Xavier MILHAUD

Aix-Marseille Université (AMU)

Wednesday 27th November, 2024

Aix Marseille Université

STATISTICS

Master 1 Mathématiques Appliquées et Statistique

Contents

Chapter 1

Reminders in Probability theory

1 Probability space

Let Ω be a set. We denote by $P(\Omega)$ the power set (i.e., the set of all subsets) of Ω.

Definition 1.1 *A sigma-algebra $\mathcal{A}$ on Ω is a collection of subsets of Ω that is non-empty, closed under taking complements, and closed under countable unions.*

In the following, we fix a sigma-algebra $\mathcal{A}$ on Ω.

Definition 1.2 *A probability measure (or law, or distribution) on $(\Omega, \mathcal{A})$ is a function $P : \mathcal{A} \to [0, 1]$ such that $P(\Omega) = 1$ and, for any sequence (A_i) of pairwise disjoint sets in $\mathcal{A}$, we have*

$$P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i).$$

Definition 1.3 *A probability space is a triple $(\Omega, \mathcal{A}, P)$, where Ω is a set, $\mathcal{A}$ is a sigma-algebra on this set, and P is a probability measure on $(\Omega, \mathcal{A})$.*

Example 1.1 *Trivial; finite or countable set; uniform measure on $[0, 1[$ and the Borel sigma-algebra; Cartesian products.*

In the following, we fix a probability space $(\Omega, \mathcal{A}, P)$. The sets in $\mathcal{A}$ will be called events.

Proposition 1.4 *(Probability of an increasing union) Let $(A_i)_{i\geq 1}$ be a sequence of sets in $\mathcal{A}$ such that for all $i \geq 1$, $A_i \subset A_{i+1}$. Then,*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i\to\infty} P(A_i).$$

We can therefore calculate the probability of a union of events when these events are disjoint or nested. In other cases, we use the inclusion-exclusion principle (formule du crible):

Proposition 1.5 *Let $(A_i)_{1\leq i\leq n}$ be a sequence of sets in $\mathcal{A}$. Then,*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} \sum_{1\leq i_1\leq \ldots \leq i_k \leq n} (-1)^{k-1} P(A_{i_1} \cap \ldots \cap A_{i_k}).$$

When $n = 2$, this enables to get the famous formula $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Exercice: give the formula when $n = 3$.

Definition 1.6 *Let A and B be two events such that $P(B) > 0$. The conditional probability of A given B is defined as*

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

The mapping $A \in \mathcal{A} \mapsto P(A \mid B)$ is a probability on $(\Omega, \mathcal{A})$ such that $P(B \mid B) = 1$.

We thus have $P(A \cap B) = P(A \mid B)P(B)$. This formula generalizes to the case of the intersection of n events:

Proposition 1.7 *(Compound probabilities) Let $(A_i)_{1 \leq i \leq n}$ be a sequence of sets in $\mathcal{A}$ such that $P(A_1 \cap \cdots \cap A_{n-1}) > 0$. Then,*

$$P(A_1 \cap \cdots \cap A_n) = P(A_n \mid A_1 \cap \cdots \cap A_{n-1})P(A_{n-1} \mid A_1 \cap \cdots \cap A_{n-2}) \cdots P(A_2 \mid A_1)P(A_1).$$

Proposition 1.8 *(Total probability formula) Let $(A_i)_{i \geq 1}$ be a partition of Ω (that is, a family of pairwise disjoint events such that $\bigcup_i A_i = \Omega$) such that $P(A_i) > 0$ for all $i \geq 1$. For any event A, we have*

$$P(A) = \sum_{i=1}^{+\infty} P(A \mid A_i)P(A_i).$$

2 Random variable

Let $(\Omega, \mathcal{A})$ be a space equipped with a sigma-algebra.

2.1 Definition

Definition 2.1 *(Discrete case) Let E be a finite or countable set, equipped with the sigma-algebra $\mathcal{E} = P(E)$. A random variable X taking values in E is a function $X : \Omega \to E$ such that, for every $x \in E$, $\{X = x\}$ is an event (i.e., an element of $\mathcal{A}$).*

Notation: $\{X = x\} = \{\omega \in \Omega : X(\omega) = x\}$

Proposition 2.2 *Let X be a function from Ω to a finite or countable set E. X is a random variable if and only if for every $A \in \mathcal{E}$, $\{X \in A\}$ is an event.*

Definition 2.3 *(Real case) Let $E = \mathbb{R}$ be the real line, equipped with $\mathcal{B}$, the Borel sigma-algebra. A real random variable is a function $X : \Omega \to \mathbb{R}$ such that, for every $A \in \mathcal{B}$, $\{X \in A\}$ is an event.*

Proposition 2.4 *Let X be a function from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$. X is a random variable if and only if for every open interval I in $\mathbb{R}$, $\{X \in I\}$ is an event.*

Definition 2.5 *(General case) A random variable taking values in a set E equipped with a sigma-algebra $\mathcal{E}$ is a function $X : \Omega \to E$ such that, for every $A \in \mathcal{E}$, $\{X \in A\}$ is an event.*

2.2 Distribution (law)

Definition 2.6 *Let $X : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ be a random variable. The mapping $A \in \mathcal{E} \mapsto \mu(A) = P(X \in A)$ defines a probability on $(E, \mathcal{E})$, called the distribution of X.*

Notation: $X \sim \mu$ means that X follows the distribution μ.

2.2.1 Discrete case

Definition 2.7 *If E is a finite or countable set, the probability mass function f of a probability measure μ on E is the function $f : E \to [0,1]$ defined for all $x \in E$ by $f(x) = \mu(\{x\})$. The probability mass function f_X of a random variable X taking values in E is the probability mass function of its distribution, given by $f_X(x) = P(X = x)$.*

Proposition 2.8 *A probability measure μ on a finite or countable set E is fully characterized by its probability mass function.*

Proposition 2.9 *Let E be finite or countable. A function $f : E \to [0,1]$ is the probability mass function of a probability measure on E if and only if $\sum_{x\in E} f(x) = 1$.*

2.2.2 Real case

Definition 2.10 *The cumulative distribution function of a probability measure μ on $\mathbb{R}$ is the function $F : \mathbb{R} \to [0,1]$ defined for all $x \in \mathbb{R}$ by $F(x) = \mu((-\infty, x])$. The cumulative distribution function F_X of a real random variable X is the cumulative distribution function of its distribution, given by $F_X(x) = P(X \leq x)$.*

Proposition 2.11 *A probability measure on $(\mathbb{R}, \mathcal{B})$ is fully characterized by its cumulative distribution function.*

Theorem 2.12 *A function $F : \mathbb{R} \to [0,1]$ is a cumulative distribution function if and only if it is non-decreasing, right-continuous, $\lim_{x\to-\infty} F(x) = 0$, and $\lim_{x\to+\infty} F(x) = 1$.*

2.3 Density

Definition 2.13 *A probability measure μ on $\mathbb{R}$ has a density function $f : \mathbb{R} \to [0, \infty[$ if, for every Borel set A in $\mathbb{R}$,*

$$\mu(A) = \int_{x\in A} f(x)dx.$$

Notation : $X \sim f$ means that X follows the distribution with density f.

Proposition 2.14 *For a density to exist and be the function f, it is sufficient to verify the above equality for all sets A of the form $(-\infty, a]$, with $a \in \mathbb{R}$.*

Proposition 2.15 *The density, if it exists, fully characterizes the probability measure.*

Proposition 2.16 *If μ has a density and if F is the cumulative distribution function of μ, then $f = F'$ is the density of μ.*

Proposition 2.17 *Any function $f : \mathbb{R} \to [0, \infty[$ such that $\int_{\mathbb{R}} f(x)\,dx = 1$ is a density of some probability measure on $\mathbb{R}$.*

2.4 Examples

Discrete case: Dirac distributions, Bernoulli distributions, binomial distributions, Poisson distributions, geometric distributions.

Real case: exponential distributions, normal distributions.

3 Expectation

Here, $(\Omega, \mathcal{A}, P)$ denotes a probability space, and X is a random variable defined on $(\Omega, \mathcal{A}, P)$ taking values in $\mathbb{R}$. The objective is to give a mathematical meaning to the intuitive definition: 'The expectation of X is the average of the values of X'.

3.1 Indicator function

Definition 3.1 *To every event $A \in \mathcal{A}$, we associate the random variable called the indicator of A, denoted 1_A, which is defined by*

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

The distribution of the variable 1_A is thus the Bernoulli distribution with parameter $P(A)$.

The notation 1 replaces conditions (the 'if' statements) with numerical values (0 or 1). This is very useful in practice: we can add and multiply numbers, which is more difficult to do with 'if' statements. Thus,

Proposition 3.2 *If A and B are two events,*

- $1_{A\cap B} = 1_A \cdot 1_B$.
- $1_{\overline{A}} = 1 - 1_A$. *More generally, if $(A_i)_{i\in I}$ is a sequence of events that forms a partition of Ω, then $\sum_{i\in I} 1_{A_i} = 1$.*
- *If A and B are disjoint, then $1_{A\cup B} = 1_A + 1_B$. More generally, $1_{A\cup B} = 1_A + 1_B - 1_{A\cap B}$.*

Definition 3.3 *Let $(A_i)_{i\in I}$ be a sequence of events that forms a partition of Ω. We say that the variable X decomposes over the partition $(A_i)_{i\in I}$ if and only if X can be expressed as*

$$X = \sum_{i\in I} x_i 1_{A_i}, \tag{1.1}$$

which means that X is constant on each A_i and takes the value x_i on A_i.

Proposition 3.4 *If X takes a finite or countable number of distinct values $(x_i)_{i\in I}$, the events $A_i = \{X = x_i\}$ form a partition of Ω over which X decomposes as*

$$X = \sum_{i\in I} x_i 1_{A_i}.$$

The expression (1.1) is not unique. For example, if A, B, C is a partition of Ω, and $X = 21_A + 1_B + 21_C$, we can also write $X = 21_{A\cup C} + 1_B$. This expression is unique (up to permutation) if the x_i are all distinct.

Definition 3.5 *X is a step variable if and only if there exists a partition $(A_i)_{i\in I}$ (with I having a finite or countable cardinality) over which X decomposes:*

$$X = \sum_{i\in I} \alpha_i 1_{A_i}.$$

In other words, a step variable is a variable whose values form a set with finite or countable cardinality.

3.2 Finite or countable universe

When the universe Ω is finite or countable, the construction of the expectation is simple to understand, as detailed below.

Definition 3.6 *Let X be a positive variable defined on (Ω, A, P). The expectation of X is given by:*

$$E(X) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) \in [0; +\infty]. \tag{1.2}$$

In the case where Ω is a countable set, $E(X)$ is a series with positive terms, which can take the value $+\infty$.
A particular (yet important!) case of (1.2): for any $A \in E$, $E(1_A) = P(A)$.

Proposition 3.7 *Let X and Y be two positive variables defined on (Ω, A, P). The following properties hold:*

1. $E(X) \geq 0$
2. $E(X) = 0$ *if and only if* $X = 0$ *almost surely (a.s.).*
3. *If* $0 \leq X \leq Y$ *almost surely, then* $0 \leq E(X) \leq E(Y)$.
4. *If α and β are positive real numbers, then* $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

Proposition 3.8 *(Writing the expectation as a function of the distribution of X) Let X be a positive variable defined on (Ω, A, P), and let $(x_i)_{i \in I}$ be the distinct values taken by X.*

$$E(X) = \sum_{i \in I} x_i P(X = x_i) \in [0; +\infty]. \tag{1.3}$$

If f is a measurable function from $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ to $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$,

$$E(f(X)) = \sum_{i \in I} f(x_i) P(X = x_i) \in [0; +\infty].$$

Proposition 3.9 *(Case of positive step variables) Let X be a positive stepped variable:*

$$X = \sum_{i \in I} \alpha_i 1_{A_i},$$

where α_i are positive real numbers and $(A_i)_{i \in I}$ form a partition of Ω,

$$E(X) = \sum_{i \in I} \alpha_i P(A_i) \in [0; +\infty]. \tag{1.4}$$

If f is a measurable function from $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ to $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$,

$$E(f(X)) = \sum_{i \in I} f(\alpha_i) P(A_i) \in [0; +\infty].$$

3.3 General case

Except for a few toy cases, the universe Ω is much larger than a set of finite or countable cardinality. In this general framework, an idea of constructing the expectation is provided in the appendix to this chapter. It follows these steps:

- We start by defining the expectation of a positive variable. This expectation is a positive number that can take the value $+\infty$.
- We then define the expectation of a random variable X by

  $$E(X) = E(X^+) - E(X^-),$$

  where X^+ and X^- are the positive and negative parts of X. This expression does not make sense if $E(X^+) = E(X^-) = +\infty$. Therefore, we define $E(X)$ only in the case where $E(|X|) < +\infty$. Since $E(|X|) = E(X^+) + E(X^-)$, this is equivalent to saying that $E(X^+) < +\infty$ and $E(X^-) < +\infty$. We say that the random variable X is integrable.

We focus here on the important properties of expectation.

