Example 6.16. Let $X_1, \dots, X_n \sim X \overset{(d)}{=} \mathcal{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$. Then

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{(P)} \mathbb{E}(X) = \mu.$$

Question: By the LLN, we thus have $\bar{X}_n = \mathbb{E}(X) + \epsilon_n$, where ϵ_n is some remainder satisfying $\epsilon_n \xrightarrow[n\to\infty]{(P)} 0$. Can we quantify how fast ϵ_n converges to 0?

6.2.3 The Central Limit Theorem

Let $X_1, X_2, \dots$ be a sequence of i.i.d. real-valued random variables. We assume that the X_i are square-integrable and denote respectively by μ and σ^2 their mean and variance. Thus, for all i,

$$\mathbb{E}(X_i) = \mu, \qquad \mathbb{V}(X_i) = \sigma^2.$$

As we saw in the previous section,

$$\mathbb{E}(\bar{X}_n) = \mu, \qquad \mathbb{V}(\bar{X}_n) = \frac{\sigma^2}{n} \tag{6.4}$$

We can perform an affine transformation on $\bar{X}_n$ in order to set its expectation and variance to 0 and 1, respectively. To do so:

1. we *center* it, by subtracting its mean $\mathbb{E}(\bar{X}_n)$
2. we *normalize* it, by dividing it by it s standard deviation $\sqrt{\mathbb{V}(\bar{X}_n)}$.

In other words, we set

$$Y_n := \frac{1}{\sqrt{\mathbb{V}(\bar{X}_n)}} \left(\bar{X}_n - \mathbb{E}(\bar{X}_n)\right) = \sqrt{\frac{n}{\sigma^2}} \left(\bar{X}_n - \mu\right).$$

With this procedure, we obtain a random variable Y_n which is centered and normalized, i.e. which satisfies

$$\mathbb{E}(Y_n) = 0, \qquad \mathbb{V}(Y_n) = 1.$$

Exercise 6.1. Verify that Y_n defined above is indeed centered and normalized, using the Z-transform (Prop 4.14).

The following theorem shows that, for n large, the distribution of Y_n is actually close to $\mathcal{N}(0,1)$.

Theorem 6.17. *[Central Limit Theorem] Let $X_1, X_2, \dots$ be a sequence of i.i.d. real-valued, square-integrable r.v. with mean μ and variance σ^2. Then*

$$\sqrt{\frac{n}{\sigma^2}}\,(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Remark 6.18. Thus, for all $a \in \mathbb{R}$,

$$\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}\,(\bar{X}_n - \mu) \le a\right) \xrightarrow[n\to\infty]{} \int_{-\infty}^{a} \frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\,\pi}}\,d\,x$$

Remark 6.19. One may loosely formulate the CLT as saying that

$$\sqrt{\frac{n}{\sigma^2}}\,(\bar{X}_n - \mu) \overset{(d)}{\approx} \mathcal{N}(0,1)$$

for n large. This equivalently means that

$$\bar{X}_n \overset{(d)}{\approx} \mu + \sqrt{\frac{\sigma^2}{n}}\,Z$$

where $Z \sim \mathcal{N}(0,1)$ i.e., in virtue of the Z-transform (Proposition 4.14,

$$\bar{X}_n \overset{(d)}{\approx} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Example 6.20. Let $X_1, \dots, X_n \sim X \overset{(d)}{=} \mathrm{Ber}(p)$ for some $p \in (0,1)$. We know from the LLN that

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{(P)} \mathbb{E}(X) = p.$$

Since $\mathbb{V}(X) = p\,(1-p)$, by the CLT, we further get

$$\sqrt{\frac{n}{p\,(1-p)}}\,(\bar{X}_n - p) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Exercise 6.2. We throw a coin of parameter $p = 0.3$ (i.e. we get Heads with prob. 0.3 at each toss). Let H denote the number of Heads we got after $n = 10^3$ tosses. Find an interval I such that

$$\mathbb{P}\,(H \in I) \approx 0.95.$$

Part II

Frequentist Inference

Chapter 7

Foundations and maximum likelihood

Assume we observe the realisations of n i.i.d random variables $X_1, \dots X_n$. Can we infer from our observations the common law of the X_i?

Very often we formulate some additional assumption on the observations $X_1, \dots, X_n$, namely that their common law belongs to a certain family of probability distributions, that we call a *model*.

7.1 Statistical model and estimator

7.1.1 Statistical model

Definition 7.1. *A statistical model $\mathcal{F}$ is a set of probability distributions.*

Example 7.2.
$$\mathcal{F} := \{\mathrm{Ber}(p) \colon p \in (0,1)\}$$

and

$$\mathcal{F}' = \{\mathcal{U}(a,b) \colon a, b \in \mathbb{R}, a < b\}$$

are both examples of statistical models. The first one consists of all Bernoulli distributions. The second one consists of all (continuous) uniform distributions on a finite interval.

Since the probability distribution of a discrete (resp. continuous) r.v. is uniquely determined by its PMF (resp. PDF), a model $\mathcal{F}$ will often be given as a set of PMFs (resp. PDFs).

Example 7.3. The model $\mathcal{F}$ in the above example may alternatively be denoted as

$$\mathcal{F} := \{\mathrm{Ber}(k; p) \colon p \in (0,1)\}$$

where we recall that $\mathrm{Ber}(k; p) = (1-p)^{1-k}\ p^k$, for $k = 0, 1$, is the PMF of the Bernoulli distribution with parameter p.

The model $\mathcal{F}'$ in the above example may alternatively be denoted as

$$\mathcal{F}' := \{\mathcal{U}(x; a, b) \colon a, b \in \mathbb{R}, a < b\}$$

where we recall that $\mathcal{U}(x;a,b) = \frac{1}{b-a}\,1_{[a,b]}(x)$ for $x \in \mathbb{R}$ is the PDF of the uniform distribution on the interval $[a,b]$.

A statistical model is said to be *parametric* if it can be parametrised by a finite number of parameters. Otherwise it is said to be non-parametric.

Example 7.4. $$\mathcal{G} := \{N(x;\mu,\sigma^2)\colon \mu \in \mathbb{R}, \sigma > 0\}$$

and

$$\mathcal{G}' = \{\text{all continuous probability distributions on } \mathbb{R}\}$$

are both examples of statistical models. The first one, consisting of all one-dimensional Gaussian distributions, is parametric: it is parametrized by $\mu \in \mathbb{R}$ and $\sigma > 0$. The second example is not parametric.

An important example of parametric model is the general linear model:

Definition 7.5. *The general linear model is defined as*

$$Y = X\beta + \epsilon,$$

where:

- $X \in \mathbb{R}^{n\times p}$ *is a fixed, deterministic matrix, called design matrix,*
- ϵ *is a* $n-dimensional$ *random vector with law* $N(0,\sigma^2 I_n)$
- $\beta \in \mathbb{R}^p$ *and* $\sigma^2 > 0$ *are unknown paramters.*

In view of the multivariate Z-transform, see the 1st statement of Proposition 4.18, the model above can be alternatively written as

$$\mathfrak{F} = \{N(y; X\beta, \sigma^2 I_n), \quad \beta \in \mathbb{R}^p, \quad \sigma^2 > 0\}. \tag{7.1}$$

7.1.2 General setting of statistical inference

We assume given a set Θ, called parameter space, and a parametric model

$$\mathfrak{F} = \{f(x;\theta), \quad \theta \in \Theta\},$$

consisting of continuous (resp. discrete) probability distributions represented by their PDF (resp. PMF). Let $\theta \in \Theta$ be a fixed, unknown parameter. We observe the realisations of n i.i.d random variables $X_1, \dots X_n$ with common distribution p_θ. From these observations, can one retrieve the value of the unknown parameter θ? There are three main ways one could use to answer this question:

1. Point estimation: we use our observed data $X_1, \dots, X_n$ to build a *point estimator* $\hat{\theta}_n$, with the hope that, for n large, $\hat{\theta}_n$ is close (in a certain sense) to the true, but unknown, value of the parameter θ.

2. Confidence interval: we use our observed data $X_1, \dots, X_n$ to build a *confidence interval* $[a_n, b_n]$ containing the true, but unkown parameter value θ with high probability.
3. Hypothesis testing: in that setting we partition the parameter space Θ into two subsets Θ_0 and Θ_1, and we just want to infer from our data whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, with a control on the error probability.

In this Chapter and the next two chapters we shall focus on point estimation. Confidence intervals and hypothesis testing will be addressed later/

7.2 Point estimator, maximum likelihood

7.2.1 Point estimator: definition

Definition 7.6. *Let $X_1, \dots, X_n$ be i.i.d random variables distributed according to p_θ. A point estimator $\hat{\theta}_n$ for θ is a random variable taking values in Θ, and of the form*

$$\hat{\theta}_n = \psi_n(X_1, \dots, X_n),$$

for some map $\psi_n : \mathbb{R}^n \to \Theta$.

Loosely speaking, $\hat{\theta}_n : \Omega \to \Theta$ is an estimator if it is a deterministic function of $X_1, \dots, X_n$. Often one encounters the abusive notation $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ to express this dependence. Note that, in the strict definition above, we do claim that $\hat{\theta}_n$ indeed does a good job at estimating θ in some sense, rather it is just an attempt at doing so. Whether or not it does indeed allow to estimate θ will be discussed later, see definition of consistency below.

7.2.2 Maximum likelihood estimator

We now present a very important example of estimator, known as the the Maximum Likelihood Estimator (or MLE).

We consider a parametric model with parameter space Θ given as

$$\mathfrak{F} = \{ f(x, \theta), \quad \theta \in \Theta \},$$

where for each θ, $f(x, \theta)$ is the PMF (resp. PDF) of a probability distribution. Let $X_1, \dots, X_n$ be i.i.d random variables with PMF (resp. PDF) $f(x; \theta)$, for some fixed, unknown, $\theta \in \Theta$.

Example 7.7. For example, you may think of $\Theta = (0, 1)$, and

$$\mathfrak{F} = \{ \mathrm{Ber}(x, p), \quad p \in (0, 1) \},$$

where for each p, $\mathrm{Ber}(x, p)$ is the PMF of a Bernoulli r.v. with parameter p:

$$\mathrm{Ber}(x, p) = \begin{cases} 1 - p & \text{if} \quad x = 0 \\ p & \text{if} \quad x = 1, \end{cases}$$

or, written more concisely,

$$\mathrm{Ber}(x,p) = (1-p)^{1-x}\ p^x, \qquad x = 0,1.$$

In this case we are considering $X_1,\dots,X_n$ to be i.i.d Bernoulli variables with some unknown parameter $p \in (0,1)$ that we'll try to estimate.

Example 7.8. As another example, you may think of $\Theta = \mathbb{R}$, and

$$\mathfrak{F} = \{\mathcal{N}(x;\mu,1), \quad \mu \in \mathbb{R}\},$$

where for each μ, $\mathcal{N}(x\colon \mu,1)$ is the PDF of a Gaussian r.v. with mean μ and spread 1. This means that we consider $X_1,\dots,X_n$ be i.i.d random Gaussian variables with spread 1 and some unknown mean $\mu \in \mathbb{R}$ that we will try to estimate.

So we have $X_1, X_2, \dots \sim f(x,\theta)$, where $\theta \in \Theta$ is an unknown parameter. In particular, for all $n \geqslant 1$, the random variables $X_1,\dots,X_n$ have a joint PMF (resp PDF) given by

$$(x_1,\dots,x_n) \mapsto \prod_{i=1}^{n} f(x_i,\theta).$$

Definition 7.9. *The likelihood function $\mathcal{L}_n\colon \Theta \to \mathbb{R}_+$ is defined by*

$$\mathcal{L}_n(\theta) := \prod_{i=1}^{n} f(X_i;\theta), \qquad \theta \in \Theta.$$

In other words, the likelihood function is obtained by plugging into the joint density of $X_1,\dots,X_n$, the observed realizations $X_1,\dots,X_n$. We also define the log-likelihood.

Definition 7.10. *The log-likelihood function $\ell_n\colon \Theta \to \mathbb{R}$ is defined by*

$$\ell_n(\theta) = \log\left(\mathcal{L}_n(\theta)\right) = \sum_{i=1}^{n} \log\left(f(X_i;\theta)\right), \qquad \theta \in \Theta.$$

It is important to note that $\mathcal{L}_n(\theta)$ and $\ell_n(\theta)$ are regarded here as functions of θ. Actually they also do depend on our observations $X_1,\dots,X_n$, but it is the former dependence which will require most of our attention. With the above definition at hand, we are now able to introduce an important example of estimator:

Definition 7.11. *The Maximum Likelihood Estimator (MLE) is defined by*

$$\hat{\theta}_n := \underset{\theta\in\Theta}{\mathit{argmax}}\ \ell_n(\theta).$$

In other words, $\hat{\theta}_n$ is the value of the parameter $\theta \in \Theta$ which maximises $\ell_n(\theta)$.

Remark 7.12. Note that, since $\ell_n(\theta) = \log(\mathcal{L}_n(\theta))$ and $\log\colon (0,\infty) \to \mathbb{R}$ is strictly increasing, $\hat{\theta}_n$ can equivalently be defined as the maximiser of $\mathcal{L}_n(\theta)$. However in applications it is often more convenient work with the log-likelihood rather than the likelihood itself.

Remark 7.13. It is not clear a priori that $\ell_n(\theta)$ should admit a unique maximiser. It will however often be the case in common examples we will encounter.

It is not obvious why the estimator $\hat{\theta}_n$ defined above does a "good job" at approximating θ. It will be verified in the next Chapters that the MLE is indeed an etsimator with several good properties which, in many cases, provides a practical way to find the true but unknown value of the parameter θ. In this Chapter we focus on explaining procedures to derive the MLE, and provide examples.

To compute the MLE, the usual way consists in three steps:

1. we compute the log-likelihood function $\ell_n(\theta)$
2. we differentiate it and search a $\theta \in \Theta$ such that $\ell_n'(\theta) = 0$
3. we check that this θ is indeed the maximum.

The above procedure will be illustrated by explicit analytical computations in a few classical examples. In practice, we rely on numerical methods, see e.g. the Newton-Raphson algorithm below.

7.2.3 Computation of the MLE: Analytical examples

Here we give specific examples where the MLE admits an explicit analytical expression.

7.2.3.1 Case of Bernoulli random variables

Assume that $X_1, X_2, \dots$ are i.i.d Bernoulli variables with some unknown parameter $p_0 \in (0,1)$. The likelihood function is given, for $p \in (0,1)$, by

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} \mathrm{Ber}(X_i, p) = (1-p)^{n(1-\bar{X}_n)}\ p^{n\bar{X}_n},$$

so the log-likelihood is given by

$$\ell_n(p) = n\,(1-\bar{X}_n)\,\log(1-p) + n\,\bar{X}_n \log(p).$$

Differentiating in p yields

$$\ell_n'(p) = n\left(-\frac{1-\bar{X}_n}{1-p} + \frac{\bar{X}_n}{p}\right) = n\,\frac{\bar{X}_n - p}{p\,(1-p)}.$$

Equating the above to 0 yields the following value for the MLE

$$\hat{p}_n = \bar{X}_n,$$

so that, in this case, the MLE coincides with the empirical mean of the X_i.

7.2.3.2 Case of Gaussian random variables (with known spread)

Assume that $X_1, X_2, \dots$ are i.i.d Gaussian variables with spread 1 and some unknown mean parameter $\mu_0 \in \mathbb{R}$. The likelihood function is given, for $\mu \in \mathbb{R}$, by

$$\mathcal{L}_n(\mu) = \prod_{i=1}^{n} \mathcal{N}(X_i; \mu, 1) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2 \right),$$

so the log-likelihood is given by

$$\ell_n(\mu) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Differentiating in μ yields

$$\ell_n'(\mu) = -\sum_{i=1}^{n} (\mu - X_i) = n(\bar{X}_n - \mu).$$

Equating the above to 0 yields the following value for the MLE

$$\hat{p}_n = \bar{X}_n,$$

so that, in this case again, the MLE coincides with the empirical mean of the X_i.

Exercise 7.1. In both examples discussed above, check that the MLE $\hat{p}_n$ (resp. $\widehat{mu}_n$) is *consistent*, i.e. that, when $n \to \infty$, it converges in probability to the true, but unknown value of the parameter p_0 (respectively μ_0). **Hint**: recall the LLN.

7.2.3.3 Generalised linear model

Consider the Generalised Linear Model $Y = X\beta + \epsilon$, where $X \in \mathbb{R}^{n \times p}$ is a given, known, design matrix, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Recall that here $\beta \in \mathbb{R}^p$ and $\sigma > 0$ are unknown parameters, so this model can be written as in (7.1). Assume we observe the vector

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

The likelihood function is then given by

$$\mathcal{L}(\beta, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (X\beta)_i)^2 \right)$$

so the log-likelihood takes the form

$$\ell_n(\beta,\sigma) \propto -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i - (X\beta)_i)^2 .$$

The MLE $(\hat{\beta},\hat{\sigma})$ is given by

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} (y_i - (X\beta)_i)^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - (X\hat{\beta})_i)^2.$$

Example 7.14. Assume that $p=2$ and the design matrix X is of the form

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n\times 2},$$

with $x_1,\dots,x_n \in \mathbb{R}$. Then, writing $\beta=(\beta_0,\beta_1)$, the previous model takes the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i=1,\dots n.$$

This is called a *simple linear regression model*. Denoting by $\bar{x}$ and $\bar{y}$ the sample means of x and y, one can check that for this special case, $\hat{\beta}$ is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2},$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.$$

In the terminology of linear regression, the above expressions are called the *least square estimates* for β_0 and β_1.

7.2.4 A numerical approach: the Newton-Raphson algorithm

Most of the time, it is not a problem to find an expression for the log-likelihood $\ell_n(\theta)$. However in many cases, finding an explicit expression for $\operatorname{argmax}_{\theta\in\Theta}\ell_n(\theta)$ is out of reach. In such cases we usually have to resort to a numerical approach, based on maximisation algorithms. There are plenty such algorithms, here we will focus on a specific one called Newton-Raphson algorithm.

The Newton-Raphson provides a general method to approach the maximiser $\operatorname{argmax}_{\theta\in\Theta} f(\theta)$ of a a function $f(\theta)$ defined over real numbers. This is done by looking for a critical point $\tilde{\theta}$ of f. The heuristics behind the Newton-Raphson algorithm is based on the following observation. For $\theta \approx \tilde{\theta}$, it should hold that

$$f(\tilde{\theta}) \approx f(\theta) + (\tilde{\theta} - \theta)\, f'(\theta).$$

But by assumption $f'(\tilde{\theta}) = 0$, so we get

$$\tilde{\theta} \approx \theta - \frac{f'(\theta)}{f''(\theta)}.$$

This suggests an update rule for the following iterative algorithm:

The Newton-Raphson algorithm:

Initialisation: Define a starting point $\theta^{(0)} \in \mathbb{R}$ and set $k := 0$.

If $f'(\theta^{(0)}) = 0$: STOP and return $\theta^{(0)}$. Else, proceed to Iterations

Iterations:

1. Set $\theta^{(k+1)} = \theta^{(k)} - \frac{f'(\theta^{(k)})}{f''(\theta^{(k)})}$.
2. If $f'(\theta^{(k+1)}) = 0$: STOP and return $\theta^{(k+1)}$. Else, go to 3.
3. Set $k := k + 1$ and go to 1.

We admit that, under appropriate assumptions on f, the above algorithm works in the sense that, for k sufficiently large, $\theta^{(k)}$ will be very close to the maximiser $\tilde{\theta}$ of f. Applying this algorithm to $f(\theta) = \ell_n(\theta)$, the log-likelihood function, thus allows to obtain a numerical approximation of the MLE $\hat{\theta}_n$.

Chapter 8
Finite-sample estimator properties

In this chapter we introduce and study properties of finite-sample estimators. Throuhgout this chapter we assume given a parametric model $\mathcal{F}=\{f(x;\theta)\colon \theta\in\Theta\}$ and a sample $X_1,\dots,X_n$ of i.i.d. random variables distributed according to the PMF (or PDF) $f(x;\theta)$, for some (unknown) parameter value $\theta\in\Theta$.

The contents from this chapter are taken from Chapter 9 in [5] and from Chapter 9 in [3].

8.1 Error, bias and unbiasedness

Definition 8.1. *Let $\hat{\theta}_n=\hat{\theta}_n(X_1,\dots,X_n)$ be an estimator for θ constructed from our observed sample $X_1,\dots,X_n$.*

- *The error is defined as*
$$\hat{\theta}_n-\theta$$
- *The bias is defined as*
$$\mathrm{bias}(\hat{\theta}_n)=\mathbb{E}_\theta(\hat{\theta}_n)-\theta$$
- *We call the estimator $\hat{\theta}_n$ unbiased if $\mathrm{bias}(\hat{\theta}_n)=0$, or in other words if*
$$\mathbb{E}_\theta(\hat{\theta}_n)=\theta$$
Otherwise we say that it is biased.

Remark 8.2. Above we write a superscript θ for $\mathbb{E}_\theta$ to stress the fact that we are working under the assumption that the X_i follow the distribution with PMF (or PDF) $f(x,\theta)$. Keep in mind that θ is supposed to be unknown.

Remark 8.3. The biasedness property is likeable, but not crucial. As we shall see some very important estimators such as the MLE are biased in general.

So far we have defined an estimator $\hat{\theta}_n$ of a parameter $\theta \in \Theta$. In many cases however we are not interested in estimating θ itself, but rather some quantity associated with the PMF (or PDF) $f(x, \theta)$, called satistical functional.

Definition 8.4. *If $f = f(x, \theta)$ is the PMF (or PDF) associated with a parameter value θ, a statistical functional is any function $T(f)$ of f.*

Example 8.5. The mean $\mu = \int_{\mathbb{R}} x\, f(x)\, \mathrm{d}x$ and the variance $\sigma^2 = \int_{\mathbb{R}} (x - \mu)^2 f(x)\, \mathrm{d}x$ are both example of statistical functionals.

Comng back to our setting above, we assume $X_1, \ldots, X_n \sim f(x, \theta)$ for some true but unkown parameter value θ. Let $\mu = \int_{\mathbb{R}} x\, f(x, \theta)\, \mathrm{d}x$ be the mean of the true (but unkown) distribution of our observations, and let $\sigma^2 = \int_{\mathbb{R}} (x - \mu)^2\, f(x; \theta) \mathrm{d}x$ be its variance. These are both relevent statistical functionals, whcih we may try to estimate.

Theorem 8.6.

1. *The sample mean $\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k$ is an unbiased estimator for μ*
2. *The sample variance $S_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X_n})^2$ is an unbiased estimator for σ^2.*

Proof.

1. We readily saw in 6.4 that $\mathbb{E}(\overline{X_n}) = \mu$, so that $\overline{X_n}$ is indeed unbiased.
2. Rewriting, for all k,

$$X_k - \bar{X}_n = (X_k - \mu) - (\overline{X_n} - \mu),$$

we can then rewrite $\sum_{k=1}^{n} (X_k - \overline{X_n})^2$ as the sum of three terms:

$$\sum_{k=1}^{n} (X_k - \mu)^2 - 2 \sum_{k=1}^{n} (X_k - \mu)(\overline{X_n} - \mu) + n\, (\bar{X}_n - \mu)^2.$$

Taking expectation of the first term gives

$$\mathbb{E}\left(\sum_{k=1}^{n} (X_k - \mu)^2 \right) = n\sigma^2.$$

In the second term we recognise - $2n\, (\overline{X_n} - \mu)^2$, taking its expectation yields

$$-2n\, \mathbb{E}\, (\overline{X_n} - \mu)^2 = -2n\, \mathbb{V}(\bar{X}_n) = -2\sigma^2.$$

Finally, expectation of the third term gives

$$\mathbb{E}(n\,(\bar{X}_n-\mu)^2)=n\,\mathbb{V}(\bar{X}_n)=\sigma^2.$$

Summing up the three, we obtain

$$\mathbb{E}\left(\sum_{k=1}^{n}\,(X_k-\overline{X_n})^2\right)=n\sigma^2-2\sigma^2+\sigma^2=(n-1)\sigma^2,$$

so that

$$\mathbb{E}(S_n^2)=\mathbb{E}\left(\frac{1}{n-1}\sum_{k=1}^{n}\,(X_k-\overline{X_n})^2\right)=\sigma^2,$$

as requested. □

Question: What estimator can we consider for the standard deviation $\sigma=\sqrt{\sigma^2}$? A natural choice would be to consider the *sample standard deviation*

$$S_n=\sqrt{S_n^2}=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\,(X_i-\bar{X}_n)^2}$$

Proposition 8.7. *Assume that $X_1,\ldots,X_n$ are not deterministic. Then S_n is a biased estimator of σ.*

Proof. Since the function $x\mapsto\sqrt{x}$ is strictly concave from $\mathbb{R}_+$ to itself, by Jensen's inequality (see A.3.7 in [2]), we have

$$\mathbb{E}_\theta(S_n)=\mathbb{E}_\theta\Big(\sqrt{S_n^2}\Big)<\sqrt{\mathbb{E}_\theta(S_n^2)}=\sqrt{\sigma^2}=\sigma.$$ □

The fact that the estimator S_n for σ is a biased is not a pathological case. It is a general fact that non-linear transformations of unbiased estimators may typically be biased.