Proposition 3.10 *(Properties of the expectation operator)*
Let X and Y be random variables defined on $(\Omega, \mathcal{A}, P)$ taking values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

1. *If $X \geq 0$ almost surely, then $E(X) \geq 0$. If $X \geq 0$ almost surely and $E(X) = 0$, then $X = 0$ almost surely.*
2. *If X and Y are positive or integrable and $X \leq Y$ almost surely, then $E(X) \leq E(Y)$.*
3. *If X and Y are integrable, for all real numbers α and β,*

   $$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$

Proposition 3.11 *(Inequalities)*

1. ***Jensen's Inequality*** *If X is integrable and ϕ is a convex function such that $\phi(X)$ is integrable, then*

   $$\phi(E(X)) \leq E(\phi(X)).$$

2. ***Markov's Inequality*** *Let X be a positive random variable. For all $x \in \mathbb{R}^+$,*

   $$P(X \geq x) \leq \frac{1}{x} E(X).$$

3. ***Bienaymé–Chebyshev Inequality*** *Let X be a random variable such that $E(X^2) < \infty$. Define $m = E(X)$ and $\sigma^2 = E((X-m)^2)$ as the mean and variance of X. Then, for any $\epsilon > 0$,*

   $$P(|X - m| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

3.4 Compute the expectancy

Proposition 3.12 *(Expectation in the discrete case) Let X be a random variable on $(\Omega, \mathcal{A}, P)$ whose values $(x_i)_{i \in I}$ form a finite or countable set in $(E, \mathcal{E})$. Let g be a measurable function from $(E, \mathcal{E})$ taking values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that*

$$\sum_{i \in I} |g(x_i)| P(X = x_i) < +\infty.$$

Then,

$$E(g(X)) = \sum_{i \in I} g(x_i) P(X = x_i).$$

Proposition 3.13 *(Expectation in the real case with density function) Let X be a random variable defined on $(\Omega, \mathcal{A}, P)$ taking values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with density f_X. Let g be a measurable function from $(R, B(R))$ to $(R, B(R))$ such that*

$$\int_R |g(x)| f_X(x)\, dx < +\infty.$$

Then,

$$E(g(X)) = \int_R g(x) f_X(x)\, dx.$$

3.5 Specific functions

3.5.1 Generating function

Definition 3.14 *Let X be a random variable taking values in $\mathbb{N}$. The generating function of X, denoted G_X, is the function*

$$G_X : [-1, 1] \to \mathbb{R}, \quad s \mapsto G_X(s) = E(s^X) = \sum_{k=0}^{+\infty} s^k P(X = k).$$

Proposition 3.15 *The properties of the generating function are:*

1. *The generating function characterizes the distribution of X. $P(X = k)$ is the coefficient of s^k in the power series expansion of G_X:*

$$P(X = k) = \frac{G_X^{(k)}(0)}{k!}$$

2. *If $E(|X|) < +\infty$, then $E(X) = \lim_{s \to 1^-} G'_X(s)$.*

3.5.2 Characteristic function

Definition 3.16 *Let X be a random vector taking values in $\mathbb{R}^d$. The characteristic function of X, denoted φ_X, is the function*

$$\varphi_X : \mathbb{R}^d \to \mathbb{R}$$

defined by

$$t \mapsto \varphi_X(t) = E(e^{it \cdot X}) = E(e^{i(t_1 X_1 + \cdots + t_d X_d)}).$$

Proposition 3.17 *(Properties of the characteristic function)*

1. *The characteristic function of X characterizes the distribution of X.*

2. *If $E(|X|) < +\infty$, then φ_X is differentiable, and $\varphi'_X(0) = iE(X)$.*

4 Sequence. Consistency

In the entire chapter, $(\Omega, \mathcal{A}, P)$ denotes a probability space. $(X_n)_{n \geq 0}$ and X are random variables defined on $(\Omega, \mathcal{A}, P)$ taking values in $\mathbb{R}^d$.

4.1 Types of convergence

Definition 4.1 ***Convergence in probability:*** *The sequence $(X_n)_{n\geq 0}$ converges in probability to X (denoted as $X_n \xrightarrow[n\to\infty]{P} X$) if and only if for every $\epsilon > 0$,*

$$\lim_{n\to+\infty} P[\|X_n - X\| > \epsilon] = 0.$$

Definition 4.2 ***Almost sure convergence:*** *The sequence $(X_n)_{n\geq 0}$ converges almost surely to X (denoted as $X_n \xrightarrow[n\to\infty]{a.s.} X$) if and only if:*

$$P\left(\left\{\omega : \lim_{n\to+\infty} X_n(\omega) \text{ exists and equals } X(\omega)\right\}\right) = 1.$$

Definition 4.3 ***Convergence in $L^p(p \geq 1)$:*** *The sequence $(X_n)_{n\geq 0}$ converge to X in L^p (denoted $X_n \xrightarrow[n\to\infty]{L^p} X$) if and only if :*

$$\lim_{n\to+\infty} E\left[\|X_n - X\|^p\right] = 0,$$

for some $p \geq 1$.

Definition 4.4 ***Convergence in distribution:*** *The sequence $(X_n)_{n\geq 0}$ converge in distribution to X (denoted $X_n \xrightarrow[n\to\infty]{d} X$) if and only if, for every bounded continuous function f on $\mathbb{R}^d$ taking values in $\mathbb{R}$, we have:*

$$\lim_{n\to+\infty} E[f(X_n)] = E[f(X)].$$

Remark 4.5 *Convergence in distribution is not strictly a convergence of random variables. This notion of convergence only concerns the sequence of* ***distributions of*** *X_n. Thus, the definition of convergence in distribution makes sense even if the variables X_n are not defined on the same space $(\Omega, \mathcal{A}, P)$, but on different spaces $(\Omega_n, \mathcal{A}_n, P_n)$. However, the space of values taken by these variables is the same for all variables.*

Remark 4.6 *Be careful with the convergence in distribution!*

One must be cautious about certain "automatisms" when working with convergence in distribution. For example, a statement such as:

$$\text{If } X \xrightarrow[n\to\infty]{*} X \text{ and } Y \xrightarrow[n\to\infty]{*} Y, \text{ then } X+Y \xrightarrow[n\to\infty]{*} X+Y.$$

remains true if we talk about convergence in probability, almost sure convergence, or convergence in L^p. However, this is not the case with the convergence in distribution!

4.2 Criteria for convergence in distribution

Proposition 4.7 *(Convergence in distribution and CDF) Let $(X_n)_{n\geq 1}$ and X be random variables taking values in $\mathbb{R}$ with distribution functions $(F_n)_{n\geq 1}$ and F. There is an equivalence between:*

1. *$X_n \xrightarrow[n\to\infty]{d} X$,*
2. *For every real t where F is continuous, $F_n(t) \xrightarrow[n\to\infty]{} F(t)$.*

Corollary 4.1 *Let $(X_n)_{n\geq 1}$ be a sequence of random variables taking values in $\mathbb{R}$, and let a be a real number. If $X_n \xrightarrow[n\to\infty]{d} a$, then $X_n \xrightarrow[n\to\infty]{P} a$.*

For this reason, we can resume our routines when we "compose" two convergences in distribution, as long as one of the two convergences occurs towards a constant. For example, the following statement is true:

Let a be a real number. If $X_n \underset{n\to\infty}{\xrightarrow{d}} a$ and $Y_n \underset{n\to\infty}{\xrightarrow{d}} Y$, then $X_n + Y_n \underset{n\to\infty}{\xrightarrow{d}} a + Y$.

Proposition 4.8 *(Convergence in distribution and characteristic functions) Let $(X_n)_{n\geq 1}$ and X be random vectors taking values in $\mathbb{R}^d$ with characteristic functions $(\varphi_n)_{n\geq 1}$ and φ. There is an equivalence between*

1. $X_n \underset{n\to\infty}{\xrightarrow{d}} X$,
2. *For every vector $t \in \mathbb{R}^d$, $\varphi_n(t) \underset{n\to\infty}{\longrightarrow} \varphi(t)$.*

The implication 1. ⇒ 2. is a direct consequence of the definition of convergence in distribution. It is the implication 2. ⇒ 1. that is useful. The latter is a weakened version of the following theorem due to Paul Lévy:

Theorem 4.9 *(Paul Lévy's theorem) Let $(X_n)_{n\geq 1}$ be random vectors taking values in $\mathbb{R}^d$ with characteristic functions $(\varphi_n)_{n\geq 1}$. If for every vector $t \in \mathbb{R}^d$, $\varphi_n(t)$ converges as $n \to +\infty$ to a function $\varphi(t)$ that is continuous at 0, then the function φ is the characteristic function of a random vector X, and we have $X_n \underset{n\to\infty}{\xrightarrow{d}} X$.*

Remark 4.10 *The theorem by Paul Lévy is a seminal result in probability theory. For instance, the proof of the Central Limit Theorem (CLT) is based on this theorem.*

4.3 Almost sure limit and expectation

The following proposition provides general and important results linking the limit of expectations and the expectation of the limit.

Proposition 4.11 1. ***Monotone Convergence Theorem (Beppo-Levi's):*** *If (X_n) is a sequence of positive random variables that is non-decreasing, and if $X = \lim_{n\to+\infty} X_n$ almost surely, then*

$$\lim_{n\to+\infty} E(X_n) = E(X).$$

2. ***Fatou's Lemma:*** *If (X_n) is a sequence of non-negative random variables, then*

$$E\left(\liminf_{n\to+\infty} X_n\right) \leq \liminf_{n\to+\infty} E(X_n).$$

3. ***Dominated Convergence Theorem of Lebesgue:*** *If (X_n) is a sequence of random variables that converges almost surely to a random variable X, and there exists an integrable random variable Y such that $|X_n| \leq Y$ for all n, then X is integrable and*

$$E\left(|X_n - X|\right) \to 0.$$

In particular,

$$\lim_{n\to+\infty} E(X_n) = E(X).$$

4.4 Relations between the different types of convergence

Proposition 4.12 *Convergence in L^p implies convergence in probability.*

Proof: By Markov's inequality, for all $\varepsilon > 0$ and for all $p \geq 1$,

$$P[\|X_n - X\| \geq \varepsilon] = P[\|X_n - X\|^p \geq \varepsilon^p] \leq \frac{1}{\varepsilon^p} E[\|X_n - X\|^p].$$

Proposition 4.13 *Almost sure convergence implies convergence in probability.*

Proof: Suppose that $X_n \underset{n\to\infty}{\xrightarrow{a.s.}} X$. Noting $Y_n = 1_{\|X-X_n\|>\varepsilon}$, we have

$$P[\|X_n - X\| > \varepsilon] = E(Y_n).$$

Since $Y_n \underset{n\to\infty}{\xrightarrow{a.s.}} 0$ and $|Y_n| \leq 1$ (and the indicator is integrable), we can deduce from the Dominated Convergence Theorem that

$$E(Y_n) \xrightarrow{n\to+\infty} 0.$$

Proposition 4.14 *Convergence in probability implies convergence in distribution.*

In summary,

a.s. convergence $\Rightarrow$ convergence in probability $\Rightarrow$ convergence in distribution

and

convergence in L^p $\Rightarrow$ convergence in probability $\Rightarrow$ convergence in distribution

5 Exercices

See the correction on manuscript papers.

Exercise 1

Let $(A_i)_{i\geq 0}$ be a sequence of events such that $P(A_i) = 1$ for all $i \geq 0$. Show that:

$$P\left(\bigcap_i A_i\right) = 1.$$

Exercise 2

Prove the following formulas: - The inclusion-exclusion formula, - The formula for compound probabilities, - The law of total probability.

Exercise 3

Calculate the cumulative distribution function for the following distributions:

(a) The binomial distribution $B(n, p)$,

(b) The Poisson distribution $P(\lambda)$,

(c) The exponential distribution $\mathrm{Exp}(\lambda)$,

(d) The distribution with density $f(x) = x\exp(-x^2/2)$ for $x \geq 0$ (and zero otherwise).

Exercise 4

Let F be the cumulative distribution function of a distribution. A number m is a median of F if:

$$\lim_{y \to m^-} F(y) \leq \frac{1}{2} \leq F(m).$$

Does such a median always exist? Is it unique?

Exercise 5

Let X be a real-valued random variable with cumulative distribution function F. For all $a, b \in \mathbb{R}$ such that $a \leq b$, express the following quantities in terms of F, a, and b:

$$P(X \in (a,b]), \quad P(X \in (a,b)), \quad P(X \in [a,b]), \quad P(X \in [a,b)), \quad P(X = a).$$

Exercise 6

Let X be a real-valued random variable with a continuous cumulative distribution function F. Let G be a continuous and strictly increasing function on $\mathbb{R}$. Find the cumulative distribution functions of the following random variables:

1. $-X$,
2. X^2,
3. $|X|$,
4. $\sin(X)$,
5. $X^+ = \max(0, X)$,
6. $X^- = -\min(0, X)$,
7. $G^{-1}(X)$,
8. $F(X)$,
9. $G^{-1}(F(X))$.

Exercise 7

Let X be a real-valued random variable with cumulative distribution function F, and $a < b$ two real numbers. Calculate the cumulative distribution functions of the random variables Y and Z defined by:

1. $Y = \begin{cases} X & \text{if } a \leq X \leq b, \\ a & \text{if } X < a, \\ b & \text{if } X > b. \end{cases}$

2. $Z = \begin{cases} X & \text{if } |X| \leq b, \\ 0 & \text{otherwise.} \end{cases}$

Exercise 8

Find (if it exists) a value for c such that the function $f(x)$ defined by the following formula is a valid density function:

$$f(x) = \begin{cases} cx^{-d} & \text{if } x > 1, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{and} \qquad f(x) = ce^x(1+e^x)^{-2}$$

Exercise 9

Let $a > 0$ and $0 < p < 1$. Let X be a random variable with a Poisson distribution $P(a)$. Let Y be a random variable with integer values such that for all integers $0 \leq k \leq n$,

$$P(Y = k \mid X = n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

What is the distribution of Y? And what is the distribution of $Z = X - Y$?

Exercise 10

Let T be a random variable taking values in $\mathbb{N}$ such that:
- For all $n \in \mathbb{N}$, $P(T \geq n) > 0$,
- For all $n, m \in \mathbb{N}$, $P(T \geq n + m \mid T \geq n) = P(T \geq m)$.

(1) T is said to have no memory. Why? Is it reasonable to model the duration of a phone call as a memoryless variable?

(2) Show that T is a geometric random variable with parameter $q = P(T = 0)$ (i.e., for all $n \in \mathbb{N}$, $P(T = n) = (1-q)^n q$).

Exercise 11

(1) Let T be a random variable with values in $\mathbb{R}_+$ such that:

$$\forall t > 0, P(T > t) > 0; \quad \forall t, s > 0, P(T > t + s \mid T > s) = P(T > t).$$

T is said to be memoryless. Why?

(2) Define the function $f(t) = \log(P(T > t))$ for $t > 0$. Show that $f(x) = xf(1)$ for all positive rational x, and then for all positive real x. Conclude that T is an exponential random variable.

(3) Let T be an exponential random variable. What is the distribution of the variable $\lfloor T \rfloor$ (the integer part of T)?

Exercise 12

Let X be a random variable taking the values $3, -1, 1$, and Y be a random variable taking the values $-1, 3$. Write the random variables $X, Y, \exp(X), X^2, X + Y$ in the form $\sum_{i \in I} \alpha_i 1_{A_i}$, where the α_i are all distinct.

Exercise 13

In a die roll, you win 2 euros if the number rolled is prime, and you lose 2 euros if the number rolled is even. Thus, if you roll a 2, you win $2 - 2 = 0$ euros. Let X be the gain obtained.

(1) Write a triplet (Ω, A, P) modeling the experiment.

(2) Write X in the form $\sum_{i\in I} \alpha_i 1_{A_i}$, where the α_i are all distinct.

(3) Calculate $\mathbb{E}(X)$.

Exercise 14

Calculate the expectation of a random variable X for the following distributions:

(1) Binomial distribution $B(n, p)$ (Answer: np),

(2) Poisson distribution $P(\lambda)$ (Answer: λ),

(3) Exponential distribution $\mathrm{Exp}(\lambda)$ (Answer: $\frac{1}{\lambda}$),

(4) Normal distribution $N(\mu, \sigma^2)$ (Answer: μ).

Exercise 15

Let X be a random variable taking values in $\mathbb{N}$. For each integer j, let $p_j = P(X = j)$ and $q_j = P(X > j)$. Show that the expectation of X is:

$$\mathbb{E}(X) = \sum_{j=0}^{\infty} q_j.$$

Exercise 16

For $a > 0$, we define the function $\Gamma(a)$ as:

$$\Gamma(a) = \int_0^{\infty} e^{-x} x^{a-1}\, dx.$$

The Gamma distribution with parameters a and λ (where $a > 0$ and $\lambda > 0$), denoted $G(a, \lambda)$, has the probability density function:

$$f(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}, \quad x > 0.$$

(1) Verify that $\Gamma(a)$ is well-defined for $a > 0$, show that $\Gamma(a + 1) = a\Gamma(a)$, and compute $\Gamma(n)$ for $n \in \mathbb{N}^*$.

(2) Let X be a r.v. with the Gamma distribution $G(a, \lambda)$. Calculate $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

(3) Let Y be a standard normal random variable $N(0, 1)$. Show that Y^2 has the Gamma distribution $G\left(\frac{1}{2}, \frac{1}{2}\right)$. From this, deduce the value of $\Gamma\left(\frac{1}{2}\right)$.

Exercise 17

For natural numbers n, we define the the function $B(n, m)$ as:

$$B(n, m) = \int_0^1 x^n (1 - x)^m\, dx.$$

This is also given by:

$$B(n, m) = \frac{n!m!}{(n + m + 1)!}.$$

Let the probability density function be:

$$f_{n,m}(x) = \frac{(n+m+1)!}{n!m!} x^n (1-x)^m 1_{[0,1]}(x).$$

(1) Show that $f_{n,m}(x)$ is a valid probability density function.

(2) Show that if X has the density $f_{n,m}(x)$, then for any integer p, $\mathbb{E}(X^p) = \frac{(n+p)!(n+m+1)!}{n!(n+m+p+1)!}$.

(3) If X has the density $f_{n,m}(x)$, what is the distribution of $1-X$, and $X/(1-X)$?

Exercise 18

We say that Z follows the Log-normal distribution with parameters m and σ ($\sigma > 0$) if Z is a random variable with values in $\mathbb{R}_+^*$ such that $\ln(Z)$ follows the normal distribution $N(m, \sigma^2)$.

(1) What is the probability density function of Z?

(2) Determine the expectation and variance of Z.

Exercise 19

(1) Calculate the generating function of a random variable X with the binomial distribution $B(n,p)$. (Answer: $(1-p+sp)^n$).

(2) Calculate the generating function of a random variable X with the Poisson distribution $P(\lambda)$. (Answer: $\exp(-\lambda(1-s))$).

Exercise 20

(1) Calculate the characteristic function of a random variable X with the exponential distribution $E(\lambda)$. (Answer: $\frac{\lambda}{\lambda - it}$).

(2) Calculate the characteristic function of a random variable X with the symmetric exponential distribution with parameter λ, and density $\frac{\lambda^2}{2}\exp(-\lambda|x|)$ on $\mathbb{R}$. (Answer: $\frac{\lambda^2}{\lambda^2+t^2}$).

(3) Using the previous result and the Fourier inversion, calculate the characteristic function of a random variable X with the Cauchy distribution with density $\frac{1}{\pi(1+x^2)}$ on $\mathbb{R}$. (Answer: $\exp(-|t|)$).

(4) Calculate the characteristic function of a random variable X with the normal distribution $N(\mu, \sigma^2)$. (Answer: $\exp(it\mu)\exp\left(-\frac{\sigma^2 t^2}{2}\right)$).

Exercise 21

Let X be a random variable with values in $\mathbb{N}$ and its generating function defined on $[-1,1]$ by $G(t) = \frac{t^2}{2-t^2}$. Determine the distributions of X and $Y := X/2$.

Exercise 22

Let $(X_n)_{n\geq 0}$ be a sequence of real random variables such that for all n, X_n has the density $f_n(x) = \frac{n}{\pi(1+n^2x^2)}$. Show that

$$X_n \xrightarrow[n\to\infty]{d} 0$$

Exercise 23

Let $(X_n)_{n\geq 0}$ be a sequence of random variables where $X_n \sim B(n, p_n)$ and $np \to \lambda > 0$. Show that X_n converges in distribution to a Poisson distribution $P(\lambda)$.

Construction of the expectation

Case of Positive Step Variables.
We no longer assume that Ω has a finite or countable cardinality. In this case, expression (1.2) does not make sense (due to the sum over an uncountable number of terms). However, expressions (1.3) and (1.4) continue to be meaningful for any positive variable taking a finite or countable number of values. These expressions then serve as definitions of expectation.

If we choose expression (1.4) as the definition of the expectation of a step variable, we must verify that this expression does not depend on the chosen representation. For example, if $X = 21_A + 1_B + 21_C = 21_{A\cup C} + 1_B$, this leads to two expressions for $E(X)$:

$$E(X) = 2P(A) + P(B) + 2P(C)$$

and

$$E(X) = 2P(A \cup C) + P(B),$$

which are indeed identical. Expression (1.3) is then a particular case of (1.4).

If we choose expression (1.3) as the definition, we must then show that the expression (1.4) for the expectation of a step variable is valid. This can be done by using the linearity of expectation, provided we have demonstrated this linearity from definition (1.3)... Try to convince yourself that this linearity holds true by considering a particular case.

With this definition, propositions 3.7, 3.8 and 3.9 remain true.

Case of positive random variables.
In the case where the set of values taken by X is not countable, neither expression (1.3) nor (1.4) makes sense. To define $E(X)$, we use a limiting process. We approximate X by a decreasing sequence (X_n) of step variables: for example,

$$X_n = \sum_{k=0}^{+\infty} \frac{k+1}{2^n} 1_{\{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\}}.$$

Since $0 \leq X_{n+1} \leq X_n$, we have $E(X_{n+1}) \leq E(X_n)$. The sequence $(E(X_n))_{n\geq 1}$ is a decreasing sequence of positive reals. Therefore, this sequence has a limit, which we define as $E(X)$:

$$E[X] \quad = \quad \lim_{n\to\infty} E[X_n].$$

To verify that this is indeed a definition, we must ensure that this limit does not depend on the decreasing sequence of step variables that approximate X.

General case: real-valued random variables.
Let X be a real-valued variable that is not necessarily positive. For any real number x, we define the positive part x^+ of x by $x^+ = \max(x, 0)$, and its negative part x^- by $x^- = \max(-x, 0)$. These

are two positive numbers, and we have $x = x^+ - x^-$ and $|x| = x^+ + x^-$. We can therefore write:

$$X = X^+ - X^-$$

and

$$|X| = X^+ + X^-.$$

We have seen how to define $E(X^+)$ and $E(X^-)$. We then define $E(X)$ by setting $E(X) = E(X^+) - E(X^-)$ as soon as this expression makes sense, i.e., as long as we do not encounter the indeterminate form $+\infty - \infty$. This is particularly the case when $E(X^+) < +\infty$ and $E(X^-) < +\infty$, or equivalently when $E(|X|) = E(X^+) + E(X^-) < +\infty$.

Chapter 2

Random vectors

1 Random vectors, joint distribution, marginals

In the entire chapter, $(\Omega, \mathcal{A}, P)$ denotes a probability space,

Let $(E_i)_{1\leq i\leq d}$ be a family of sets, each equipped with a σ-algebra $\mathcal{E}_i$. For every $i \in \{1, \dots, d\}$, we define a function $X_i : (\Omega, \mathcal{A}) \to (E_i, \mathcal{E}_i)$.

Definition 1.1 *The family $X = (X_i)_{1\leq i\leq d} : \Omega \to \times_{1\leq i\leq d} E_i$ is a random vector if, for every $i \in \{1, \dots, d\}$, X_i is a random variable taking values in E_i.*

Definition 1.2 *The product σ-algebra $\mathcal{E} = \bigotimes_{i\in I} \mathcal{E}_i$ on $E = \times_{1\leq i\leq d} E_i$ is the smallest σ-algebra on E that makes these two points equivalent for any family $(X_i)_{1\leq i\leq d}$:*

1. *$(X_i)_{1\leq i\leq d}$ is a random vector;*
2. *$X = (X_i)_{1\leq i\leq d}$ is a variable taking values in $(E, \mathcal{E})$ in the sense of definition 2.6 of Chapter 1.*

It is the smallest σ-algebra containing all sets of the form $\times_{1\leq i\leq d} A_i$, where $A_i \in \mathcal{E}_i$ for all $i \in \{1, \dots, d\}$.

Definition 1.3 *The joint distribution of the random vector X is the distribution of X viewed as a random variable taking values in $(E, \mathcal{E})$. The marginal distributions of X are the distributions of X_i, for $i \in I$.*

Proposition 1.4 *If we know the joint distribution μ of X, then we know all the marginal distributions μ_i of X. The reverse is false.*

1.1 Discrete case

In this subsection, the E_i are finite or countable. Therefore, E is finite or countable, and the distribution of X is characterized by its mass function $f : E \to [0, 1]$ defined for every $x = (x_1, \dots, x_d) \in E$ by

$$f(x_1, \dots, x_d) = P(X_1 = x_1, \dots, X_d = x_d)$$

Proposition 1.5 *(Marginal distributions in the discrete case). Let j be fixed between 1 and d. If f is the mass function of X, and f_j is the mass function of X_j, then for all $x_j \in E_j$, we have*

$$f_j(x_j) = \sum_{x_1\in E_1} \cdots \sum_{x_{j-1}\in E_{j-1}} \sum_{x_{j+1}\in E_{j+1}} \cdots \sum_{x_d\in E_d} f(x_1, \dots, x_d).$$

1.2 Real case

In this subsection, we assume that all E_i are equal to $\mathbb{R}$ and $E = \mathbb{R}^d$.

Definition 1.6 *The cumulative distribution function of X is the function $F : \mathbb{R}^d \to [0,1]$ defined for all $x = (x_1, \ldots, x_d)$ by*

$$F(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

Proposition 1.7 *(Cumulative distribution function of marginal distributions). Let j be fixed between 1 and d. Let F be the cumulative distribution function of X and F_j that of X_j. Then, for all $x_j \in \mathbb{R}$, we have*

$$F_j(x_j) = \lim_{x_1 \to +\infty} \ldots \lim_{x_{j-1} \to +\infty} \lim_{x_{j+1} \to +\infty} \ldots \lim_{x_d \to +\infty} F(x_1, \ldots, x_d).$$

and these limits commute.