8.2 Variance, standard error and mean-squared error.

Let $\hat{\theta}_n=\hat{\theta}_n(X_1,\ldots,X_n)$ be an estimator. In the previous section we introuced the error $\hat{\theta}_n-\theta$. Heuristically the estimator $\hat{\theta}_n$ is good if the error is small. In this section we shall introduce ways to quantify more precisely the error.

Definition 8.8.

1. *The* variance *of* $\hat{\theta}_n$ *is defined as*

$$\mathbb{V}_\theta(\hat{\theta}_n) := \mathbb{E}_\theta((\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n))^2)$$

2. *The* standard error of $\hat{\theta}_n$ *is defined as*

$$\mathrm{se}(\hat{\theta}_n) := \sqrt{\mathbb{V}_\theta(\hat{\theta}_n)}.$$

Remark 8.9. The above are nothing other than the *variance* and *standard deviation* of the random variable $\hat{\theta}_n$.

Example 8.10. Let $p \in (0,1)$ and $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$. We consider the MLE $\hat{p}_n$ for p. Recall that in this case $\hat{p}_n = \bar{X}_n$. The variance of $\hat{p}_n$ is then given by

$$\mathbb{V}_p(\hat{p}_n) = \mathbb{V}_p(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}_p(X_i) = \frac{p(1-p)}{n}.$$

The standard error is then given by

$$\mathrm{se}(\hat{p}_n) = \sqrt{\mathbb{V}_p(\hat{p}_n)} = \sqrt{\frac{p(1-p)}{n}}. \tag{8.1}$$

Note that p is in a priori unknonwn to us: it is precisely the parameter we are trying to estimate. So a priori we do not have access to the quantity $\sqrt{\frac{p(1-p)}{n}}$. However, we can approximate the true standard error by the *estimated standard error*

$$\hat{\mathrm{se}}(\hat{p}_n) := \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}.$$

That quantity in turn is written completely in terms of our observed data, and we can thus evaluate it.

Definition 8.11. *Let* $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ *be an estimator for* θ*. The mean-squared error of* $\hat{\theta}_n$ *is defined as*

$$\mathrm{MSE}(\hat{\theta}_n) := \mathbb{E}_\theta((\hat{\theta}_n - \theta)^2)$$

Remark 8.12. The mean-squared error quantifies how far our estimator lies away from the true but unknown parameter value θ. As we will see in the next Chapter, if we know that $\mathrm{MSE}(\hat{\theta}_n) \to 0$ as $n \to \infty$, we can deduce that $\hat{\theta}_n \to \theta$ in probability (a property we will call consistency).

Remark 8.13. Be careful not to mix up the standard error $\mathrm{se}(\hat{\theta}_n)$ and the mean-squared error $\mathrm{MSE}(\hat{\theta}_n)$! See the recap chart below.

error	$\hat{\theta}_n - \theta$
$\text{bias}(\hat{\theta}_n)$	$\mathbb{E}_\theta(\hat{\theta}_n) - \theta$
$\mathbb{V}_\theta(\hat{\theta}_n)$	$\mathbb{E}_\theta((\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n))^2)$
$\text{se}(\hat{\theta}_n)$	$\sqrt{\mathbb{V}_\theta(\hat{\theta}_n)}$
$\text{MSE}(\hat{\theta}_n)$	$\mathbb{E}_\theta((\hat{\theta}_n - \theta)^2)$

The different quantities above are related by the below:

Theorem 8.14. *(Mean squared error decomposition)*
Let $\hat{\theta}_n$ be an estimator for θ. We have

$$\text{MSE}(\hat{\theta}_n) = \mathbb{V}_\theta(\hat{\theta}_n) + \text{bias}(\hat{\theta}_n)^2 \tag{8.2}$$

Proof. We have

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}_\theta((\hat{\theta}_n - \theta)^2) = \mathbb{V}_\theta(\hat{\theta}_n - \theta) + \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2,$$

where the second equality follows by the translation variance Theorem 5.17 applied to the random variable $X = \hat{\theta}_n - \theta$. Now, the first term equals $\mathbb{V}(\hat{\theta}_n)$ while in the second term we recognise $\text{bias}(\hat{\theta}_n)^2$. This yields the claim. □

Remark 8.15. Recalling that $\mathbb{V}_\theta(\hat{\theta}_n) = \text{se}(\hat{\theta}_n)^2$, we can alternatively write the above decomposition as

$$\text{MSE}(\hat{\theta}_n) = \text{se}(\hat{\theta}_n)^2 + \text{bias}(\hat{\theta}_n)^2$$

Remark 8.16. In the case where $\hat{\theta}_n$ is *unbiased,* the above decomposition reduces to $\text{MSE}(\hat{\theta}_n) = \mathbb{V}_\theta(\hat{\theta}_n)$.

Motivation for (8.2) : Assume that you want to prove that $\hat{\theta}_n$ is a consistent estimator for θ, see chapter below for the definition. It suffices to check that $\hat{\theta}_n \to \theta$ in quadratic mean, because this will entail convergence in probability of $\hat{\theta}_n$ to θ. To prove the former convergence, it suffices by the above Theorem to check that $\text{bias}(\hat{\theta}_n)$ and $\mathbb{V}_\theta(\hat{\theta}_n)$ both converge to 0 when $n \to \infty$.

8.3 The Cramér-Rao bound

We saw above that if $\hat{\theta}_n$ is an *unbiased* estimator for θ, then $\text{MSE}(\hat{\theta}_n) = \mathbb{V}_\theta(\hat{\theta}_n)$. It is therefore tempting to choose $\hat{\theta}_n$ so that $\mathbb{V}_\theta(\hat{\theta}_n)$ is small. How small could it possibly be? In order to answer the question we need to introduce a couple of definitions.

Let us again assume $X_1, \dots, X_n \sim f(x, \theta)$, where $x \mapsto f(x, \theta)$ is a PMF or PDF, and $\theta \in \Theta$ is an unkown parameter value. Let $\ell_n \colon \Theta \to \mathbb{R}$ be the log-likelihood function.

Definition 8.17.

- *The* score function *of the random sample $X_1, \dots, X_n$ is defined as*
$$s_n(\theta) := \frac{\mathrm{d}}{\mathrm{d}\theta} \ell_n(\theta)$$
- *The (expected)* Fisher information *of the random sample $X_1, \dots, X_n$ is defined as*
$$I_n(\theta) := \mathbb{E}_\theta\bigg(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ell_n(\theta) \bigg)$$

Remark 8.18. In Chapter 9 of [3], the quantity
$$\mathbb{E}_\theta\bigg(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ell_n(\theta) \bigg)$$
is referred to as the expected Fisher information and denoted by J_n. Here we are following the notations of [5].

For $n = 1$ we shall write
$$s(\theta) := s_1(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_1, \theta), \qquad I(\theta) := I_1(\theta) = \mathbb{E}_\theta\bigg(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X_1, \theta) \bigg)$$

Remark 8.19. We have $I_n(\theta) = n\, I(\theta)$.

Proof. We have
$$\begin{aligned} I_{n(\theta)} &= \mathbb{E}_\theta\bigg(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \sum_{i=1}^{n} \log f(X_i, \theta) \bigg) \\ &= \sum_{i=1}^{n} \mathbb{E}_\theta\bigg(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X_i, \theta) \bigg) \\ &= n\, \mathbb{E}_\theta\bigg(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X_1, \theta) \bigg) \end{aligned}$$
where the last equality follows from the fact that the X_i all have the same distribution. The claim follows. □

Lemma 8.20. *We have*
$$\mathbb{E}(s(\theta)) = 0, \qquad \mathbb{V}_\theta(s(\theta)) = I(\theta)$$

We refer to pages 390-392 of [3] for the proof.

We are now able to state the main result of this section:

Theorem 8.21. *[Cramér-Rao lower bound]*
Under appropriate assumptions (known as "Fisher regularity conditions"), the following holds. Let g be a differentiable function, and let $\hat{\theta}_n$ be an unbiased estimator for $g(\theta)$. Then

$$\mathbb{V}_\theta(\hat{\theta}_n) \geqslant \frac{(g'(\theta))^2}{I_n(\theta)}$$

In particular, for $g(\theta) = \theta$, we have

$$\mathbb{V}_\theta(\hat{\theta}_n) \geqslant \frac{1}{I_n(\theta)}.$$

The Fisher regularity conditions are a standard set of assumptions that will be satisfied in the examples we shall be considering. We refer to Definition 4.1 in [2] for their precise statement, see also p. 395 in [3].

Remark 8.22. Recalling Remark 8.16, in the case where we are estimating $g(\theta) = \theta$, the above theorem shows that

$$\mathrm{MSE}(\hat{\theta}_n) \geqslant \frac{1}{I_n(\theta)}.$$

Thus, a highly Fisher informative unbiased estimator will have a small varinace, and hence a small mean-squared error.

Proof. We apply the correlation inequality (5.2) to the random variables $X = s_n(\theta)$ and $Y = \hat{\theta}_n$. We get

$$-1 \leqslant \frac{\mathrm{Cov}_\theta(s_n(\theta), \hat{\theta}_n)}{\sqrt{\mathbb{V}_\theta(s_n(\theta))}\sqrt{\mathbb{V}_\theta(\hat{\theta}_n)}} \leqslant 1$$

Taking the square yields

$$\frac{\mathrm{Cov}_\theta(s_n(\theta), \hat{\theta}_n)^2}{\mathbb{V}_\theta(s_n(\theta))\mathbb{V}_\theta(\hat{\theta}_n)} \leqslant 1,$$

i.e.

$$\frac{\mathrm{Cov}_\theta(s_n(\theta), \hat{\theta}_n)^2}{\mathbb{V}_\theta(s_n(\theta))} \leqslant \mathbb{V}_\theta(\hat{\theta}_n).$$

But, by Lemma 8.20, $\mathbb{V}_\theta(s_n(\theta)) = I_n(\theta)$, hence

$$\mathbb{V}_\theta(\hat{\theta}_n) \geqslant \frac{\mathrm{Cov}_\theta(s_n(\theta), \hat{\theta}_n)^2}{I_n(\theta)}.$$

So to obtain the requested lower bound it only remains to prove that

$$\mathrm{Cov}_\theta(s_n(\theta), \hat{\theta}_n) = \frac{d}{d\theta} g(\theta).$$

In view of Lemma 8.20 $s_n(\theta)$ is centered, hence the covariance above equals

$$\begin{aligned}\mathbb{E}_\theta(s_n(\theta)\,\hat{\theta}_n) &= \mathbb{E}_\theta\Big(\frac{\mathrm{d}}{\mathrm{d}\theta}\log(\mathcal{L}_n(\theta))\,\hat{\theta}_n\Big)\\ &= \mathbb{E}_\theta(\frac{1}{\mathcal{L}_n(\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{L}_n(\theta)\hat{\theta}_n),\end{aligned}$$

where $\mathcal{L}_n(\theta)=\prod_{i=1}^n f(X_i;\theta)$ is the likelihood function. Recalling that $X_1,\ldots,X_n$ have joint density given by

$$(x_1,\ldots,x_n)\mapsto\prod_{i=1}^n f(x_i;\theta),$$

we see that the above takes the form of the n-fold integral

$$\int\cdots\int\frac{1}{\prod_{i=1}^n f(x_i;\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\prod_{i=1}^n f(x_i;\theta)\Big)\hat{\theta}_n(x_1,\ldots,x_n)\prod_{i=1}^n f(x_i;\theta)\,\mathrm{d}x_1\ldots\mathrm{d}x_n$$

(here we assume for simplicity that we are in the case where $f(x;\theta)$ is a PDF, the case where it is a PMF is similar). So

$$\begin{aligned}\mathrm{Cov}_\theta(s_n(\theta),\hat{\theta}_n) &= \int\cdots\int\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\prod_{i=1}^n f(x_i;\theta)\Big)\hat{\theta}_n(x_1,\ldots,x_n)\,\mathrm{d}x_1\ldots\mathrm{d}x_n\\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\int\cdots\int\Big(\prod_{i=1}^n f(x_i;\theta)\Big)\hat{\theta}_n(x_1,\ldots,x_n)\,\mathrm{d}x_1\ldots\mathrm{d}x_n\Big)\\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta(\hat{\theta}_n).\end{aligned}$$

Since $\hat{\theta}_n$ is an unbiased estimator for $g(\theta)$, the last quantity equals $g'(\theta)$, and the claim follows. □

Example 8.23. Assume that $X_1,\ldots,X_n\sim\mathrm{Ber}(p)$ for $p\in(0,1)$. We consider the maximum-likelihood estimator $\hat{p}_n$ for p. Recall that in this case the MLE coincides with the sample mean:

$$\hat{p}_n=\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i.$$

By Theorem 8.6, $\hat{p}_n$ is an unbiased estimator for p. Moreover, as seen above,

$$\mathbb{V}_p(\hat{p}_n)=\frac{p(1-p)}{n}. \tag{8.3}$$

Denoting by $\ell_n\colon(0,1)\to\mathbb{R}$ the log-likelihood function, we have

$$\ell_n(p)=n(1-\bar{X}_n)\log(1-p)+n\,\bar{X}_n\log(p).$$

Differentiating twice we get

$$\ell_n''(p) = \frac{n(\bar{X}_n - 1)}{(1-p)^2} - \frac{n\,\bar{X}_n}{p^2}.$$

Taking expectation we get

$$I_n(p) = \mathbb{E}_p(-\ell_n''(p)) = \frac{n(1-p)}{(1-p)^2} + \frac{n\,p}{p^2} = \frac{n}{p(1-p)}. \tag{8.4}$$

By (8.3) and (8.4) we see that $\mathbb{V}_p(\hat{p}_n) = \frac{1}{I_n(p)}$, i.e. in the Bernoulli case the MLE achieves the equality in the Cramér-Rao lower bound.