Definition 1.8 *The measure μ has the joint density function $f : \mathbb{R}^d \to [0, +\infty[$ if, for every Borel set $A \subset \mathbb{R}^d$,*

$$\mu(A) = \int \ldots \int_{x \in A} f(x)\, dx.$$

Proposition 1.9 *If μ has a density f, and if F is the cumulative distribution function of μ, then almost everywhere (a.e.) we have*

$$f = \frac{\partial^d F}{\partial x_1 \cdots \partial x_d}.$$

Proposition 1.10 *(Densities of marginals). Let j be fixed between 1 and d. If the distribution of X has a joint density f, then the distribution of X_j has a density f_j, referred to as the marginal density, and for (almost) every $x_j \in \mathbb{R}$,*

$$f_j(x_j) = \int \cdots \int_{(x_1, \ldots, x_{j-1}) \in \mathbb{R}^{j-1}} \int \cdots \int_{(x_{j+1}, \ldots, x_d) \in \mathbb{R}^{d-j}} f(x)\, dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_d$$

1.3 Others

In a random vector, some coordinates can be discrete random variables, while others can be real-valued random variables.
Examples: Complex random variables.

1.4 Computations

It is often of interest to study the distribution of a vector Y defined from a vector X with a known distribution, through a relation $Y = \varphi(X)$. One can then attempt to calculate the cumulative distribution function or the characteristic function of Y, as these two functions characterize the distribution of Y. More generally, one may try to express $\mathbb{E}(h(Y))$ for a bounded measurable function h (methode de la fonction muette). If X has density f, then we have

$$\mathbb{E}(h(Y)) = \mathbb{E}(h \circ \varphi(X)) = \int \ldots h \circ \varphi(x) f(x)\, dx_1 \cdots dx_d.$$

We then seek to express this integral in the form

$$\int h(y) g(y)\, dy.$$

If this is possible, g is the density of Y. In this context, we recall the change of variables formula.

Proposition 1.11 *(Change of variables). Let $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be a diffeomorphism (i.e., a differentiable bijection with differentiable inverse φ^{-1}) that associates $x = (x_1, \dots, x_d)$ with $y = (y_1, ..., y_d) = \varphi(x) = (\varphi_1(x), \dots, \varphi_d(x))$. We denote by $J_g(x)$ the Jacobian matrix of g, defined by*

$$J_g(x) = \begin{pmatrix} \frac{\partial g_1(x)}{\partial x_1} & \dots & \frac{\partial g_1(x)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_d(x)}{\partial x_1} & \dots & \frac{\partial g_d(x)}{\partial x_d} \end{pmatrix}.$$

Then, we have

$$\int \dots \int_{\mathbb{R}^d} h \circ \varphi(x)\, dx_1 \cdots dx_d = \int \dots \int_{\mathbb{R}^d} h(y) \left|\det J_{\varphi^{-1}}(y)\right| dy_1 ... dy_d.$$

Example 1.1 *(In dimension 1, thus with random variable and not random vectors). Take the lognormal distribution, defined from the Gaussian one. Indeed, the lognormal distribution Y is such that:*

$$Y = e^X,$$

where X follows a Gaussian distribution, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$.

Here, $\phi(x) = e^x$, then $\phi^{-1}(x) = \ln(x)$. It follows that $(\phi{-}1(x))' = 1/x$. We then obtain

$$f_Y(y) = f_X(\ln(y)) \times (\phi{-}1(y))' = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}} \frac{1}{y} = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}.$$

We recognize the density of the lognormal distribution.

More generally, still considering the transformation $Y = \phi(X)$, and looking in terms of the CDF:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\phi(X) \leq y) = P(X \leq \phi^{-1}(y)) = F_X(\phi^{-1}(y)), \quad \textit{then} \\ f_Y(y) &= (F_Y(y))', \quad \textit{we thus have} \\ f_Y(y) &= (F_X(\phi^{-1}(y)))' = (\phi^{-1}(y))'\, f_X(\phi^{-1}(y)) \end{aligned}$$

The latter formula applied to the lognormal example above obviously gives the same result.

2 Covariance matrix

2.1 Variance, covariance

Let X and Y be two real random variables such that $\mathbb{E}(X^2) < +\infty$ and $\mathbb{E}(Y^2) < +\infty$ (in this case, we say that X and Y are square-integrable).

Definition 2.1

$$\text{var}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

$$\text{cov}(X, Y) := \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Proposition 2.2 *(Properties of the variance)*

1. $\text{var}(X) \geq 0$.
2. $\text{var}(X) = 0$ *if and only if X is almost surely constant (equal to $\mathbb{E}(X)$).*
3. *For all $m \in \mathbb{R}$, $\text{var}(X) \leq \mathbb{E}((X - m)^2)$. Thus, $\mathbb{E}(X)$ is the best prediction (in the sense of quadratic risk) one can make for the variable X.*

4. *For any real* α*,* $\operatorname{var}(\alpha X) = \alpha^2 \operatorname{var}(X)$*.*
5. $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2 \operatorname{cov}(X, Y)$*.*

2.2 Extension to random vectors

Let $X = (X_1, \ldots, X_d)$ be a random vector such that $\mathbb{E}(\|X\|^2) < +\infty$.

Definition 2.3 *The covariance matrix of* X *is the square matrix of dimension* d*, denoted by* Γ*, defined by*

$$\forall i, j \in \{1, \ldots, d\}, \quad \Gamma_{i,j} = \operatorname{cov}(X_i, X_j).$$

Proposition 2.4 *(Properties of the covariance matrix)*

1. *For any vector* $\alpha \in \mathbb{R}^d$*, we have* $\alpha^T \Gamma \alpha = \mathbb{E}\left[\left(\sum_{i=1}^d \alpha_i (X_i - E[X_i])\right)^2\right] \geq 0$*.*
2. Γ *is a symmetric positive semidefinite matrix.*
3. Γ *has an eigenvalue of zero if and only if the variables* X_i *are linearly dependent, i.e., there exist real numbers* $\alpha_0, \alpha_1, \ldots, \alpha_d$ *such that*

$$\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_d X_d = 0 \quad \textit{almost surely.}$$

4. *If* A *is an* $l \times d$ *matrix, and* $Y = AX$*, the covariance matrix of* Y *is the square matrix of dimension* l *given by*

$$A \Gamma A^T, \quad \textit{where } A^T \textit{ is the transpose matrix of } A.$$

3 Independence

3.1 Independence of events

Definition 3.1 *Two events* A *and* B *are independent if and only if* $P(A \cap B) = P(A)P(B)$ *(or equivalently, when* $P(B) > 0$*,* $P(A|B) = P(A)$*).*

Definition 3.2 *A family of events* $(A_i)_{i \in I}$ *is a family of independent events if and only if for every finite subset of indices* $J \subset I$*, we have*

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j).$$

3.2 Independence of random variables

Let X and Y be two random variables defined on $(\Omega, \mathcal{A}, P)$ with values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$, respectively.

Definition 3.3 X *and* Y *are independent if and only if for all* $A \in \mathcal{E}$ *and* $B \in \mathcal{F}$*, we have*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Proposition 3.4 *Measurable functions of independent variables are independent. Let* f *be a measurable function from* $(E, \mathcal{E})$ *to* $(E', \mathcal{E}')$*. Let* g *be a measurable function from* $(F, \mathcal{F})$ *to* $(F', \mathcal{F}')$*. If* X *and* Y *are independent, then* $f(X)$ *and* $g(Y)$ *are independent.*

Proposition 3.5 *Case of discrete variables. Independence in terms of the mass function. If the sets E and F are finite or countable, the variables X and Y are independent if and only if for all $x \in E$ and $y \in F$, we have*

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

(the mass function of the pair is the product of the mass functions of each variable)

Proposition 3.6 *Case of real variables. Independence and non-correlation. Suppose $E = F = \mathbb{R}$ and $\mathcal{E} = \mathcal{F} = \mathcal{B}(\mathbb{R})$, and that the variables X and Y are square-integrable. If X and Y are independent, then $\mathrm{cov}(X, Y) = 0$. The converse is false.*

Proposition 3.7 *Case of real variables. Independence in terms of the cumulative distribution function. Suppose $E = F = \mathbb{R}$ and $\mathcal{E} = \mathcal{F} = \mathcal{B}(\mathbb{R})$. The variables X and Y are independent if and only if for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$, we have*

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

(the cumulative distribution function of the pair is the product of the cumulative distribution functions of each variable)

Proposition 3.8 *Case of real variables. Independence in terms of the characteristic function. Suppose $E = F = \mathbb{R}$ and $\mathcal{E} = \mathcal{F} = \mathcal{B}(\mathbb{R})$. Let $\varphi(X, Y)$, φ_X, and φ_Y be the characteristic functions of the pair (X, Y) and of the variables X and Y, respectively. The variables X and Y are independent if and only if for all $\theta = (\theta_x, \theta_y) \in \mathbb{R}^2$, we have*

$$\varphi(X, Y)(\theta) = \varphi_X(\theta_x)\varphi_Y(\theta_y).$$

(the characteristic function of the pair is the product of the characteristic functions of each variable)

Proposition 3.9 *Case of real variables with density. Independence in terms of the density. Suppose $E = F = \mathbb{R}$ and $\mathcal{E} = \mathcal{F} = \mathcal{B}(\mathbb{R})$, and that the pair (X, Y) has density $h(x, y)$. Let f and g be the densities of X and Y (which exist according to Proposition 1.10). The variables X and Y are independent if and only if for almost every $(x, y) \in \mathbb{R}^2$, we have*

$$h(x, y) = f(x)g(y).$$

(the density of the pair is the product of the densities of each variable)

All of the above generalizes to more than two random variables.

4 Special case: Gaussian vectors

Definition 4.1 *Let $X = (X_1, \dots, X_d)$ be a random vector of dimension d. X is a Gaussian vector if and only if every linear combination of the coordinates of X is a Gaussian real variable: for every $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$,*

$$\langle \alpha, X \rangle := \alpha_1 X_1 + \cdots + \alpha_d X_d$$

is a Gaussian variable.

Proposition 4.2 *Characteristic function of a Gaussian vector. X is a Gaussian vector if and only if there exists $m \in \mathbb{R}^d$ and a positive symmetric matrix Γ of dimension $d \times d$, such that the characteristic function φ_X of X is given by*

$$\varphi_X(t) = \exp(i\langle t, m \rangle) \exp\left(-\frac{1}{2}\langle t, \Gamma t \rangle\right).$$

Here, m is the mean vector of X and Γ is its covariance matrix:

$$m := \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_d) \end{pmatrix}, \quad \Gamma := \begin{pmatrix} \mathrm{var}(X_1) & \cdots & \mathrm{cov}(X_1, X_d) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(X_d, X_1) & \cdots & \mathrm{var}(X_d) \end{pmatrix}.$$

We denote this by $X \sim \mathcal{N}_d(m, \Gamma)$.

Proposition 4.3 *Linear transformation of a Gaussian vector. Let X be a Gaussian vector of dimension d: $X \sim \mathcal{N}_d(m, \Gamma)$. Let A be a matrix of dimension $l \times d$ and b a vector of dimension l. The vector $AX + b$ is a Gaussian vector with distribution $\mathcal{N}_l(Am + b, A\Gamma A^T)$.*

Proposition 4.4 *Independence in a Gaussian vector. Let $X = (X_1, \dots, X_d)$ be a Gaussian vector. For all index subsets I, J of $\{1, \dots, d\}$, $(X_i)_{i \in I}$ and $(X_j)_{j \in J}$ are independent if and only if* $\mathrm{cov}(X_i, X_j) = 0$ *for all $i \in I$ and all $j \in J$.*

Proposition 4.5 *Density of a Gaussian vector. Let X be a Gaussian vector of dimension d: $X \sim \mathcal{N}_d(m, \Gamma)$. X has a density on $\mathbb{R}^d$ if and only if Γ is positive definite. In this case, the density of X is given by*

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Gamma)}} \exp\left(-\frac{1}{2}(x-\mu)\Gamma^{-1}(x-\mu)^T\right).$$

5 Conditioning

5.1 Conditional distribution

Let (X, Y) be a pair of random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively. The goal of this section is to answer the question: if I observe the variable Y, what can I say about the distribution of X?

5.1.1 Independence or not

Proposition 5.1 *The variables X and Y are independent if and only if for all $y \in F$, the conditional distribution of X given $Y = y$ does not depend on y:*

$$\forall A \in E, \quad P(X \in A | Y = y) = P(X \in A).$$

5.1.2 Discrete conditioning variable

For every possible value y of the variable Y such that $P(Y = y) > 0$, the conditional distribution of X given $Y = y$ is the probability on $(E, \mathcal{E})$ that associates with an event $A \in \mathcal{E}$ the following:

$$P(X \in A | Y = y) = \frac{P(X \in A; Y = y)}{P(Y = y)}. \tag{2.1}$$

This depends on the fixed value y.