Example 8.24. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$, with $\mu \in \mathbb{R}$. We consider the MLE $\hat{\mu}_n$, here also given by the sample mean. We know from Theorem 8.6 that $\hat{\mu}_n$ is unbiased, and we have

$$\mathbb{V}_\mu(\hat{\mu}_n) = \mathbb{V}_\mu\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_\mu(X_i) = \frac{1}{n}.$$

We shall compute the Fisher information using a slightly simpler route than used in the previous example. We compute the log-likelihood just for $n = 1$. We have

$$\begin{aligned} \ell_1(\mu) &= \log\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(X_1-\mu)^2}{2}\right)\right) \\ &= -\frac{1}{2}\log(2\pi) - \frac{(\mu - X_1)^2}{2}. \end{aligned}$$

Differentiating twice we get $\ell_1''(\mu) = -1$, so that $I(\mu) = 1$. In view of Remark 8.19 we deduce that $I_n(\mu) = n$. Thus, we have

$$\mathbb{V}_\mu(\hat{\mu}_n) = \frac{1}{n} = \frac{1}{I_n(\mu)},$$

so, also in that case, the equality in the Cramér-Rao lower bound is achieved.

Example 8.25. Consider $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \theta)$, where both μ and θ are unknown. To estimate θ we consider the estimator given by the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X_n})^2.$$

By Theorem 8.6 this is an unbiased estimator for the variance of the X_i, here given by θ. One can prove that

$$\mathbb{V}(S_n^2) = \frac{2\theta^2}{n-1},$$

while the Cramér-Rao lower bound is given by $\frac{2\theta^4}{n}$, see Example 7.3.14 in [1].

Chapter 9
Asymptotic estimator properties

In this Chapter we study properties of estimators $\hat{\theta}_n$ for a large sample size n. Such properties are referred to as asymptotic properties. We will mainly focus on 4 properties known as:

1. Asymptotic unbiasedness
2. Consistency
3. Asymptotic normality
4. Asymptotic efficiency

As an example, we shall see that the MLE satsifies all of the above.

Here and below, we assume given a parametric model $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ and a sample $X_1, \ldots, X_n \sim \theta$, where $\theta \in \Theta$ is the true, but unkwon, parameter value.

9.1 Asymptotic unbiasedness

Definition 9.1. *An estimator $\hat{\theta}_n$ for θ is* asymptotically unbiased *if*

$$\lim_{n\infty} \mathbb{E}_\theta(\hat{\theta}_n) = \theta\,.$$

Comparing this with Definition 8.1 we see that $\hat{\theta}_n$ being asymptotically unbiased means that it is unbiased after sending $n \to \infty$. In practice, it means that, if we have a *large sample size*, we expect $\hat{\theta}_n$ to be *approximately* unbiased, i.e. $\mathbb{E}_\theta(\hat{\theta}_n) \approx \theta$. In particular, by definition, an unbiased estimator is always asymptotically unbiased.

Example 9.2. Consider $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$, $\sigma > 0$. We consider the MLE $(\hat{\mu}_n, \hat{\sigma}_n)$ for (μ, σ). This is obtained by maximising the log-likelihood function $\ell_n : \mathbb{R} \times (0,\infty) \to \mathbb{R}$, which is here given by

$$\begin{aligned} \ell_n(\mu,\sigma) &= -\sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)\right) \\ &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2} \end{aligned}$$

Note that here ℓ_n is a function of two variables. Its maximiser has to solve the equations:

$$\frac{\partial}{\partial\mu}\ell_n(\mu,\sigma)=\frac{\partial}{\partial\sigma}\ell_n(\mu,\sigma)=0.$$

Differentiating the expression above for ℓ_n in μ and σ, we obtain the expressions

$$\frac{\partial}{\partial\mu}\ell_n(\mu,\sigma)=-\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu-X_i)=-\frac{n}{\sigma^2}(\mu-\bar{X}_n),$$

where $\bar{X}_n$ denotes the sample mean, and

$$\frac{\partial}{\partial\sigma}\ell_n(\mu,\sigma)=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(\mu-X_i)^2=-\frac{n}{\sigma}\left(1-\frac{1}{n\,\sigma^2}\sum_{i=1}^{n}(\mu-X_i)^2\right).$$

Equating these to 0 yields the following expression for the MLE:

$$\hat{\mu}_n=\bar{X}_n,\qquad \hat{\sigma}_n=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

In particular, note that $\hat{\mu}_n$ is given by the sample mean of the random variables X_i^2. In view of Theorem 8.6, we know that $\hat{\mu}_n$ is an unbiased estimator for μ. Regarding the estimator $\hat{\sigma}_n^2$ for σ^2, note that $\hat{\sigma}_n^2$ is NOT the sample variance of the X_i, compare with Definition 5.3. Actually, $\hat{\sigma}_n^2$ is a *biased* estimator for σ^2. Indeed, note that $\hat{\sigma}_n^2=\frac{n-1}{n}S_n^2$, where S_n^2 is the sample variance of the X_i. In virtue of Theorem 8.6, $\mathbb{E}(S_n^2)=\sigma^2$, hence it follows that

$$\mathbb{E}(\hat{\sigma}_n^2)=\mathbb{E}\left(\frac{n-1}{n}S_n^2\right)=\frac{n-1}{n}\mathbb{E}(S_n^2)=\frac{n-1}{n}\sigma^2\neq\sigma^2,$$

so that $\hat{\sigma}_n^2$ is indeed a biased estimator for s^2. We claim however that $\hat{\sigma}_n^2$ is an asymptotically *unbiased* estimator for σ^2. Indeed

$$\mathbb{E}(\hat{\sigma}_n^2)=\frac{n-1}{n}\sigma^2\underset{n\to\infty}{\longrightarrow}\sigma^2.$$

In fact, below we will see that in general the MLE is an *asymptotically unbiased* estimator, although it is often biased for fixed sample size n.

9.2 Consistency

Definition 9.3. *An estimator $\hat{\theta}_n$ for θ is said to be consistent if $\hat{\theta}_n\underset{n\to\infty}{\longrightarrow}\theta$ in probability.*

In other words, an estimator $\hat{\theta}_n$ for θ is consistent if, for all $\varepsilon>0$, we have

$$\lim_{n\to\infty}\mathbb{P}_\theta(|\hat{\theta}_n-\theta|>\varepsilon)=0.$$

Theorem 9.4. *[Mean-squared error criterion for consistency] Let $\hat{\theta}_n$ be an estimator for θ. Assume that*

$$\lim_{n\to\infty} \mathrm{MSE}(\hat{\theta}_n) = 0.$$

Then $\hat{\theta}_n$ is consistent.

Proof. By assumption we have

$$\lim_{n\to\infty} \mathbb{E}_\theta(|\hat{\theta}_n - \theta|^2) = \lim_{n\to\infty} \mathrm{MSE}(\hat{\theta}_n) = 0.$$

Thus $\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta$ in quadratic mean, and hence (in view of the 1st point in Theorem 6.10), in probability.

□

Corollary 9.5. *[Bias and variance criterion for consistency] Let $\hat{\theta}_n$ be an estimator for θ. Assume that*

$$\lim_{n\to\infty} \mathbb{V}_\theta(\hat{\theta}_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} \mathrm{bias}(\hat{\theta}_n) = 0$$

Then $\hat{\theta}_n$ is consistent.

Proof. The above follows from Theorem 9.4 and the mean-squared error decomposition (8.2). □

Remark 9.6. Recalling that $\mathbb{V}_\theta(\hat{\theta}_n) = \mathrm{se}(\hat{\theta}_n)^2$, an alternative way of writing the above criterion is

$$\lim_{n\to\infty} \mathrm{se}(\hat{\theta}_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} \mathrm{bias}(\hat{\theta}_n) = 0.$$

Example 9.7. Let $X_1, \ldots, X_n$ be i.i.d., square-integrable, random variables with mean μ and variance σ^2. Then the sample mean $\bar{X}_n$ is a consistent estimator for μ. Indeed, by Theorem 8.6, $\mathrm{bias}(\bar{X}_n) = 0$, while

$$\mathbb{V}(\bar{X}_n) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i),$$

where the last equality follows from the independence of the X_i. So we get

$$\mathbb{V}(\bar{X}_n) = \frac{n\,\sigma^2}{n^2} = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0.$$

The claim follows by the bias and variance criterion for consistency.

9.3 Asymptotic normality

Definition 9.8. *An estimator $\hat{\theta}_n$ for θ is said to be asymptotically normal if*

$$\frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Thus, the estimator $\hat{\theta}_n$ for θ is asymptotically normal if, for n large, the distribution of $\hat{\theta}_n$ is approximately equal to $\mathcal{N}(\theta, \mathrm{se}(\hat{\theta}_n)^2)$.

Example 9.9. Let $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$, for some unkown $p \in (0,1)$. We consider the MLE $\hat{p}_n$ for p, which is here given by the sample mean $\bar{X}_n$. Then $\hat{p}_n$ is asymptotically normal. Indeed, in this case the X_i have mean p and variance $p(1-p)$, so, in virtue of the CLT, we have

$$\sqrt{\frac{n}{p(1-p)}}(\bar{X}_n - p) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

Recalling from (8.1) that $\mathrm{se}(\hat{\theta}_n) = \sqrt{\frac{p(1-p)}{n}}$, we deduce that

$$\frac{(\hat{p}_n - p)}{\mathrm{se}(\hat{p}_n)} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

as requested.

9.4 Asymptotic efficiency

Recall the Cramér-Rao lower bound for an unbiased estimator $\hat{\theta}_n$ for θ:

$$\mathrm{se}(\hat{\theta}_n) \geqslant \frac{1}{\sqrt{I_n(\theta)}},$$

which we here rewrote in terms of the standard error of $\hat{\theta}_n$. It is likeable for $\hat{\theta}_n$ to achieve this lower bound, but this will in general not be the case for fixed n. However, we can still hope that

$$\mathrm{se}(\hat{\theta}_n) \approx \frac{1}{\sqrt{I_n(\theta)}}$$

for n large. Combining this property with the notion of asymptotic normality seen above motivates the following definition:

Definition 9.10. *An estimator $\hat{\theta}_n$ for θ is said to be asymptotically efficient if*

$$\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

In other words, $\hat{\theta}_n$ is a symtotically efficient if it is asymptotically normal *and* it satisfies $\mathrm{se}(\hat{\theta}_n) \approx \frac{1}{\sqrt{I_n(\theta)}}$ for n large, or equivalently if

$$\hat{\theta}_n \overset{(d)}{\approx} \mathcal{N}\left(\theta, \frac{1}{I_n(\theta)}\right)$$

when n is large.

Remark 9.11. The term *efficient* is used in different ways in the literature. Definition above is the one followed by [3], see p.441 therein, see also section 9.8 in [5].

Remark 9.12. If $\hat{\theta}_n$ is asymptotically efficient, it is in particular asymptotically unbiased and asymptotically normal.

Example 9.13. Let again $X_1, \dots, X_n \sim \mathrm{Ber}(p)$, for some unkown $p \in (0,1)$. We consider the MLE $\hat{p}_n$ for p, given by the sample mean $\bar{X}_n$. We have seen previously that $\hat{p}_n$ is asymptotically normal:

$$\frac{(\hat{p}_n - p)}{\mathrm{se}(\hat{p}_n)} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

In addition, we have seen in the previous Chapter, see (8.4) and (8.3), that

$$\mathrm{se}(\hat{p}_n) = \sqrt{\frac{p(1-p)}{n}} = \frac{1}{I_n(p)}.$$

It follows at once that $\hat{p}_n$ is asymptotically efficient.