If X is a discrete variable, it is characterized by its conditional distribution that associates to $x \in E$ the probability $P(X = x | Y = y)$. We have

$$P(X = x | Y = y) = \frac{P(X = x; Y = y)}{P(Y = y)} = \frac{P(X = x; Y = y)}{\sum_{z \in E} P(X = z, Y = y)}.$$

5.1.3 Continuous conditioning (with density)

In this case, the variable Y also has a density, and for all $y \in \mathbb{R}$, $P(Y = y) = 0$. Similarly, for every event $A \in \mathcal{B}(\mathbb{R})$ and every $y \in \mathbb{R}$, $P(X \in A; Y = y) = 0$. Thus, the expression (2.1) is an indeterminate form of the type "0/0". To define the conditional probability, we must use a limit, and define $P(X \in A|Y = y)$ as the limit as ϵ approaches 0 of $P(X \in A|Y \in [y - \epsilon; y + \epsilon])$. If $f(x, y)$ denotes the density of (X, Y), then we have

$$P(X \in A|Y = y) = \frac{\int_{x\in A} f(x,y)\,dx}{\int_{\mathbb{R}} f(x,y)\,dx}.$$

Thus, the conditional density of X given $Y = y$ is given by the expression

$$f_{X|Y=y}(x) = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dx}.$$

5.2 Conditional expectation

Let X be a random variable or a random vector and Y be a real random variable. We seek to define the conditional expectation of Y given X as a mean of Y with X fixed. Here is the general definition.

Definition 5.2 *Suppose that $E(|Y|) < +\infty$. The conditional expectation of Y given X, denoted $E(Y|X)$, is an integrable random variable Z such that*

1. *there exists a measurable function η such that $Z = \eta(X)$,*
2. *for any bounded measurable function ζ,*

$$E[Y\zeta(X)] = E[Z\zeta(X)].$$

Proposition 5.3 *Two random variables Z_1 and Z_2 that satisfy the two conditions above are equal almost surely. The function η is unique (up to almost everywhere) on the support of the distribution of X. We then denote $\eta(x)$ by $E(Y|X = x)$.*

Note that, quite often, $P(X = x) = 0$ and thus $E(Y|X = x) = \frac{E(Y 1_{X=x})}{P(X=x)}$ is not well-defined. This formula is valid if the distribution of X is discrete (see the proposition below).

If A is an event, then $P(A|X)$ is the conditional expectation of $Y = 1_A$ given X, i.e.,

$$P(A|X) = E(1_A|X) \text{ p.s.}$$

The definition above is not intuitive. However, if we assume that Y is in L^2, we obtain a clearer characterization of $E(Y|X)$.

Proposition 5.4 ***Characterization in the L^2 case.*** *If $E(Y^2) < \infty$, then $E(Y|X)$ is the best approximation of Y by a function of X in the least squares sense. In other words, for any square-integrable random variable $Z' = \eta'(X)$,*

$$E\left(Y - E(Y|X)\right)^2 \le E\left(Y - \eta'(X)\right)^2.$$

Proposition 5.5 ***Calculation in the case where X is discrete.*** *Assume that the set $\{x_i, i \in I\}$ of values taken by X is finite or countable. Then,*

$$E(Y|X) = \sum_{i\in I} E(Y|X = x_i)\,\mathbb{I}_{X=x_i},$$

where $E(Y|X = x_i)$ is given by

$$E(Y|X = x_i) = \frac{E(Y\mathbb{I}_{X=x_i})}{P(X = x_i)}.$$

Proposition 5.6 ***Calculation in the case where*** (X, Y) ***has a density.*** *Assume that the couple (X, Y) has a density $f(x, y)$ on $\mathbb{R}^2$. Let g be the density of the marginal distribution of X, that is,*

$$g(x) = \int f(x, y)\, dy \quad \textit{for all } x.$$

Then, for any x,

$$E(Y|X) = \int y \frac{f(X, y)}{g(X)}\, dy \quad a.s.$$

This can be rewritten as:

$$E(Y|X = x) = \int y \frac{f(x, y)}{g(x)}\, dy = \int y f_{Y|X=x}(y)\, dy.$$

Theorem 5.7 *(Desintegration). Let Y be a random variable independent of X, and let T be a random variable that is a function of X. If $\phi(T, Y)$ is integrable, then*

$$E[\phi(T, Y)|X] = \Phi(T) \quad p.s.$$

where, for all t,

$$\Phi(t) = E[\phi(t, Y)].$$

Proposition 5.8 ***Properties of the conditional expectation of*** Y ***given*** X. *Whenever the random variables within the conditional expectations are integrable, we have:*

1. $E[E(Y|X)] = E(Y)$
2. *If $Y \geq 0$ almost surely, then $E(Y|X) \geq 0$ almost surely; if $Y_1 \leq Y_2$ almost surely, then $E(Y_1|X) \leq E(Y_2|X)$ almost surely.*
3. *If α and β are two constants, then $E(\alpha Y_1 + \beta Y_2|X) = \alpha E(Y_1|X) + \beta E(Y_2|X)$ almost surely.*
4. *If X and Y are independent, then $E(Y|X) = E(Y)$ almost surely.*
5. *(Conditional Jensen's inequality) If ϕ is a convex function such that $\phi(Y)$ is integrable, then*

$$\phi(E(Y|X)) \leq E[\phi(Y)|X] \quad \textit{almost surely.}$$

6. *If T is a function of X, then $E(TY|X) = TE(Y|X)$ almost surely.*
7. *(Conditional Beppo-Levi) If (Y_n) is a non-decreasing sequence of positive random variables, and if $Y_\infty = \lim_{n\to\infty} Y_n$, then $\lim_{n\to\infty} E(Y_n|X) = E(Y_\infty|X)$ almost surely.*
8. *(Conditional Fatou's lemma) If (Y_n) is a sequence of positive random variables, then*

$$E\left[\liminf_{n\to\infty} Y_n|X\right] \leq \liminf_{n\to\infty} E(Y_n|X) \quad \textit{almost surely.}$$

9. *(Conditional Lebesgue's dominated convergence) If (Y_n) is a sequence of random variables that converges almost surely to Y_∞ and if there exists an integrable random variable Z such that for all n, $|Y_n| \leq Z$, then*

$$E[\,|Y_n - Y_\infty| \quad | \quad X] \to 0 \quad \textit{almost surely.}$$

5.3 The case of Gaussian random vectors

Proposition 5.9 *Let X and Y be two random vectors (of dimensions n and d respectively) such that (X, Y) is a Gaussian vector of dimension $n + d$, with mean vector $(m_X, m_Y)^T$ and covariance matrix*

$$\Gamma = \begin{pmatrix} \Gamma_{XX} & \Gamma_{XY} \\ \Gamma_{YX} & \Gamma_{YY} \end{pmatrix}.$$

We assume that Γ_{XX} is positive definite. The conditional distribution of the vector Y given X is the distribution of a Gaussian vector with mean

$$m_Y + \Gamma_{YX}(\Gamma_{XX})^{-1}(X - m_X)$$

and covariance matrix

$$\Gamma_{YY} - \Gamma_{YX}(\Gamma_{XX})^{-1}\Gamma_{XY}.$$

In particular, $E(Y|X) = m_Y + \Gamma_{YX}(\Gamma_{XX})^{-1}(X - m_X)$ is an affine transformation of X.

6 Asymptotic results about empirical mean

The two theorems in this section provide information about the behavior of empirical means as n approaches infinity. Therefore, they are at the heart of statistical theory in the context of large sample limits.

Throughout the rest of the chapter, we consider a sequence of independent and identically distributed random vectors $(X_n)_{n\geq 1}$ (abbreviated as "i.i.d." for "independent and identically distributed"), taking values in $\mathbb{R}^d$.

We assume that $E(\|X_1\|) < +\infty$ and denote m as the mean vector: $m = E(X_1)$. When $E(\|X_1\|^2) < +\infty$, we can also define the covariance matrix Γ of X_1.

We set

$$S_n := X_1 + \cdots + X_n \in \mathbb{R}^d \qquad \text{and} \qquad \bar{X}_n := \frac{1}{n}(X_1 + \cdots + X_n) = \frac{S_n}{n} \in \mathbb{R}^d,$$

where the latter is called "the empirical mean" of X_i.

Proposition 6.1 *Mean and covariance matrix of the empirical mean.*

1. *If $E(\|X_1\|) < +\infty$, then $E(\bar{X}_n) = m$.*

2. *If $E(\|X_1\|^2) < +\infty$, the covariance matrix $\bar{\Gamma}_n$ of $\bar{X}_n$ is equal to $\frac{\Gamma}{n}$. In particular,*

$$E\left[\|\bar{X}_n - m\|^2\right] = Trace(\bar{\Gamma}_n) = \frac{1}{n}\,Trace(\Gamma),$$

and

$$\bar{X}_n \underset{n\to\infty}{\overset{L^2}{\longrightarrow}} m.$$

Thus, the empirical average converges in quadratic mean (mean square) to the theoretical mean. This last result can be significantly improved:

6.1 Strong Law of large numbers (SLLN)

Theorem 6.2 *Let $E(\|X_1\|) < +\infty$. Then,*

$$\bar{X}_n \xrightarrow{p.s.} m \quad \text{as } n \to +\infty.$$

Thus, outside a set of probability zero, a realization of the empirical average (i.e., the computation of the empirical average from a sample) converges to the theoretical mean.

Remark 6.3 *There also exists the weak version of this law, where the convergence is not an almost sure convergence. This is a convergence in probability. However, the almost sure consistency ensures the convergence in probability, which explains why we often work with the SLLN.*

6.2 Central Limit Theorem (CLT)

The following result specifies how fast the convergence of the empirical average to the theoretical mean occurs and what the statistical distribution of the fluctuations of the empirical average around the theoretical mean is. It thus allows for the construction of confidence intervals for the theoretical mean.

Theorem 6.4 *Central Limit Theorem.*
Assume that $E(\|X_1\|^2) < +\infty$. Then,

$$\sqrt{n}(\bar{X}_n - m) \overset{d}{\to} Z \sim N(0, \Gamma) \quad \text{as } n \to +\infty.$$

The central limit theorem generalizes to sufficiently smooth functions of the empirical average (thanks to the Delta method):

Corollary 6.1 *Let f be a function from $\mathbb{R}^d$ to $\mathbb{R}$ of class C^1. Then,*

$$\sqrt{n}(f(\bar{X}_n) - f(m)) \overset{d}{\to} Z \sim N(0, \langle \nabla f(m), \Gamma \nabla f(m) \rangle).$$

7 Exercices

Exercise 1

What is the density of the pair (X, Y) whose joint cumulative distribution function is given by:

$$F(x,y) = \begin{cases} 0 & \text{if } x < 0, \\ (1 - e^{-x}) \left(\frac{1}{2} + \frac{1}{\pi}\arctan(y)\right) & \text{if } x \geq 0. \end{cases}$$

Exercise 2

For the two functions F defined below, which one (or both) is a valid cumulative distribution function for a pair (X, Y)?

$$F(x,y) = \begin{cases} 1 - e^{-x-y} & \text{if } x, y \geq 0, \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad F(x,y) = \begin{cases} 1 - e^{-x} - xe^{-y} & \text{if } 0 \leq x \leq y, \\ 1 - e^{-y} - ye^{-x} & \text{if } 0 \leq y \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

If the answer is yes, what are the two marginal cumulative distribution functions?

Exercise 3

Let (X, Y, Z) be the real-valued random triplet whose density function is defined, when non-zero, by

$$f(x,y,z) = (y-x)^2 \exp\left(-(1+z)(y-x)\right),$$

if $0 \leq x \leq 1$, $y \geq x$, and $z \geq 0$.
Let $U = X$, $V = (Y - X)$, and $W = Z(Y - X)$. What is the distribution of (U, V, W)?

Exercise 4

Let $a > 0$ and $0 < p < 1$. Let X be a random variable following the Poisson distribution $\mathcal{P}(a)$. Let Y be a random variable with integer values such that, for all integers $0 \leq k \leq n$,

$$P(Y = k \mid X = n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Define $Z = X - Y$. What is the distribution of the pair (Y, Z)?

Exercise 5

Two dice, one red and one blue, are rolled, and their respective outcomes are recorded. Let the events be:

- A: "The red die shows an even number."
- B: "The blue die shows an even number."
- C: "The sum of the numbers is even."

Compute the probabilities of A, B, and C. Verify that A, B, and C are pairwise independent but not mutually independent.

Exercise 6

Let A and B be two incompatible events. Show that A and B are independent if and only if $P(A)$ or $P(B)$ is zero.

Exercise 7

Independent trials with a constant probability of success p, where $0 < p < 1$, are conducted until a fixed number m of successes is obtained. Let X denote the number of trials required.

1. Compute the probability distribution of X.
2. Show that $\mathbb{E}\left[\frac{m-1}{X-1}\right] = p$ and that $\mathbb{E}\left[\frac{m}{X}\right] \neq p$ (assume $m > 1$).

Exercise 8

Let X_1, X_2, X_3, X_4 be independent random variables following the Bernoulli distribution $\mathcal{B}(p)$. Consider the matrix

$$M = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix},$$

and let D be its determinant. Compute $\mathbb{E}[D]$.

Exercise 9

Two numbers are drawn at random from the set $\{-2, -1, 0, 1, 2\}$, and their product is denoted by X. Compute $\mathbb{E}[X]$:

1. when the draw is with replacement;
2. when the draw is without replacement.

Exercise 10

Let X_1 and X_2 be two independent Bernoulli random variables with respective parameters p_1 and p_2. Let $Y_i = 2X_i - 1$ for $i = 1, 2$. Are Y_1 and Y_2 independent? Are Y_1 and Y_1Y_2 independent?

Exercise 11

Let X be a random variable uniformly distributed on the interval $[-2, 1]$. Define $Y = |X|$ and $Z = \max(X, 0)$. Find the cumulative distribution functions of Y and Z. Are Y and Z random variables with densities? Are Y and Z independent?

Exercise 12

Let (X, Y) be a pair of random variables taking values in $\{-2, 0, 1\} \times \{-\frac{1}{2}, 0, 1\}$. The table below gives the values of $P[X = x, Y = y]$ for the different values of x and y.

y / x	-2	0	1
-1/2	1/10	a	0
0	3/10	0	3/10
1	1/10	1/10	0

1. What is the value of a?
2. What is the distribution of Y?
3. Is X independent from Y?

Exercise 13

Let (X, Y) be a pair of random variables with the density function:

$$f_{(X,Y)}(x, y) = \exp(-y) \cdot \mathbf{1}_{0 \le x \le y}.$$

1. Verify that $f_{(X,Y)}$ is indeed a density function.
2. Are the variables X and Y independent?
3. Are the variables X and $Y - X$ independent?
4. What is the distribution of the pair $(Y - X, X/Y)$?

Exercise 14

1. If $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$ are two independent variables, what is the distribution of $X + Y$?
2. If $X \sim P(\lambda)$ and $Y \sim P(\mu)$ are two independent variables, what is the distribution of $X + Y$?

Exercise 15

Two independent telephone exchanges receive daily call volumes X and Y, which follow Poisson distributions with parameters λ and μ, respectively.

1. What is the probability that the total number of calls received by both exchanges is at most 3, given $\lambda = 2$ and $\mu = 4$?
2. What is the probability that $X = k$ given that $X + Y = n$, for two integers k and n? What is the distribution in this case?

3. Assuming $\lambda = 2$ and $\mu = 4$, and knowing that the total number of calls received by both exchanges is 8, what is the probability that the first exchange received k calls? For which value of k is this conditional probability maximized?

Exercise 16

Let (X, Y) be a pair of random variables. The table below gives the values of $P[X = x, Y = y]$ for the different values of x and y.

X / Y	0	1	2
0	1/9	2/9	0
1	0	1/9	2/9
2	2/9	0	1/9

1. Show that $\text{cov}(X, Y) = 0$, but X and Y are not independent.
2. Determine the moment-generating functions G_X, G_Y, G_{X+Y} of X, Y, and $X + Y$, and verify that $G_X G_Y = G_{X+Y}$.
3. Calculate $\mathbb{E}(X \mid Y)$.

Exercise 17

Let X be a real-valued random variable with a symmetric distribution (i.e., X and $-X$ have the same distribution). Let ε be an independent random variable such that $P(\varepsilon = 1) = p = 1 - P(\varepsilon = -1)$, where $p \in (0, 1)$.

1. Find the distribution of εX.
2. Under what condition on p is the covariance between X and εX equal to zero? In this case, are X and εX independent?
3. Let $Y = 1_{X>0} - 1_{X<0}$. Find the distribution of Y and XY. Calculate the covariance between $|X|$ and Y. Are these two variables independent?

Exercise 18

Let X be a 3-dimensional random vector following a $\mathcal{N}(0, \Sigma)$ distribution, where

$$\Sigma = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Find a vector of the form $a(X)$, where a is a linear map from $\mathbb{R}^3$ to $\mathbb{R}^3$, such that the components are independent.

Exercise 19

Let $X \sim \mathcal{N}(0, 1)$, and let ε be an independent random variable with $P(\varepsilon = 1) = P(\varepsilon = -1) = \frac{1}{2}$.

1. Find the distribution of $Y = \varepsilon X$.
2. Is the pair (X, Y) jointly Gaussian?
3. Calculate $\text{cov}(X, Y)$.

Exercise 20

1. Show that the matrix
$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$
is a covariance matrix if and only if $\rho \in [-1, 1]$.

   In what follows, we assume that this condition is fulfilled, and we consider (X, Y) a centered Gaussian vector with covariance matrix Σ.

2. Find the value a_0 of a that minimizes $\mathbb{E}[(Y - aX)^2]$.

3. What are the distributions of X, $Y - a_0X$, and $(X, Y - a_0X)$? Show that X and $Y - a_0X$ are independent.

4. Calculate $\mathbb{E}(Y \mid X)$.

Exercise 21

Let $\Sigma = (\sigma_{ij})$ be a 3×3 symmetric matrix defined by $\sigma_{ii} = 1$, and $\sigma_{ij} = a$ for $i \neq j$.

1. For which values of a is Σ a covariance matrix?

2. Assume that Σ is a covariance matrix, and let (X, Y, Z) be a centered Gaussian vector with covariance matrix Σ. Calculate $\mathbb{E}(Z \mid (X, Y))$ and $\mathbb{E}(Z \mid X + Y)$.

3. Calculate $\mathbb{E}[X^2Y^2]$ and $\mathbb{E}[X^2Y^4]$.

Exercise 22

Let M and X be two random variables. Assume M follows a Gaussian distribution, and for all $t \in \mathbb{R}$,
$$\mathbb{E}\left[e^{itX} \mid M\right] = \exp\left(itM - \frac{\sigma^2 t^2}{2}\right).$$

1. Show that (X, M) is a Gaussian vector.

2. Calculate $\mathbb{E}(X \mid M)$.

Exercise 23

The number Y of breakdowns of a machine during a year is related to the age X (in years) of the machine. For a fixed age x, the distribution of Y is a Poisson distribution with parameter $\mu_Y(x) = 1 + \ln(x)$:
$$P[Y = k \mid X = x] = \frac{\mu_Y(x)^k}{k!} \exp(-\mu_Y(x)), \quad \forall k \in \mathbb{N}.$$

For this type of machine, the distribution of ages is given by:

x	1	2	3	4
$p_X(x)$	0.1	0.2	0.3	0.4

1. What does the parameter $\mu_Y(x)$ represent? Comment on its expression as a function of x.
2. What is the joint distribution of the pair (X, Y)?
3. I decide to buy this second-hand machine from a dealer who cannot specify its age. What is the distribution of the annual number of breakdowns?

Exercise 24

The height X (in meters) of a species of trees found in a given forest depends on the age Y (in years) of the tree. For a given age y, the height follows a normal distribution, where the mean $\mu_X(y)$ and the variance $\sigma_X^2(y)$ depend on the age. For trees with ages y ranging from 20 to 150 years, we have:

$$\mu_X(y) = 40\,(1 - \exp(-y/50))\,, \quad \sigma_X(y) = \mu_X(y)/10.$$

It is assumed that the distribution of ages follows an exponential distribution with parameter 0.02: $Y \sim \text{Exp}(0.02)$.

1. Comment on the assumptions made (choice of distributions, choice of parameters, etc.).
2. What is the proportion of trees with ages between 20 and 150 years?
3. What is the distribution of the height of trees aged between 20 and 150 years?
4. Suppose we have a tree with a measured height of x meters and know its age is between 20 and 150 years. What can we infer about its age?
5. This exercise is actually a statistics problem. What is the parameter of interest in the last question? Show that this exercise can be framed as a Bayesian model. What is the prior distribution, the statistical model, and the posterior distribution in this case?

Exercise 25

An urn contains an unknown proportion p of white balls. We perform a series of n draws of a ball with replacement, and estimate p by the proportion Y_n of white balls obtained in the n draws. Show that for any $\epsilon > 0$,

$$P(|Y_n - p| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}.$$

From this, deduce a condition on n for the approximation to provide a value of p to within 0.01 with a probability greater than or equal to 95%.

Exercise 26 (Chernoff Inequality).

Let ϵ be a symmetric random variable ($P(\epsilon = 1) = P(\epsilon = -1) = \frac{1}{2}$) and $(\epsilon_i)_{i\geq 0}$ be a sequence of independent random variables with the same distribution as ϵ. We define $S_n = \sum_{i=1}^n \epsilon_i$.

1. Using Markov's inequality on $\exp(tS_n)$, show that for all $t > 0$ and all $\delta \geq 0$,

$$P(S_n > \delta) \leq \exp(-\delta t) \cdot ch(t)^n.$$

2. By optimizing this inequality with respect to t, deduce that for all $\delta \geq 0$,

$$P\left(\frac{S_n}{n} > \delta\right) \leq \exp(-nh(\delta)),$$

where the function $h(\delta)$ is defined as:

$$h(\delta) = \begin{cases} \delta \text{arctanh}(\delta) + \frac{1}{2}\log(1-\delta^2) & \text{if } \delta \in [0, 1[\\ +\infty & \text{if } \delta \geq 1. \end{cases}$$

Exercise 27

Let X be a Cauchy random variable, and $(X_n)_{n\geq 0}$ be a sequence of i.i.d. random variables with the same distribution as X. We define $S_n = \sum_{i=1}^n X_i$.

Show that S_n/n converges in distribution.

Exercise 28

Let $(X_n)_{n\geq 0}$ be a sequence of i.i.d. random variables with the density $f(x) = \frac{3}{4}(1-x^2)1_{|x|\leq 1}$. Define $\xi_n = \max(X_1, \dots, X_n)$.

1. Show that $\xi_n \xrightarrow{n\to\infty} 1$ in probability.

2. Show that $\sqrt{n}(1-\xi_n)$ converges in distribution, and give the expression for the density of the limiting distribution.

Exercise 29

We generate 100,000 random numbers $u_1, \dots, u_{100000}$ according to a uniform distribution on $[0,1]$, and calculate their geometric mean:

$$(u_1 u_2 \dots u_{100000})^{1/100000}.$$

This value will be very close to a certain number a. What is this number a?

Exercise 30

Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with a common distribution having density $f(x) = 1_{[1/2,3/2]}(x)$. Define, for all $n \geq 1$, $Y_n = \prod_{i=1}^n X_i$.

1. Show that the sequence $(Y_n)_{n\geq 0}$ converges almost surely to 0. (Hint: consider the sequence $\log Y_n$).

2. Does the sequence $(Y_n)_{n\geq 0}$ converge in L^1?

Exercise 31

Let $X_1, \dots, X_{1000}$ be random variables following the uniform distribution on the interval $[0,1]$. Let M be the number of these variables that lie between $1/4$ and $3/4$. Use a normal approximation to determine $P(|M-500| > 20)$.

Exercise 32

Suppose we have tossed a fair coin 10,000 times.

1. Find a symmetric interval around 5000 that contains the number of heads with a probability greater than 0.99.

2. Compare the result with the one obtained using Chebyshev's inequality.

Exercise 33

Let $X \sim \mathcal{P}(\lambda)$ and $Y = \frac{X-\lambda}{\sqrt{\lambda}}$.

1. Show that Y converges in distribution as $\lambda \to \infty$, and determine its limiting distribution.

2. Derive this result as a consequence of the Central Limit Theorem, assuming $\lambda \to \infty$ within $\mathbb{N}$.

3. Show that $e^{-n}\sum_{k=0}^n \frac{n^k}{k!} \to \frac{1}{2}$ as $n \to \infty$.

Chapter 3

Inferential Statistics

1 Introduction

1.1 Forewords

Statistics can be divided into three types of studies: descriptive statistics, inferential statistics, and Bayesian statistics.

Descriptive statistics aim to describe data. They summarize the information contained within the data through simple statistical indicators (such as mean, dispersion, quantiles, etc.) and graphical representations like histograms or bar charts.

The goal of **inferential statistics** is to enable decision-making and forecasting based on observed data. There are two main categories of techniques for achieving this:

- Parameter estimation,
- Statistical tests.

Estimation can be either point-based or set-based, encompassing two primary methods: estimation by the method of moments and estimation by the method of maximum likelihood. Hypothesis tests, on the other hand, are used for decision-making.

Finally, **bayesian statistics** aim to incorporate the experience of observed data to update prior knowledge. This prior knowledge, or expert opinion, is represented by a probability distribution that models uncertainty about the parameter of the probability law governing the phenomenon of interest.

1.2 Parametric versus nonparametric

In general, parametric statistics is distinguished from non-parametric statistics. Parametric statistics assumes the existence of a known model with an unknown parameter that one seeks to find, whereas non-parametric statistics does not rely on the existence of a specified (or known class of) model. A typical example of parametric statistics is assuming a given probability distribution to model some observed random phenomenon, whereas a typical example of nonparametric statistical model is Machine Learning. Of course, pros and cons are associated to parametric and nonparametric statistical models. In general,

- statisticians have developed a lot of theory underlying parametric statistics, with asymptotic results that guarantee good properties of the obtained results;
- it is often less important to have a large dataset when using parametric statistical models than when using nonparametric ones;

- however, nonparametric models allow for more flexibility, since there is no specific constraint that "shapes" the behaviour of the random phenomenon. They are usually called "data-driven" models;
- nonparametric statistical models also rely on tuning parameters which are sometimes not easy to deal with;
- the necessary trade-off between fit-to-data (adequacy) and predictive performance leads to choices in the complexity of models (number of parameters for parametric models, estimator size in nonparametric ones): these choices are often easier to make in the parametric setup.