9.5 Properties of the Maximum Likelihood Estimator

In this section, we show that the Maximum Likelihood Estimator satisfies many of the asymptotic properties defined above.

Here and below, we fix a parametric model $\mathcal{F} = \{ f(x;\theta) : \theta \in \Theta \}$ which we assume to fulfill a standard set of assumptions known as the Fisher regularity conditions, see Definition 4.1 in [2] for their precise statement, as well as p. 395 in [3]. We denote by $\ell_n : \Theta \to \mathbb{R}$ the associated log-likelihood function.

We are given $X_1, \dots, X_n \sim f(x;\theta)$ for some unkown $\theta \in \Theta$ and denote by $\hat{\theta}_n$ the associated MLE. We will prove the following:

Theorem 9.14. *The MLE is a consistent, asymptotically efficient estimator for θ.*

Proof. We refer to p. 135 of [5] for the proof of consistency. Here we detail the proof of asymptotic efficiency, along the lines of pp. 136 and 137 in [5]. To do so, we introduce the log-likelihood function $\ell_n : \Theta \to \mathbb{R}$ as well as the score function

$$s_n(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \ell_n(\theta), \qquad \theta \in \Theta$$

and the Fisher information

$$I_n(\theta) := \mathbb{E}_\theta\left(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ell_n(\theta) \right) = \mathbb{E}_\theta\left(-\frac{\mathrm{d}}{\mathrm{d}\theta} s_n(\theta) \right).$$

Since $\hat{\theta}_n$ is a consistent estimator for θ (as we admitted), it is legitimate to think of $\hat{\theta}_n$ as being close to θ when n is large, so we may write the following approximation

$$s_n(\hat{\theta}_n) \approx s_n(\theta) + (\hat{\theta}_n - \theta)\,\frac{\mathrm{d}s_n}{\mathrm{d}\theta}(\theta).$$

Now, the left-hand side above vanishes, as by definition $\hat{\theta}_n$ maximises the likelihood function, and hence $s_n(\hat{\theta}_n) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\hat{\theta}_n) = 0$. So we get

$$s_n(\theta) + (\hat{\theta}_n - \theta)\frac{\mathrm{d}}{\mathrm{d}\theta}s_n(\theta) \approx 0,$$

i.e.

$$\hat{\theta}_n - \theta \approx \frac{s_n(\theta)}{\left(-\frac{\mathrm{d}}{\mathrm{d}\theta}s_n(\theta)\right)}.$$

With this approximation, and recalling that $I_n(\theta) = n\,I(\theta)$, we get

$$\begin{aligned}\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) &= \sqrt{n\,I(\theta)}\,(\hat{\theta}_n - \theta)\\ &= \frac{\frac{1}{\sqrt{n\,I(\theta)}}\,s_n(\theta)}{\frac{1}{n\,I(\theta)}\left(-\frac{\mathrm{d}}{\mathrm{d}\theta}s_n(\theta)\right)}.\end{aligned}$$

We will study convergence of the numerator and of the denominator of the last expression separately.

Numerator:

Note that $s_n(\theta) = \sum_{i=1}^n Y_i = n\,\bar{Y}_n$, where

$$Y_i := \frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(X_i;\theta)), \qquad i = 1, \ldots, n.$$

Note that the Y_i are i.i.d. We now compute their mean and variance. We have

$$\begin{aligned}\mathbb{E}_\theta(Y_i) &= \mathbb{E}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(X_i;\theta))\right]\\ &= \mathbb{E}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(X_1;\theta))\right]\\ &= \mathbb{E}_\theta[s_1(\theta)]\\ &= 0,\end{aligned}$$

where the last equality follows from Lemma 8.20. So the Y_i are centered. Their variance is given by

$$\begin{aligned}\mathbb{V}_\theta(Y_i) &= \mathbb{V}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(X_i;\theta))\right]\\ &= \mathbb{V}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(X_1;\theta))\right]\\ &= \mathbb{V}_\theta[s_1(\theta)]\\ &= I(\theta)\end{aligned}$$

where the last equality follows from Lemma 8.20. So the Y_i are i.i.d., centered r.v.'s with variance $I(\theta)$. By the CLT, we have

$$\sqrt{\frac{n}{I(\theta)}}\,\bar{Y}_n \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Rewriting this in terms of $s_n(\theta)$ yields

$$\sqrt{\frac{1}{n\,I(\theta)}}\,s_n(\theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Denominator:

We have

$$\frac{1}{n\,I(\theta)}\left(-\frac{\mathrm{d}}{\mathrm{d}\theta}s_n(\theta)\right) = \bar{Z}_n,$$

where $Z_1,\ldots,Z_n$ are the i.i.d. random variables given by

$$Z_i := \frac{1}{I(\theta)}\left(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(X_i;\theta))\right).$$

The Z_i have mean given by

$$\begin{aligned}\mathbb{E}_\theta(Z_i) &= \frac{1}{I(\theta)}\mathbb{E}_\theta\left(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(X_i;\theta))\right)\\ &= \frac{}{I(\theta)}I(\theta)\\ &= 1,\end{aligned}$$

where the second equality follows by te very definition of $I(\theta)$. Hence, by the LLN, we have $\bar{Z}_n \xrightarrow[n\to\infty]{(P)} 1$, i.e.

$$\frac{1}{n\,I(\theta)}\left(-\frac{\mathrm{d}}{\mathrm{d}\theta}s_n(\theta)\right)\xrightarrow[n\to\infty]{(P)} 1.$$

Conclusion:

We have

$$\sqrt{I_n(\theta)}(\hat{\theta}_n-\theta) = \frac{\frac{1}{\sqrt{n\,I(\theta)}}\,s_n(\theta)}{\frac{1}{n\,I(\theta)}\left(-\frac{\mathrm{d}}{\mathrm{d}\theta}s_n(\theta)\right)}$$

with

$$\frac{1}{\sqrt{n\,I(\theta)}}\,s_n(\theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

and

$$\frac{1}{n\,I(\theta)}\left(-\frac{\mathrm{d}}{\mathrm{d}\theta}s_n(\theta)\right)\xrightarrow[n\to\infty]{(P)} 1.$$

The claimed result now follows by the Lemma below. □

Lemma 9.15. *Let $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ be sequences of real-valued r.v.'s such that*

$$X_n \xrightarrow[n\to\infty]{(d)} X$$

for some random variable X and

$$Y_n \xrightarrow[n\to\infty]{(P)} c,$$

for some strictly positive constant c. Then

$$\frac{X_n}{Y_n} \xrightarrow[n\to\infty]{(d)} \frac{X}{c}.$$

Proof. This is an easy consequence of the continuous mapping Theorem and Slutsky's Theorem, see p. 451 in [3]. □

Chapter 10
Confidence intervals

In this Chapter again, we assume given a parametric model $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ and a sample $X_1, \ldots, X_n \sim f(x;\theta)$, where $\theta \in \Theta$ is the true, but unknown, parameter value. In the previous Chapters we have introduced the notion of an estimator $\hat{\theta}_n$ for θ. Hopefully, if we constructed this estimator in a reasonable way, it should provide a "good guess" for the true parameter value of θ. Now we would like to make this a bit more quantitative: how "good" is this guess, i.e. how large is the error? A common way to answer this is by the notion of confidence interval.

The content of this chapter is based mainly on on Section 6.3.2 and parts of Chapter 9 of [5], see also Chapter 11 in [3].

10.1 Definition and interpretation

Definition 10.1. *Let $\alpha \in (0,1)$ be a fixed number. A $1-\alpha$ confidence interval for θ is an interval $C_n = (a,b)$, where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are random numbers depending only on our observed data $X_1, \ldots, X_n$, such that*

$$\mathbb{P}_\theta(C_n \ni \theta) \geqslant 1 - \alpha. \tag{10.1}$$

Remark 10.2. Note that the interval $C_n = (a,b)$ is random, as a and b are random variables, but depends only on our observations $X_1, \ldots, X_n$, while θ is deterministic but unkown.

Remark 10.3. We commonly use 95% confidence intervals, which correspond to choosing $\alpha = 0.05$.

The information provided by a confidence interval is crucial. There are two possible ways to use it:

1. If one repeats the experiment over and over again, then Property (10.1) guarantees that, a fraction $1-\alpha$ of the times, our confidence interval will contain our true, but unkown, parameter value.
2. If we run a sequence of experiments with unrelated parameters θ_1, θ_2, ... and construct a $1-\alpha$ confidence interval for each, then a fraction $1-\alpha$ of the times, our confidence interval will contain our true, but unkown, parameter value.

Example 10.4. [see Wasserman, example 6.13]

Everyday, newspapers report opinion polls. For instance, one day they might say that 83% of people prefer French wine to Californian wine, adding a statement of the form "this poll is accurate to within 4 points 95% of the time", which means they provide $(79, 87)$ as a 95% confidence interval of the true percentage of people with this opinion. Another day, they will say that 76% of people prefer Italian cheese to French cheese, adding a statement of the form "this poll is accurate to within 2 points 95% of the time", which means they provide $(74, 78)$ as a 95% confidence interval. Etc. That means that, if you read the polls every single day, for about 95% of the polls, the true poll result will be contained in the given confidence interval.

10.2 Exact confidence intervals

Constructing exact confidence intervals can be a complicated task. In this section we show how such confidnece intervals can be constructed in the case of Bernoulli random variables. Our construction will rely on the inequality below, which is not given in classe.

Theorem 10.5. ** Hoeffding's inequality*

Let $Y_1, \dots, Y_n$ be independent random variables such that, for all $i = 1, \dots, n$,

1. $\mathbb{E}(Y_i) = 0$
2. $a_i \leqslant Y_i \leqslant b_i$, *for some* ***deterministic*** $a_i, b_i \in \mathbb{R}$

Let $\varepsilon > 0$. Then, for any $t > 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geqslant \varepsilon\right) \leqslant e^{-t\varepsilon} \prod_{i=1}^{n} e^{t^2 (b_i - a_i)^2 / 8}.$$

A very useful consequence of the above is the following result, wee Theorem 4.5 in [5].

Theorem 10.6. *Let $p \in (0, 1)$ and $X_1, \dots, X_n \sim \mathrm{Ber}(p)$. Then, denoting by $\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$ the sample mean, we have, for all $\varepsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - p| > \varepsilon) \leqslant 2e^{-2n\varepsilon^2}. \tag{10.2}$$

Proof. *We set $Y_i := X_i - p$ and apply Hoeffding's inequality to the Y_i, see p. 67 of [5] for a detailed proof.* □

The inequality (10.2) is often stronger than Chebyshev's inequality (6.1). For instance, taking $n = 100$ and $\varepsilon = 0.2$, inequality (10.2) gives

$$\mathbb{P}(|\bar{X}_n - p| > \varepsilon) \leqslant 2\, e^{-2 \times 100 \times (0.2)^2} = 2\, e^{-8} \approx 6.7 \times 10^{-4},$$

while Chebyshev's inequality would only give the much coraser bound

$$\mathbb{P}(|\bar{X}_n - p| > \varepsilon) \leqslant \frac{\mathbb{V}(\bar{X}_n)}{\varepsilon^2} = \frac{p(1-p)}{n\,\varepsilon^2} \leqslant \frac{1}{4\,n\,\varepsilon^2} = \frac{1}{4 \times 100 \times (0.2)^2} = 6.25 \times 10^{-2},$$

were we used that $p(1-p) \leqslant 1/4$ for all p in the middle inequality.

Application: Confidence intervals for Bernoulli model.

We throw n times a coin with an unkown probability p of hitting Heads and record our results. These are modelled by $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$. We estimate p by the sample mean $\hat{p}_n := \bar{X}_n$, which are also corresponds to the MLE. Then, for any $\alpha \in (0,1)$ setting $\varepsilon_n := \sqrt{\frac{\log(2/\alpha)}{2n}}$, we claim that the interval $C_n := (\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n)$ is a $1-\alpha$ confidence interval for p. Indeed, in view of Hoeffding's inequality,

$$\begin{aligned} \mathbb{P}_p(p \notin C_n) &= \mathbb{P}_p(|\hat{p}_n - p| > \varepsilon_n) \\ &\leqslant 2\,e^{-2n\varepsilon_n^2} \\ &= 2\,e^{-\log(2/\alpha)} \\ &= \alpha, \end{aligned}$$

so that $\mathbb{P}_p(p \in C_n) \geqslant \alpha$ as required. We record below the approximated value of ε_n (up to 2 decimals) for different values of n, for 95 and 97.5% intervals, respectively.

n	ε_n (95% confidence interval)	ε_n (97.5% confidence interval)
10	0.43	0.47
20	0.30	0.33
50	0.19	0.21
100	0.14	0.15
500	0.06	0.07
10^3	0.04	0.05

Note the likeable feature that these confidence intervals are valid whatever the value of p.