One of the problems with parametric statistics is the error resulting from a poor model choice. The advantage of non-parametric statistics is that it is not subject to this risk. However, if the observations indeed come from a specific model, the statistical methods that utilize this model are more efficient than those that do not.

In terms of notation, a parametric model is one where it is assumed that the type of distribution of X is known but depends on an unknown parameter θ of dimension greater than or equal to 1. The family of possible probability distributions for X can then be expressed as follows:

$$\mathcal{P} = \{P_\theta;\, \theta \in \Theta \subset \mathbb{R}^d\}.$$

In this context, when we aim to make statistical inference through estimation and testing on θ, we say that we are performing parametric statistics. For the whole course, we will focus hereafter on parametric statistical models.

In contrast, a non-parametric model is one where $\mathcal{P}$ cannot be expressed in this form. In this framework, $\mathcal{P}$ may include, for example, the set of continuous probability distributions on $\mathbb{R}$, or the set of symmetric probability distributions around the origin. The objects on which estimation and testing procedures are based on are no longer parameters of a probability distribution. We may want to estimate quantities such as the mean and variance using empirical estimators. For instance, conducting a test on the value of a mean, or the goodness-of-fit to a distribution, falls under non-parametric statistics.

1.3 Statistical model

The link between statistics and probability theory is established through the statistical model. A statistical model is a mathematical object associated with the observation of data resulting from a random phenomenon. A statistical experiment collects an observation x of a random variable X with values in a space $\mathcal{X}$, where the exact probability distribution P of X is unknown. It is then assumed that P belongs to a family of possible probability distributions $\mathcal{P}$.

Definition 1.1 *The statistical model associated with this experiment is the triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where*

- *$\mathcal{X}$ is the observation space (the set of all possible observations),*
- *$\mathcal{A}$ is the set (σ-algebra) of observable events associated with it, and*
- *$\mathcal{P}$ is a family of possible probability distributions defined on $\mathcal{A}$.*

The model is discrete if $\mathcal{X}$ is finite or countable; in this case, the σ-algebra $\mathcal{A}$ represents all the (sub)sets of $\mathcal{X}$: $\mathcal{A} = P(\mathcal{X})$. This occurs when the random variable X follows a discrete probability distribution.

The model is continuous when $\mathcal{X} \subset \mathbb{R}^p$, and for each P belonging to the set of admissible distributions $\mathcal{P}$, there exists a density with respect to the Lebesgue measure on $\mathbb{R}^p$. Here, $\mathcal{A}$ is the Borel σ-algebra on $\mathcal{X}$ (open intervals on $\mathbb{R}$): $\mathcal{A} = \mathcal{B}(\mathcal{X})$.

It may happen that the observation of the random variable X includes both discrete and continuous elements. In this case, $\mathcal{X}$ and $\mathcal{A}$ are more complex.

Most often, the observed random elements consist of independent random variables, and random variables that have the same distribution are said to be independent and identically distributed (i.i.d.). We denote these as $(X_1, ..., X_n)$. This refers to a **sample**. In this case, denoting by $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ the statistical model of a sample of size 1, $(\mathcal{X}, \mathcal{A}, \mathcal{P})^n$ corresponds to a sample of size n.

Example 1.1 *Consider lifetimes of mechanical parts in automobile. We collect the lifetimes of these parts, assumed to be independent and identically distributed according to an exponential distribution. Thus, we have a statistical model of the following form:* $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \{exp(\lambda), \lambda \in \mathbb{R}^+\})^n$.

Example 1.2 *Now, considering the quality of certain automobile parts, we are interested in the proportion of defective parts. The observation is denoted as $x = (x_1, ..., x_n)$, where x is the realization of the i.i.d. sample $X = (X_1, ..., X_n)$, where $X_i = 0$ if the part functions properly and $X_i = 1$ if it is defective. Thus, we obtain the following statistical model:* $(\{0; 1\}, \mathcal{P}(\{0; 1\}), \{Bern(p), p \in [0, 1]\})^n$.

2 Point-based estimation

2.1 Introduction

In this section, we will assume that we are performing parametric statistics in a unidimensional framework ($d = 1$). We will assume that the data $x_1, ..., x_n$ are independent realizations of the same random variables $X_1, ..., X_n$. Thus, we are in a **sample**, and it is equivalent to assume that the observations $x_1, ..., x_n$ are realizations of independent and identically distributed random variables $X_1, ..., X_n$.

Descriptive statistics allow us to make hypotheses about the type of probability distribution of X_i. More sophisticated techniques, such as goodness-of-fit tests, enable us to validate or reject these hypotheses.

Here, we assume that these techniques have allowed us to select a specific family of probability distributions for the distribution of X_i, with an unknown parameter θ. We will aim to estimate this parameter. Based on the observations $x_1, ..., x_n$, we would like to provide the closest possible approximation to the true unknown value. This approximation can be made in the form of a single value, known as a point estimate, or a set of plausible values, known as a confidence region.

In the following, let $F(x; \theta)$ be the cumulative distribution function of X_i. In the discrete case, we denote $\mathbb{P}(X = x; \theta)$ as the elementary probabilities. In the continuous case, we denote $f(x; \theta)$ as the probability density function.

Example 2.1 *For instance, if $X \sim \mathcal{E}(\lambda)$, then*

$$F(x : \lambda) = (1 - e^{-\lambda x})\, 1_{\mathbb{R}^+}(x) \quad \textit{and} \quad f(x; \theta) = \lambda e^{-\lambda x}\, 1_{\mathbb{R}^+}(x).$$

2.2 Methods

Several estimation techniques exist for a parameter θ. For example, we can mention probability plots or the fact that a probability can be estimated by a proportion. Here, we present the two most well-known estimation methods: the method of moments and the maximum likelihood method.

2.2.1 What is an estimator?

We observe the data $x_1, ..., x_n$. To estimate θ, we only have this information. This means that an estimator of θ will necessarily be a function of these observations.

Definition 2.1 *A statistic t is a function of the observations $x_1, ..., x_n$:*

$$t : \mathbb{R}^n \to \mathbb{R}^m$$
$$(x_1, ..., x_n) \to t(x_1, ..., x_n)$$

We now give some examples:

- the observed empirical mean $\bar{x}_n = (1/n)\sum_i x_i$ is a statistic.
- x_1^* (the lowest observation, once the sample sorted) is a statistic.

We have already seen that $x_1, ..., x_n$ are realizations of $X_1, ..., X_n$. Therefore, $t(x_1, ..., x_n)$ is a realization of $t(X_1, ..., X_n)$. For example, $\bar{X}_n = (1/n)\sum_i X_i$ has a realization $\bar{x}_n$. In terms of notations, we often simplify and denote $t_n = t(x_1, ..., x_n)$ the realized value of $T_n = t(X_1, ..., X_n)$.

Definition 2.2 *An estimator of θ is a statistic T_n which is a random variable taking values in the set of possible values of θ. Being an estimate of θ means that it is a realization t_n of the estimator T_n.*

When the sample x changes, the estimate t_n also changes. However, the expression of the estimator T_n remains the same.

2.2.2 Moment estimation

The basic idea of the method of moments is to estimate a mathematical expectation by an empirical mean, or a variance by an empirical variance. We then proceed by identification after establishing the link between the parameter of the chosen distribution and these quantities.

For example, if the parameter to be estimated is the expectation of the distribution of X_i, such as in a Poisson distribution, we can estimate it using the empirical mean of the sample. More formally, if θ equals the expectation of X_i, then the estimator using the method of moments is:

$$\tilde{\theta}_n = \bar{X}_n.$$

More generally, for θ, if the expectation of X_i is equal to $\phi(\theta)$, where ϕ is an invertible function, then the estimator of θ using the method of moments is:

$$\tilde{\theta}_n = \phi^{-1}(\bar{X}_n).$$

If the distribution of X_i has two parameters θ_1 and θ_2, such as in the case of the Gaussian distribution, we then use the first two moments in this way:

$$(\tilde{\theta}_{1n}, \tilde{\theta}_{2n}) = \phi^{-1}(\bar{X}_n, S_n^2),$$

where $S_n^2 = (1/n)\sum_i (X_i - \bar{X}_n)^2$ is an estimator of the variance of X.

Example 2.2 *Let us try some examples where you have to provide the estimator by the method of moments:*

- *Bernoulli distribution, $X_i \overset{iid}{\sim} \mathcal{B}(p)$: recall that $E[X_i] = p$.*
- *Exponential distribution, $X_i \overset{iid}{\sim} \mathcal{E}(\lambda)$: recall that $E[X_i] = 1/\lambda$.*
- *Gaussian distribution, $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$: recall that $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$.*
- *Gamma distribution, $X_i \overset{iid}{\sim} \mathcal{G}(\alpha, \lambda)$: recall that $E[X_i] = \alpha/\lambda$ and $Var(X_i) = \alpha/\lambda^2$.*

2.2.3 Maximum likelihood estimation

The likelihood function, as its name suggests, is used to assess how plausible a particular model is given the data.

Definition 2.3 *When the observations are all discrete or all continuous, the likelihood function, or likelihood for the sample $x = (x_1, ..., x_n)$, is the function of the parameter θ:*

$$L(\theta; x_1, ..., x_n) = \begin{cases} \mathbb{P}(X_1 = x_1, ..., X_2 = x_2; \theta) & \textit{in the discrete case} \\ f_{(X_1,...,X_n)}(x_1, ..., x_n; \theta) & \textit{in the continuous case.} \end{cases}$$

Using the assumption of independent and identically distributed observations and underlying random variables, the likelihood can also be expressed as follows

$$L(\theta; x_1, ..., x_n) = \begin{cases} \prod_i \mathbb{P}(X_i = x_i; \theta) & \text{in the discrete case} \\ \prod_i f_{X_i}(x_i; \theta) & \text{in the continuous case.} \end{cases}$$

We have moved from a multivariate distribution to a product of univariate distributions. The likelihood function is therefore considered a function of the parameter θ, and depends on the observations $x_1, ..., x_n$.

To provide intuition, let's start with an example.

Example 2.3 *Considering a single observation, let's assume that X_1 follows a Binomial distribution with parameter p, where p is unknown. We have $X_1 \sim \mathcal{B}inom(15, p)$. We observe $x_1 = 5$, and want to estimate p. Theoretically, the likelihhod is given by*

$$L(p; 5) = \mathbb{P}(X_1 = 5; p) = C_{15}^5 p^5 (1-p)^{15-5}.$$

This is the probability to observe 5 when the parameter equals p.
Let's vary p to determine for which value this probability is maximized.

p	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$L(p;5)$	0,01	0,1	0,21	0,19	0,09	0,02	0,003	10^{-4}	2×10^{-7}

For example, we observe that the probability of observing 5 successes when $p = 0.8$ is $1/10,000$. This probability is maximized at $p = 0.3$ in this grid. This result makes sense, given that there are 15 Bernoulli trials. It is therefore much more likely that $p = 0.3$. We thus look for the value that maximizes this likelihood function among all possible values of p.

For optimization, and more specifically for maximizing a function in the search for better parameters by maximizing the likelihood, this is done by finding the value that nullifies the derivative of the likelihood. Furthermore, since the likelihood is defined as a product, it is easier to maximize a sum than a product. Therefore, we will maximize the log-likelihood (denoted l further) in the example above. It is given by:

$$\ln L(p; x_1) = l(p; x_1) = \ln C_{15}^{x_1} + x_1 \ln p + (15 - x_1)\ln(1-p).$$

The derivative of this function thus equals

$$\frac{\partial}{\partial p} l(p; x_1) = \frac{x_1}{p} - \frac{(15 - x_1)}{1-p} = \frac{x_1 - 15p}{p(1-p)}.$$

We easily see that the derivative equals 0 when $p = (1/3)$. The maximum likelihood estimate of p is therefore one-third, with the associated maximum likelihood value $L(1/3; 5) = 0.214$.

Definition 2.4 *The maximum likelihood estimation of θ is the value $\hat{\theta}_n$ of θ that maximizes the likelihood function. The maximum likelihood estimator of θ is the corresponding random variable.*

Therefore, we write $\hat{\theta}_n$ as

$$\hat{\theta}_n = \arg\max_{\theta} L(\theta; x_1, ..., x_n) = \arg\max_{\theta} l(\theta; x_1, ..., x_n).$$

When θ is multidimensional (say with dimension d), and all the partial derivatives exist, the estimator is the solution to a system of equations called the likelihood equations. Hence,

$$\forall j \in \{1, ..., d\}, \qquad \frac{\partial}{\partial \theta_j} L(\theta; x_1, ..., x_n) = 0.$$

Remark 2.5 *A solution to this optimization could a priori be a minimum, but the very nature of a likelihood function leads to a maximum.*

Remark 2.6 *When the system does not have an explicit solution, a numerical method is used. The most well-known and common one is the Newton-Raphson algorithm.*

Example 2.4 *Consider a sample of size n. Getting back to the previous example, we can give the maximum likelihhod estimator for the parameter p of a Bernoulli distribution. Given that the Bernoulli sample has the following likelihood:*

$$L(p; x_1, ..., x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}.$$

The log-likelihood is therefore extremely simple and equals to

$$l(p; x_1, ..., x_n) = \ln p \sum_i x_i + \ln(1-p)(n - \sum_i x_i),$$

which gives the derivative

$$\frac{\partial}{\partial p} l(p; x) = \frac{\sum_i x_i}{p} - \frac{n - \sum_i x_i}{1-p}.$$

Thus, the maximum likelihood estimator (MLE) $\hat{p}_n$ is a random variable, and satisfies

$$\frac{\sum_i X_i}{\hat{p}_n} - \frac{n - \sum_i X_i}{1 - \hat{p}_n} = 0.$$

This comes up with $\hat{p}_n = \bar{X}_n$.