Example 10.7. A poll is carried out in Berlin, asking people if they approve transforming the Tempelhofer Feld into a huge shopping mall. A sample of $n = 1000$ people is chosen at random, and asked their opinion. Among them, 570 people approve of the project. Denoting by p the true proportion of the Berlin population approving the project, and modelling the sampled people's opinions by i.i.d. Bernoulli random variables with parameter p, we see from the above chart that $C_n := (0.53, 0.61)$ provides a 95 % confidence interval for p.

Open Question: do you think it reasonable to model sampled people's opinions by i.i.d. random variables, and why?

10.3 Asymptotic confidence intervals

In most situations, constructing an exact confidence interval for θ is too difficult. In such cases we resort to asymptotic confidence intervals. Often we will use normal-based confidence intervals.

10.3.1 Large probability interval for a normal random variable

Let us first state a result for standard normal random variables. Let us denote by Φ^{-1} the quantile function of the standard normal distribution. For $p \in [0,1]$, let us denote $z_p := \Phi^{-1}(1-p)$, so that, if $Z \sim \mathcal{N}(0,1)$, we have $\mathbb{P}(Z \leqslant z_p) = 1-p$.

Lemma 10.8. *Let $Z \sim \mathcal{N}(0,1)$. Then, for all $\alpha \in (0,1)$, we have*

$$\mathbb{P}(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}) = 1-\alpha.$$

Proof. We have

$$\mathbb{P}(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}). \tag{10.3}$$

Note now that, for all $x \in \mathbb{R}$, we have

$$\Phi(-x) = 1 - \Phi(x),$$

indeed using the fact that PDF f_Z of $Z \sim \mathcal{N}(0,1)$ is even, we have

$$\begin{aligned}
\Phi(-x) &= \int_{-\infty}^{-x} f_Z(x)\,\mathrm{d}x \\
&= \int_{x}^{+\infty} f_Z(x)\,\mathrm{d}x \\
&= 1 - \int_{-\infty}^{x} f_Z(x)\,\mathrm{d}x \\
&= 1 - \Phi(x).
\end{aligned}$$

Applying this to (10.3), we get

$$\begin{aligned}
\mathbb{P}(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}) &= \Phi(z_{\alpha/2}) - (1 - \Phi(z_{\alpha/2})) \\
&= 2\,\Phi(z_{\alpha/2}) - 1 \\
&= 2\left(1 - \frac{\alpha}{2}\right) - 1 \\
&= 1 - \alpha.
\end{aligned}$$

□

Recall that

$$\Phi^{-1}(0.95) \approx 1.645,$$

$$\Phi^{-1}(0.975) \approx 1.960.$$

Hence, for $\alpha = 0.1$, $z_{\alpha/2} = \Phi^{-1}\big(1 - \frac{\alpha}{2}\big) = \Phi^{-1}(0.95) \approx 1.645$, so that

$$\mathbb{P}(-1.645 \leqslant Z \leqslant 1.645) \approx 0.9,$$

while for $\alpha = 0.05$, $z_{\alpha/2} = \Phi^{-1}\big(1 - \frac{\alpha}{2}\big) = \Phi^{-1}(0.975) \approx 1.960$, so that

$$\mathbb{P}(-1.960 \leqslant Z \leqslant 1.960) \approx 0.95.$$

In particular we certainly have the following, easy and very useful to remember, fact:

$$\mathbb{P}(-2 \leqslant Z \leqslant 2) \geqslant 0.95.$$

We now show how to use the above Lemma to compute approximate confidence intervals.

10.3.2 Approximate, normal-based, confidence intervals: the general principle

Assume that we have an asymptoticall normal estimator $\hat{\theta}_n$ for θ: that means that

$$\frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1). \tag{10.4}$$

That is, for n large, the distribution of $\frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)}$ is approximately the standard normal distribution. We are then on good tracks to derive a confidence interval for θ using Lemma 10.8, but there is a caveat, which is that $\mathrm{se}(\hat{\theta}_n)$ often depends on the unkown parameter value θ and is thus impossible for us to evaluate. Luckily, we can often define an *estimated standard error* $\hat{\mathrm{se}}(\hat{\theta}_n)$, which depends only on our observations $X_1, \ldots, X_n$ (so we can evaluate it), and which still satisfies the property that

$$\frac{\hat{\theta}_n - \theta}{\hat{\mathrm{se}}(\hat{\theta}_n)} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1). \tag{10.5}$$

In other words, in many cases, (10.4) is still true if we replace the true standard error $\mathrm{se}(\hat{\theta}_n)$ (usually unkown to us) with the estimated standard error $\hat{\mathrm{se}}(\hat{\theta}_n)$ (which can be evluated using our observed data $X_1, \ldots, X_n$). We then have the following result:

Lemma 10.9. *If (10.5) holds, then for all $\alpha \in (0,1)$, the interval*

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\,\hat{\mathrm{se}}(\hat{\theta}_n), \hat{\theta}_n + z_{\alpha/2}\,\hat{\mathrm{se}}(\hat{\theta}_n))$$

is an approximate $1-\alpha$ confidence interval for θ, in the sense that

$$\mathbb{P}(\theta \in C_n) \approx 1 - \alpha,$$

for n large.

Proof. We have, for n large, in view of (10.5),

$$\begin{aligned}\mathbb{P}(\theta \in C_n) &= \mathbb{P}(|\theta - \hat{\theta}_n| \leqslant z_{\alpha/2}\,\hat{\mathrm{se}}(\hat{\theta}_n)) \\ &= \mathbb{P}\left(-z_{\alpha/2} \leqslant \frac{(\hat{\theta}_n - \theta)}{\hat{\mathrm{se}}(\hat{\theta}_n)} \leqslant z_{\alpha/2}\right) \\ &\approx \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\ &= 1-\alpha.\end{aligned}$$

where the last equality follows by Lemma 10.8. □

Example 10.10. For $\alpha = 0.05$, $z_{\alpha/2} \approx 1.960 \leq 2$, so $\hat{\theta}_n \pm 2\,\hat{\mathrm{se}}(\hat{\theta}_n)$ is a 95% confidence interval.

We give below a few examples where the above machinery for constructing approximate confidence intervals can be applied.

10.3.3 Case of the sample mean

Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let us first assume for simplicity that σ is known. Assume that we want to estimate the mean μ of $X_1, \ldots, X_n$. We do so by considering the estimator $\hat{\mu}_n$ defined as the sample mean

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

By the CLT, we know that

$$\sqrt{\frac{n}{\sigma^2}}(\hat{\mu}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1). \tag{10.6}$$

So (10.5) is satisfies, with $\theta = \mu$, $\hat{\theta}_n = \hat{\mu}_n$ and $\hat{\mathrm{se}}(\hat{\mu}_n) = \sqrt{\frac{\sigma^2}{n}}$. In virtue of Lemma 10.9, $C_n := [\mu - \varepsilon_n, \mu + \varepsilon_n]$, with $\varepsilon_n := z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$, is an *approximate* $1-\alpha$ confidence interval for μ, i.e.

$$\mathbb{P}(\hat{\mu}_n \in C_n) \approx 1-\alpha,$$

for n large. Below we give a few values of ε_n, up to 2 decimals, for different values of n and for $\sigma = 1$.

n	ε_n (90% confidence interval)	ε_n (95% confidence interval)
10	0.52	0.62
20	0.37	0.44
50	0.23	0.28
100	0.16	0.20
500	0.07	0.09
10^3	0.05	0.06

Now, what if σ is unknown to us? It then seems natural to try and replace, in (10.6), the true but unkown variance σ^2, by the *sample variance* $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. As it turns out, we can do so without losing the convergence to the standard normal distribution, as the following result show (see Theorem 5.10 in [5])

Theorem 10.11. *There holds the convergence in distribution*

$$\sqrt{\frac{n}{S_n^2}}(\hat{\mu}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Hence, setting $\hat{\text{se}}(\hat{\mu}_n) := \sqrt{\frac{S_n^2}{n}}$, we have

$$\frac{(\hat{\mu}_n - \mu)}{\hat{\text{se}}(\hat{\mu}_n)} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1),$$

and in virtue of Lemma 10.9, we deduce that $C_n := [\mu - \varepsilon_n, \mu + \varepsilon_n]$ is an *approximate* $1-\alpha$ confidence interval for μ, with $\varepsilon_n := z_{\alpha/2}\sqrt{\frac{S_n^2}{n}}$.

10.3.4 Construction via the MLE

We now give a more general class of examples of normal-based confidence intervals, which relies on the MLE.

We assume that our model $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ fulfills Fisher's regularity condition. We consider $X_1, \dots, X_n \sim f(x;\theta)$ with θ unknown, and consider the MLE $\hat{\theta}_n$ for θ. We aim at obtaining a confidence interval for θ. Recall from Theorem 9.14 that $\hat{\theta}_n$ is an asymptotically efficient estimator for θ, i.e. we have

$$\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1),$$

which can equivalently be written as the following approximate equality in law, for n large:

$$\hat{\theta}_n \overset{(d)}{\approx} \mathcal{N}\left(\theta, \frac{1}{I_n(\theta)}\right).$$

In particular, we see that, for n large, $\text{se}(\hat{\theta}_n) \approx \frac{1}{\sqrt{I_n(\theta)}}$. Here we face the caveat that θ is unkown to us and we thus cannot evaluate $I_n(\theta)$. Nevertheless, it turns out that, setting

$$\hat{\text{se}}(\hat{\theta}_n) := \frac{1}{\sqrt{I_n(\hat{\theta}_n)}},$$

we have that

$$\frac{(\hat{\theta}_n - \theta)}{\hat{\text{se}}(\hat{\theta}_n)} = \sqrt{I_n(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1), \tag{10.7}$$

see (9.13) in [5]. We now deduce the following:

Theorem 10.12. *Let $\alpha \in (0,1)$. Then the interval*

$$C_n := (\hat{\theta}_n - z_{\alpha/2}\,\hat{\mathrm{se}}(\hat{\theta}_n), \hat{\theta}_n + z_{\alpha/2}\,\hat{\mathrm{se}}(\hat{\theta}_n))$$

is a $1-\alpha$ confidence interval for θ.

Proof. This follows from (10.7) and Lemma 10.9.

□

Chapter 11
Hypothesis testing

In the previous chapters we saw how to estimate the parameter θ from our observations $X_1, \dots, X_n$. There are situations where, rather than asking "what" the value of θ is, we are testing an hypothesis on the value of θ.

11.1 Definitions

11.1.1 Notations

We assume given a parametric model $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ and a sample $X_1, \dots, X_n \sim f(x;\theta)$, where $\theta \in \Theta$ is the true, but unknown, parameter value. Assume further that we have a partition of Θ into two (disjoint) subsets Θ_0 and Θ_1, so that $\Theta = \Theta_0 \cup \Theta_1$. We would like to test two hypotheses

$$H_0 : \theta \in \Theta_0, \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1.$$

We call H_0 the *null hypothesis,* and H_1 the *alternative hypothesis*.

Example 11.1. Let $X_1, \dots, X_n \sim \mathrm{Ber}(p)$, with $p \in (0,1)$ unknown. This corresponds to throwing a coin n times and recording the outcomes. We would like to test the hypothesis that the coin is fair, i.e.

$$H_0 : p = 1/2, \qquad \text{versus} \qquad H_1 : p \neq 1/2.$$

This corresponds to partitioning $\Theta = (0,1)$ into $\Theta_0 = \{\frac{1}{2}\}$ and $\Theta_1 = (0,1) \setminus \{\frac{1}{2}\}$.

Example 11.2. Let again $X_1, \dots, X_n \sim \mathrm{Ber}(p)$, but we would like now to test the hypotheses

$$H_0 : p < 1/2, \qquad \text{versus} \qquad H_1 : p \geqslant 1/2.$$

This corresponds to partitioning $\Theta = (0,1)$ into $\Theta_0 = \left(0, \frac{1}{2}\right)$ and $\Theta_1 = \left[\frac{1}{2}, 1\right)$.

Definition 11.3. * *Let $\theta_0 \in \Theta$. A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis, while a hypothesis of the form $\theta < \theta_0$ or $\theta > \theta_0$ is called a composite hypothesis. Further, a test of the form*

$$H_0: \theta = \theta_0, \qquad \text{versus} \qquad H_1: \theta \neq \theta_0$$

is called a two-sided test. A test of the form

$$H_0: \theta < \theta_0, \qquad \text{versus} \qquad H_1: \theta \geqslant \theta_0$$

or

$$H_0: \theta > \theta_0, \qquad \text{versus} \qquad H_1: \theta \leqslant \theta_0$$

is called a one-sided test.

For instance, the test of Example 11.1 above is a two-sided test, while the test of Example 11.2 is a one-sided test.

11.1.2 General setup

In practice, to perform a test from our data $X := (X_1, \ldots, X_n)$, we define a function T, called *test statistic*, and a number c, called *critical value*, from which we define the *rejection region*

$$R = \{x: \quad T(x) > c\}.$$

We will then proceed as follows: if our observed data X are such that $X \in R$, i.e. $T(X) > x$, then we shall *reject* the null hypothesis H_0. On the other hand, as long as $X \notin R$, i.e. $T(X) \leqslant x$, we shall *retain* (not reject) the null hypothesis H_0.

Remark 11.4. A hypthesis test is like a legal trial. We assume someone is innocent unless there is strong evidence that they are guilty. Likewise, when performing hypothesis testing, we retain the null hypothesis H_0 unless there is strong evidence against it.

There are two types of errors one could make. Rejecting H_0 when H_0 is true is called a **type I error**. Retaining H_0 when H_0 is false is called a **type II error**.

	Retain H_0	Reject H_0
H_0 True	Correct answer	**type I error**
H_0 False	**type II error**	Correct answer

A major point consists in controlling the error we can make.

Definition 11.5. *We consider a test with rejection region R.*

- *The power function of the test is defined as*

$$\beta(\theta) := \mathbb{P}_\theta(X \in R), \qquad \theta \in \Theta$$

- *The size of the test is defined to be*

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta) \tag{11.1}$$

Thus the power function is the probability that the null hypothesis H_0 is rejected: this probability depends a priori on θ, hence the name **power function**. Recall that H_0 is the hypothesis that $\theta \in \Theta_0$, hence the size of the test defined in (11.1) measures the largest probability of error of type I that may occur (i.e. the largest probability of rejecting H_0 while H_0 is true). Hopefully we would like to construct tests with a size that is under control. This is quantified by the following notion.

Definition 11.6. *Let $\alpha \in (0,1)$. A test is said to be of level α if its size is less than or equal to α.*

Example 11.7. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where $\sigma > 0$ is known and μ is unknown. We want to test H_0: $\mu < 0$ against H_1: $\mu \geqslant 0$. To do so, we denote by $\bar{X}_n$ the sample mean, and consider the statistic $T(X) = \bar{X}_n$. The rejection rejection will be given by

$$R := \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) > c\},$$

where $c \in \mathbb{R}$ is a critical value, and $T(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. The size of this test is $\sup_{\mu<0} \beta(\mu)$, where, for $\mu \in \mathbb{R}$,

$$\begin{aligned} \beta(\mu) &= \mathbb{P}_\mu(X \in R) \\ &= \mathbb{P}_\mu(\bar{X}_n > c) \\ &= \mathbb{P}_\mu\left(Z > \frac{\sqrt{n}\,(c-\mu)}{\sigma} \right) \end{aligned}$$

where $Z := \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. Note that $Z \sim \mathcal{N}(0,1)$, hence, for all $\mu \in \mathbb{R}$, we have the expression $\beta(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right)$, where Φ is the CDF of $\mathcal{N}(0,1)$. Note in particular that β is a non-decreasing function from $\mathbb{R}$ to (0,1). Thus the test constructed above has size

$$\alpha := \sup_{\mu<0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}\,c}{\sigma} \right). \tag{11.2}$$

Suppose now that the value of $\alpha \in (0,1)$ is imposed, e.g. $\alpha = 0.05$, and we are being asked to devise a hypothesis test with size equal to the prescribed value α. To do so we can choose a critical value c such that (11.2) is fulfilled: this is achieved by choosing $c = \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1-\alpha) = \frac{\sigma}{\sqrt{n}} z_\alpha$. So a test of size α is given by rejecting H_0: $\mu < 0$ when $\bar{X}_n > \frac{\sigma}{\sqrt{n}} z_\alpha$, i.e. we reject H_0: $\mu < 0$ when $\frac{\sqrt{n}\bar{X}_n}{\sigma} > z_\alpha$. For $\alpha = 0.05$, $z_\alpha = \Phi^{-1}(0.95) \approx 1.645$.

Numerics: Assume for instance $\sigma = 1$, $n = 100$ and we observe the sample mean $\bar{X}_n$=0.1. Then

$$\frac{\sqrt{n}\bar{X}_n}{\sigma} = \frac{\sqrt{100} \times 0.1}{1} = 1 < 1.645,$$

so we retain H_0. On the other hand, with the same numbers as above but 4 times as many samples, i.e. $\sigma = 1$, $n = 400$ and we still observe a sample mean $\bar{X}_n$=0.1, then

$$\frac{\sqrt{n}\bar{X}_n}{\sigma} = \frac{\sqrt{400} \times 0.1}{1} = 2 > 1.645,$$

so we reject H_0. Note that in both cases above, the observed value of the statistic (the sample mean) was the same, but the decision of rejecting or not was affected by the sample size. Intuitively, with a large sample size, the fact that $\bar{X}_n > 0$ is "more significant" than with a small sample size.

11.2 Examples of tests

We now present a few important examples of tests.

11.2.1 The Z test

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where $\sigma > 0$ is known and μ is unknown. Let $\mu_0 \in \mathbb{R}$. We want to test

$$H_0: \mu = \mu_0 \qquad \text{versus} \qquad H_1: \mu \neq \mu_0.$$

To do so, we consider the sample mean $\bar{X}_n$ as well as $Z := \sqrt{\frac{n}{\sigma^2}}(\bar{X}_n - \mu_0)$. Given $\alpha \in (0,1)$, the size-α Z test rejects H_0 when $|Z| > z_{\alpha/2}$. Note that this test has indeed size α as, under H_0, Z is a standard normal random variable, so the probability that it falsely rejects H_0 is given by

$$\mathbb{P}_{\mu_0}(|Z| > z_{\alpha/2}) = \alpha.$$

Remark 11.8. The size α Z test rejects H_0: $\mu = \mu_0$ when $|\bar{X}_n - \mu_0| > \sqrt{\frac{\sigma^2}{n}} z_{\alpha/2}$, that is when $\mu_0 \notin C_n$ where $C_n = \bar{X}_n \pm \sqrt{\frac{\sigma^2}{n}} z_{\alpha/2}$ is a $(1-\alpha)$ confidence interval for μ.

The Z test is applicable when σ is known, but is is no longer available when σ is unknown. For n large σ^2 may be estimated by the sample variance S_n^2, however for n small this estimation produces an error which needs to be controlled carefully. In such cases we resort to the T test.

11.2.2 The T Test

Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where $\sigma > 0$ and $\mu \in \mathbb{R}$ are both unknown. We still want to test $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$. A key lemma is the following:

Lemma 11.9. *If $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then $\sqrt{\frac{n}{S_n^2}}(\bar{X}_n - \mu)$ follows the t distribution with $n-1$ degrees of freedom, which is the continuous distribution on $\mathbb{R}$ with PDF given by*

$$f(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1 + \frac{t^2}{n-1}\right)^{-n/2}, \qquad t \in \mathbb{R}.$$

Let us consider the test statistic $T := \sqrt{\frac{n}{S_n^2}}(\bar{X}_n - \mu_0)$, where $\bar{X}_n$ and S_n^2 denote the sample mean and sample variance, respectively. Given $\alpha \in (0, 1)$, let further $t_{\alpha/2}$ be the $1 - \alpha/2$ percentile of the t distributions with $n-1$ degrees of freedom. Note that, by symmetry of the PDF f of that distribution, we have $\int_{-t_{\alpha/2}}^{t_{\alpha/2}} f(t)\mathrm{d}t = 1 - \alpha$. The size α T test consists in rejecting H_0 when $|T| > t_{\alpha/2}$. It is easily checked that this test indeed has size α (exercise).

Remark 11.10. The size α T test rejects $H_0: \mu = \mu_0$ when $|\bar{X}_n - \mu_0| > \sqrt{\frac{S_n^2}{n}} t_{\alpha/2}$, that is when $\mu_0 \notin C_n$ where $C_n = \bar{X}_n \pm \sqrt{\frac{S_n^2}{n}} z_{\alpha/2}$ is a $(1-\alpha)$ confidence interval for μ.

Remark 11.11. For n large, $S_n^2 \approx \sigma^2$ and $t_{\alpha/2} \approx z_{\alpha/2}$, so the Z test and the T test are equivalent.

n	t_α (for $\alpha = 0.05$)
5	2.57
10	2.23
50	2.01
100	1.98

11.2.3 The Wald Test

Assume that we have an asymptotically normal estimator $\hat{\theta}_n$ for θ, and that we can estimate its standard error by an estimated standard error $\hat{\mathrm{se}}(\hat{\theta}_n)$ in such a way that

$$\frac{\hat{\theta}_n - \theta}{\hat{\mathrm{se}}(\hat{\theta}_n)} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, 1). \tag{11.3}$$

This is the case, e.g., for the MLE, see Theorem 9.14. Let $\theta_0 \in \Theta$ be a fixed (known) parameter value. Given a prescribed $\alpha \in (0,1)$ we would like to test

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0,$$

with size $\alpha \in (0,1)$. For sufficiently large n, this can be achieved using the Wald test:

Definition 11.12. *For any $\alpha \in (0,1)$, the size α Wald test consists in **rejecting** $H_0 : \theta = \theta_0$ whenever $|W| > z_{\alpha/2}$, where*

$$W := \frac{\hat{\theta}_n - \theta_0}{\hat{\mathrm{se}}(\hat{\theta}_n)}.$$

Proposition 11.13. *The test defined above has indeed, asymptotically, size α, in the sense that*

$$\mathbb{P}_{\theta_0}(|W| > z_{\alpha/2}) \underset{n\to\infty}{\longrightarrow} \alpha.$$

Remark 11.14. Note that in this case $\Theta_0 = \{\theta_0\}$, so the probability above does indeed represent the test size as defined in Definition 11.1.

Proof. In virtue of (11.3), denoting by Z a standard normal variable, we have

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}_{\theta_0}(|W| > z_{\alpha/2}) &= \mathbb{P}(|Z| > z_{\alpha/2}) \\ &= 1 - \mathbb{P}(|Z| \leqslant z_{\alpha/2}) \\ &= \alpha \end{aligned}$$

where the last equality follows from Lemma 10.8. □

Remark 11.15. Recall that the interval C_n given by $\hat{\theta}_n \pm \hat{\mathrm{se}}(\hat{\theta}_n)\, z_{\alpha/2}$ is an approximate $1-\alpha$ confidence interval for θ. Thus, the size α Wald test consists in rejecting H_0 when θ_0 lands outside the $1-\alpha$ confidence interval for θ.

Example 11.16. We poll the opinions of two different groups of population. We sample $n = 1000$ answers from the two groups that we represent by independent Bernoulli random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ with respective parameters p_1 and p_2. We set $\delta := p_1 - p_2$ and wish to test

$$H_0 : \delta = 0 \qquad \text{versus} \qquad H_1 : \delta \neq 0.$$

To do so we consider the estimator

$$\hat{\delta}_n := \bar{X}_n - \bar{Y}_n,$$

which is also the sample mean of the r.v's $X_i - Y_i$. The standard error is given by

$$\mathrm{se}(\hat{\delta}_n) = \sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}},$$

which can be estimated by

$$\hat{\mathrm{se}}(\hat{\delta}_n) = \sqrt{\frac{\bar{X}_n(1-\bar{X}_n) + \bar{Y}_n(1-\bar{Y}_n)}{n}}.$$

The size α Wald test consist in rejecting $H_0 : \delta = 0$ when $|W| > z_{\alpha/2}$ where

$$W = \frac{\hat{\delta}_n - 0}{\hat{\mathrm{se}}(\hat{\delta}_n)} = \frac{\hat{\delta}_n}{\hat{\mathrm{se}}(\hat{\delta}_n)}.$$

Note that the size of the test controls the probability of making a type I error. But it is also important to control the probability of making a type II error, or conversely to say how likely we are to correctly reject H_0 when $H_1 : \theta \neq \theta_0$ is true. This is done by the following result.

Theorem 11.17. *Suppose the true parameter value is equal to $\theta_* \neq \theta_0$. Then the power $\beta(\theta_*)$ of the size α Wald test, i.e. the probability of correctly rejecting the hypothesis $H_0 : \theta = \theta_0$, is approximately equal, for n large, to*

$$\mathbb{P}(|Z - \kappa| > z_{\alpha/2}) \;=\; 1 - \Phi(\kappa + z_{\alpha/2}) + \Phi(\kappa - z_{\alpha/2})$$

where $Z \sim \mathcal{N}(0,1)$ and $\kappa = \frac{\theta_0 - \theta_*}{\hat{\mathrm{se}}(\hat{\theta}_n)}$.

Remark 11.18. If κ is close to 0, $\mathbb{P}(|Z - \kappa| > z_{\alpha/2})$ will not be too far from $\mathbb{P}(|Z| > z_{\alpha/2}) = \alpha$, resulting in a relatively low power for the test. On the other hand, if κ is very far from 0, the probability $\mathbb{P}(|Z - \kappa| > z_{\alpha/2})$ that Z deviates from κ by more than $z_{\alpha/2}$ will be quite close to 1, resulting in a high power. So, in order for the test to be powerful, it is desirable to have κ large, which will be the case if (i) θ_0 is far from θ_* and (ii) the sample size is large enough so that $\hat{\mathrm{se}}(\hat{\theta}_n)$ is small.

Proof. We have

$$\begin{aligned}
\beta(\theta_*) &= \mathbb{P}_{\theta_*}(|W| > z_{\alpha/2}) \\
&= \mathbb{P}_{\theta_*}\left(\left| \frac{\hat{\theta}_n - \theta_*}{\hat{\mathrm{se}}(\hat{\theta}_n)} - \frac{\theta_0 - \theta_*}{\hat{\mathrm{se}}(\hat{\theta}_n)} \right| > z_{\alpha/2} \right) \\
&\underset{n\to\infty}{\approx} \mathbb{P}_{\theta_*}\left(\left| Z - \frac{\theta_0 - \theta_*}{\hat{\mathrm{se}}(\hat{\theta}_n)} \right| > z_{\alpha/2} \right)
\end{aligned}$$

where the last approximate equality follows from the convergence (11.3). The claim follows. □

Example 11.19. Considering again 11.16 and assuming that the true parameter value $\delta_* = p_1 - p_2$ is not 0, so the hypothesis $H_0: \delta = 0$, is false, the power function is approximately equal, for n large, to $\mathbb{P}(|Z - \kappa| > z_{\alpha/2})$ with

$$\kappa = \frac{-\delta_*}{\hat{\mathrm{se}}(\hat{\delta}_n)} = -(p_1 - p_2)\sqrt{\frac{n}{p_1(1-p_1) + p_2(1-p_2)}}.$$

The power will be large if $|\kappa|$ is large, which is achieved when p_1 is far from p_2 and when the sample size n is large.

11.2.4 * The likelihood ratio test (not given in lecture)

The likelihood ratio test is suited when considering a parametric model $\mathcal{F} = \{f(x;\theta): \theta \in \Theta\}$ with parametrised by more than 1 parameter, i.e. where Θ consist of vectors rather than scalars. Let $X_1, \dots, X_n \sim f(x;\theta)$ and let $\mathcal{L}_n: \Theta \to \mathbb{R}$ denote the likelihood function. Assume that we want to test

$$H_0: \theta = \theta_0 \qquad \text{versus} \qquad H_1: \theta \neq \theta_0.$$

Definition 11.20. *The lieklihood ratio statistic is*

$$\lambda = 2\log\left(\frac{\sup_{\theta\in\Theta}\mathcal{L}_n(\theta)}{\mathcal{L}_n(\theta_0)}\right) = 2\log\left(\frac{\mathcal{L}_n(\hat{\theta}_n)}{\mathcal{L}_n(\theta_0)}\right)$$

where $\hat{\theta}_n$ denotes the MLE. Thelikelihood ratio test consists in **rejecting** H_0 if $\lambda > c$, where c is a fixed critical value.

Question. Given a prescribed $\alpha \in (0,1)$, how to choose c so that the above test has size α?

Assume that $\Theta \subset \mathbb{R}^r$ for some $r \geqslant 2$, so that Θ consists in a set of r-dimensional vectors. In particular $\theta_0 \in \mathbb{R}^r$. We claim that the following holds.

Theorem 11.21. *When $n \to \infty$, λ converges in distribution the $\chi(r)$ distribution*

Part III

Bayesian inference

Chapter 12
Introduction to Bayesian inference

12.1 The Bayesian approach

So far we have been addressing *frequentist* inferencewhich relies on three postulates:

F1. Probabilities refers to *limiting relative frequencies:* it is an *objetcive* property of the world,

F2. The paramaters to estimate are fixed, *deterministic*, unknown constants,

F3. Statistical procedures should be designed to have well-defined *long-run frequency* properties.

For instance, in frequentist hypothesis testing, a 95% confidence interval should trap the true parameter value with limiting frequency larger or equal to 95%.

Another approach exists: Bayesian[12.1] inference. It relies on the following three postulates:

B1. Probability describes a *degree of belief*, not a limiting frequency

B2. We can make *probability statements about parameters*

B3. We make inferences about a parameter θ by producing a *probability distribution for* θ.

Thus, in Bayesian inference, we can for instance say that "the probability that it will rain tomorrow in Dahlem is 0,4". This describes our degree of belief of the event "it will rain tomorrow in Dahlem", not a limiting frequency. Taking a more familiar example, let's trhow a coin with unknown parameter $p \in (0, 1)$. In Bayesian inference, we may represent p as a *random variable* (which was not the case in the framework of frequentist inference!). The Bayesian method gives a toolbox to obtain a *posterior distribution* for p, that we can use to make inferences, i.e. produce point estimators, posterior intervals, etc.

12.1. named after Thomas Bayes, c. 1701-1761

12.2 The Bayesian method

Let $\mathcal{F} = \{f(x,\theta)\colon \theta \in \Theta\}$ be a parametric model, where, for each $\theta \in \Theta$, $f(x;\theta)$ is a PDF (for simplicity). How to estimate θ? The Bayesian approach consists in 3 main steps:

1. We *choose* a probability density $f(\theta)$ for θ, called *prior distribution*.
2. Given observations $x = (x_1, \ldots, x_n) \sim f(x;\theta)$, we compute the likelihood function $\mathcal{L}(\theta)$, $\theta \in \Theta$.
3. We update the distribution of θ in accordance with our observations, and obtain the *posterior distribution* $f(\theta; x)$.

In Step 3, the update is performed using Bayes' Theorem:

Theorem 12.1. (Bayes' Theorem for continuous r.v.'s) *Let X, Y be two r.v.'s admitting a joint PDF $f_{X,Y}$. Then, for all x_0 fixed,*

$$f_{X|Y}(x_0|y) = \frac{f_{X(x_0)} f_{Y|X}(y|x_0)}{\int f_{X(x)} f_{Y|X}(y|x)\mathrm{dx}}$$

provided that $f_{Y(y)} > 0$.

Bayesian method:

1. We *postulate* a probability density $f(\theta)$ for θ, called *prior distribution*. **This is a subjective choice**.
2. Given observations $x^n = (x_1, \ldots, x_n) \sim f(x;\theta)$, we compute $f(x^n|\theta)$, which coincides with the likelihood function $\mathcal{L}_n(\theta)$.
3. We compute the ***posterior distribution*** from the prior ditsribution and the likelihood using Bayes' Theorem,:

$$f(\theta|x^n) = \frac{f(x^n|\theta)\, f(\theta)}{\int f(x^n|\tilde{\theta})\, f(\tilde{\theta})\, d\tilde{\theta}}. \tag{12.1}$$

Equation (12.1) can be rewritten as

$$f(\theta|x^n) = \frac{\mathcal{L}_n(\theta) f(\theta)}{\int \mathcal{L}_n(\theta) f(\tilde{\theta})\, d\tilde{\theta}}.$$

Warning 12.2. Do not mix up the prior distribution $f(\theta)$, the likelihood $f(x^n|\theta) = \mathcal{L}_n(\theta)$, and the posterior distribution $f(\theta|x^n)$.

Note 12.3. the quantity

$$c = \int \mathcal{L}_n(\theta) f(\tilde{\theta})\, d\tilde{\theta} \tag{12.2}$$

is a constant (depending on the data x^n) called *evidence*.

So the rule of thumb is

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

Remark 12.4. The evidence can be retrieved from the likelihood $\mathcal{L}_n(\theta)$ and the prior $f(\theta)$ using (12.2). Thus in practice it is sufficient to compute $\mathcal{L}_n(\theta)\, f(\theta)$. The update rule is then

$$f(\theta|x^n) \propto \mathcal{L}_n(\theta)\, f(\theta),$$

i.e.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

12.3 Bayesian point estimators, confidence intervals, and hypothesis tests

Once the posterior distribution $f(\theta|x^n)$ has been otained, it can be used to produce Bayesian point estimators, confidence intervals, and hyopthesis tests, etc.

12.3.1 Bayesian point estimator

Using the posterior distribution $f(\theta|x^n)$, we define the associated Bayesian point estimator $\bar{\theta}_n$ as the *posterior mean*, that is the mean of the posterior distribution:

$$\begin{aligned} \bar{\theta}_n &:= \int \theta\, f(\theta|x^n)\, d\theta \\ &= \frac{\int \theta\, \mathcal{L}_n(\theta) f(\theta)\, d\theta}{\int \mathcal{L}_n(\theta) f(\theta)\, d\theta} \end{aligned}$$

12.3.2 Bayesian confidence interval

We can also define a Bayesian interval estimate. Given a prescribed $\alpha \in (0,1)$, let a, b such that

$$\int_{-\infty}^{a} f(\theta|x^n)\, d\theta = \int_{b}^{+\infty} f(\theta|x^n)\, d\theta = \frac{\alpha}{2},$$

and let C denote the interval (a,b). Then C has the property that

$$\mathbb{P}(\theta \in C|x^n) = \int_{a}^{b} f(\theta|x^n)\, d\theta = 1-\alpha.$$

We call C a $1-\alpha$ posterior interval.

12.4 Examples

Example 12.5. Let $x^n = (x_1, \ldots, x_n)$ be an iid sample of Bernoulli random variables with unknown parameter $p \in (0,1)$. Without any information on p it is tempting to choose a uniform prior distribution for p, that is $f(p) = 1$ for all $p \in (0,1)$ (also called a *flat* prior). Then the posterior is

$$\begin{aligned} f(p|x^n) &\propto \mathcal{L}_n(p)\, f(p) \\ &= (1-p)^{n-s} p^s \end{aligned}$$

where $s = \sum_{i=1}^n x_i$ denotes the number of sucesses. So

$$f(p|x^n) = \frac{(1-p)^{n-s} p^s}{\int (1-\tilde{p})^{n-s} \tilde{p}^s \, d\tilde{p}}. \tag{12.3}$$

Question 1. What distribution is this?

Recall that we denote, for all $\alpha > 0$:

$$\Gamma(\alpha) := \int_0^{+\infty} t^{\alpha-1} e^{-t} \, \mathrm{dt}.$$

Definition 12.6. *Let $\alpha, \beta > 0$. The Beta distribution with parameters α and β is the distribution with PDF*

$$\mathrm{Beta}(x; \alpha, \beta) = \mathbf{1}_{[0,1]}(x) \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

for all $x \in \mathbb{R}$.

Thus, the posterior distribution appearing in (?) is a Beta distribution, namely:

$$\begin{aligned} f(p|x^n) &= \mathrm{Beta}(p; s+1, n-s+1) \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\,\Gamma(n-s+1)} p^s (1-p)^{n-s}. \end{aligned}$$

We write

$$p|x^n \sim \mathrm{Beta}(s+1, n-s+1).$$

With this information we can now compute a Bayes estimator for p. It is given by

$$\begin{aligned} \bar{p}_n &:= \int p\, f(p|x^n)\, dp \\ &= \int p\, \mathrm{Beta}(p; s+1, n-s+1)\, dp. \end{aligned}$$

Exercise 12.1. If $\alpha, \beta > 0$, then

$$\int x\, \mathrm{Beta}(x; \alpha, \beta) = \frac{\alpha}{\alpha+\beta}.$$

So here we get

$$\bar{p}_n = \frac{s+1}{(s+1)+(n-s+1)} = \frac{s+1}{n+2}.$$

Remark 12.7. Recall that the MLE for p is given by

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{s}{n}.$$

In particular $\bar{p}_n \neq \hat{p}_n$ but

$$\frac{\bar{p}_n}{\hat{p}_n} \approx 1$$

for n large.

Chapter 13
Numerical Methods I

Chapter 14
Numerical Methods II

Chapter 15
Data assimilation and filtering

Bibliography

[1] G. Casella and R. L. Berger. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[2] L. Held and D. Sabanés-Bové. *Applied statistical inference*, volume 10. Springer, 2014.

[3] D. Ostwald. *"Statistics for Data Science", lecture slides*. 2020.

[4] R. L. Schilling. *Measures, integrals and martingales*. Cambridge University Press, 2017.

[5] L. Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.