As an exercice, find the maximum likelihood estimators in the following cases:

- *Exponential distribution $\mathcal{E}(\lambda)$;*
- *Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$;*
- *Gamma distribution $\mathcal{G}(\alpha, \lambda)$.*

Remark 2.7 *Show an illustration of maximum likelihood optimization in the software R.*

In general, we obtain two different estimators depending on the method used to estimate the parameter of the distribution. This raises a natural question: which one is better?

2.3 Quality of estimators

We recall that we are particularly interested here in a one-dimensional parameter θ, which belongs to the set of real numbers. The estimators T_n are therefore real random variables. For $\theta \in \mathbb{R}^d$ with $d \geq 2$, all these notions can be generalized, but they become more complex. For example, the concept of variance is replaced by that of a covariance matrix

2.3.1 Bias and variance

An estimator T_n of θ is considered a good estimator if it is close to θ. This requires defining a measure of the deviation between T_n and θ. This deviation is known as the risk of the estimator, which we seek to minimize.

Example 2.5 *Examples of risks: $T_n - \theta$, $|T_n - \theta|$, $(T_n - \theta)^2$, ...*

This quantity is random since it depends on random variables. We are actually interested in the deterministic version of these quantities obtained by taking the expectation. This expectation is taken with respect to the distribution of X_i.

Let us now introduce two important measures for the quality of estimators: the bias and the variance.

Definition 2.8 *The bias of the estimator T_n, denoted by $b(T_n)$, equals*

$$b(T_n) = E[T_n - \theta] = E[T_n] - \theta.$$

We say that the estimator underestimates θ when the bias is negative, otherwise it overestimates θ.

Definition 2.9 *The quadratic risk, also known as the Mean Squarred Error (MSE), equals*

$$MSE(T_n) = E[(T_n - \theta)^2].$$

Take care, the MSE is not the variance of the estimator!

Definition 2.10 *An estimator T_n is said to be unbiased if and only if (iff) $E[T_n] = \theta$. It is biased iff $E[T_n] \neq \theta$.*

The bias measures a **systematic error** in the estimation of θ by T_n.

To understand the components of the mean squared error, we can rewrite it as the sum of two terms:

$$\begin{aligned} MSE(T_n) &= E[(T_n - \theta)^2] = E[(T_n - E[T_n] + E[T_n] - \theta)^2] \\ &= E[(T_n - E[T_n])^2 + 2(T_n - E[T_n])(E[T_n] - \theta) + (E[T_n] - \theta)^2] \\ &= E[(T_n - E[T_n])^2] + 2E[(T_n - E[T_n])(E[T_n] - \theta)] + E[(E[T_n] - \theta)^2] \\ &= Var(T_n) + 0 + (b(T_n))^2 \end{aligned}$$

In other words, the MSE of any estimator is the sum of its variance and its bias to the square! Therefore, if T_n is an unbiased estimator, then its MSE is just the variance of this estimator. Finally, it is in our best interest for an estimator to be unbiased and have low variance.

Remark 2.11 *Between two unbiased estimators, the best one is the one with the smallest variance.*

Remark 2.12 *The variance of an estimator measures its variability when confronted with several similar and independent datasets. An estimator with low variance will produce estimates of θ that are close to each other. To estimate θ accurately, it is important that the estimator's variance is not too large.*

We expect that as the data size increases, the estimation will improve. Indeed, with more information, we should be able to estimate θ with no error in the limit of infinite observations. Therefore, we expect the estimator's risk to asymptotically approach 0. In other words, the estimator T_n converges to θ, in a certain sense.

Definition 2.13 *The estimator T_n converges to θ in quadratic mean if and only if it MSE tends to 0 when n tends to infinity, i.e.*

$$T_n \xrightarrow{QM} \theta \quad \Leftrightarrow \quad \lim_{n\to\infty} E[(T_n - \theta)^2] = 0.$$

Remark 2.14 *The following remarks are important in the light of using estimators in practice.*

- *The convergence in quadratic mean implies the convergence in probability. This is important to show that an estimator T_n is consistent, meaning that it converges in probability to θ.*
- *Disclaimer: nothing guarantees that if T_n is a good estimator of θ, then $\phi(T_n)$ would be a good estimator of $\phi(\theta)$. Indeed, we often have T_n as an unbiased estimator of θ, and $\phi(T_n)$ is a biased estimator of $\phi(\theta)$!*
- *The best estimator among all possible estimators of θ is an unbiased estimator with minimum variance. It does not necessarily exist.*
- *If T_n is unbiased, then* $\quad T_n \xrightarrow{QM} \theta \quad \Leftrightarrow \quad Var(T_n) \underset{n\to\infty}{\longrightarrow} 0.$
- *Some notions that will be detailed further, like sufficient statistics, sometimes allows to determine an unbiased estimator with minimal variance.*
- *To show that an estimator is unbiased with minimal variance, another useful tool is the Fisher information (see coming sections).*

2.3.2 Efficiency

The quantity of information (Fisher information) is a valuable tool for assessing the quality of an estimator. It is defined only under certain regularity conditions. These conditions are linked to the density function $f(x;\theta)$ (and thus the likelihood) of the observations:

- The support of $f(x;\theta)$ does not depend on θ;
- The partial derivative of $f(x;\theta)$ with respect to θ exists almost everywhere (it can fail to exist on a null set, as long as this set does not depend on θ);
- The integral of $f(x;\theta)$ can be differentiated under the integral sign with respect to θ.

Definition 2.15 *For $\theta \in \mathbb{R}$, if the distribution of the observations fulfills some regularity constraints, then it is possible to define the quantity of information (Fisher information) on θ given by the observation $x = (x_1, ..., x_n)$:*

$$I_n(\theta) = Var\left(\frac{\partial}{\partial\theta} \ln L(\theta; X)\right).$$

The variance operator is taken under the distribution of X_i, as well as the expectation for the bias. Knowing that the score $U(\theta; X)$ is defined as the first derivative of the log-likelihood, the Fisher information is therefore the variance of the score. It measures how sensitive is the slope of the likelihood to some changes of the observation sample. The higher the sensitivity, the more Fisher information. This means that what we observe brings a lot of information to infer θ.

Remark 2.16 *There exist two other formulas for the Fisher information:*

- *As the expectation of the score equals 0 ($E\left[\frac{\partial}{\partial\theta}\ln L(\theta;X)\right] = 0$), we have also*

$$I_n(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \ln L(\theta; X)\right)^2\right].$$

- *For the computations, it is sometimes useful to use the next formula which involves the second derivative:*

$$I_n(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln L(\theta; X)\right].$$

The latter formula is more convenient for distributions involving more than 1 parameter (ex: Gaussian).

The significance of the quantity of information lies in the fact that it provides a lower bound for the variance of any unbiased estimator.

Proposition 2.17 *(Inequality of Frechet-Darmois-Cramer-Rao) If the regularity conditions mentioned at the beginning of this section are satisfied, then for all estimator T_n of θ, we have*

$$Var(T_n) \quad \geq \quad \frac{(\frac{\partial}{\partial\theta}E[T_n])^2}{I_n(\theta)}.$$

In case of an unbiased estimator, we have $\frac{\partial}{\partial\theta}E[T_n] = 1$, which means that

$$Var(T_n) \quad \geq \quad \frac{1}{I_n(\theta)}.$$

The latter bound is called the **Cramer-Rao bound**. The variance of any unbiased estimator of θ is therefore necessarily greater than or equal to this bound.
If this bound is large, it is impossible to estimate accurately θ!

Definition 2.18 *The efficiency of an estimator T_n of θ is the quantity*

$$Eff(T_n) = \frac{(\frac{\partial}{\partial\theta}E[T_n])^2}{I_n(\theta)Var(T_n)}.$$

The efficiency is such that $0 \leq Eff(T_n) \leq 1$.

An estimator T_n is said to be efficient if $Eff(T_n) = 1$.
It is asymptotically efficient if $\lim_{n\to\infty} Eff(T_n) = 1$.

Remark 2.19 *From the definition of the efficiency, several remarks can be formulated:*

- *If T_n is an unbiased estimator, we have*

$$Eff(T_n) = \frac{1}{I_n(\theta)Var(T_n)}.$$

- *If T_n is unbiased and efficient, then the variance of T_n equals the Cramer-Rao bound. It is therefore an **unbiased estimator with minimal variance**.*
- *Sometimes there is no efficient estimator existing: in this case, an **unbiased estimator with minimal variance** has a variance strictly greater than the Cramer-Rao bound.*
- *For unbiased estimators, their quality can therefore be studied looking at their efficiency.*

Another important thing to notice is that

$$I_n(\theta) = nI_1(\theta).$$

This is easy to demonstrate. Indeed,

$$\begin{aligned}
I_n(\theta) &= Var\left(\frac{\partial}{\partial\theta}\ln L(\theta;X)\right) = Var\left(\frac{\partial}{\partial\theta}\ln\prod_i f(X_i;\theta)\right)\\
&= Var\left(\frac{\partial}{\partial\theta}\sum_i \ln f(X_i;\theta)\right) = Var\left(\sum_i \frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right)\\
&= \sum_i Var\left(\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right)\\
&= nI_1(\theta).
\end{aligned}$$

This simplification reveals often very useful for computations.

2.3.3 Sufficient and minimal statistic (exhaustivity)

Let us consider a statistical model $(\mathcal{X},\mathcal{A},\mathcal{P}_\theta)$. We aim to obtain as much knowledge as possible about θ from observing $x\in\mathcal{X}$, knowing that x is a high-dimensional vector (with dimension n). Specifically, in cases of large dimension n, it can be useful to summarize the data into lower-dimensional sufficient statistics $t(x)$ with dimension less than n.

Naturally, $t(x)$ will contain less information about θ than x does, but a sufficient statistic retains the entirety of the information about θ.

Definition 2.20 *A statistic t is sufficient for θ if and only if the conditional distribution of $X|T=t$ is independent from θ.*

In terms of interpretation, if the distribution of X conditional on $T=t$ does not depend on θ, it means that the summary statistic $t(x)$ alone suffices to estimate θ, as knowing x provides no additional information about θ. Consequently, we can rely solely on $t(x)$ to estimate θ.

Example 2.6 *Let $(\{0;1\},\mathcal{P}(\{0;1\}),\{Bern(p),p\in[0,1]\})^n$ be our statistical model, for automobile parts. Intuitively, to know the proportion of detfective part in the sample is sufficient to estimate p. We thus expect $\bar{X}_n$ to be a sufficient statistic in this model. Even $\sum_i X_i$ would be a sufficient statistic, as it contains the same quantity of information as $\bar{X}_n$. Let us check whether that works. We know that $T=\sum_i X_i\sim\mathcal{B}inom(n,p)$. Then,*

$$\begin{aligned}
\mathbb{P}(X=x|T=t) &= \mathbb{P}(X_1=x_1,...,X_n=x_n|\sum_i X_i=t) = \frac{\mathbb{P}(X_1=x_1,...,X_n=x_n,\sum_i X_i=t)}{\mathbb{P}(\sum_i X_i=t)}\\
&= \begin{cases} 0 & \text{if } \sum_i X_i\neq t\\ \dfrac{\mathbb{P}(X_1=x_1,...,X_n=x_n)}{\mathbb{P}(\sum_i X_i=t)} & \text{if } \sum_i X_i=t\end{cases}
\end{aligned}$$

Consider the case when $\sum_i X_i=t$:

$$\begin{aligned}
\mathbb{P}(X=x|T=t) &= \frac{\mathbb{P}(X_1=x_1,...,X_n=x_n)}{\mathbb{P}(T=t)} = \frac{\prod_i p^{x_i}(1-p)^{1-x_i}}{C_n^t p^t(1-p)^{n-t}}\\
&= \frac{p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}}{C_n^t p^t(1-p)^{n-t}} = \frac{1}{C_n^t}
\end{aligned}$$

Clearly, this distribution does not depend on θ!

This example shows that verifying this property is not always easy. We then use the following result to characterize sufficiency.

Theorem 2.21 *(Factorization of Fisher-Neyman) t is a sufficient statistic if and only if there exist 2 measurable functions f and g such that: $\forall x \in \mathcal{X}, \forall \theta \in \Theta, L(\theta; x) = g(t(x); \theta)h(x)$.*

Considering the previous example, $L(p; x) = \prod_i p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$. The two functions are thus $g(t(x); p) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$ and $h(x) = 1$. By identification, we can say that $t(X) = \sum_i X_i$ is a sufficient statistic for the estimation of p.

Definition 2.22 *A sufficient statistic $s(x)$ is minimal if, for any sufficient statistic $t(x)$, we can express t as a function of s.*

In practice, we obtain a minimal sufficient statistic by examining the likelihood up to a multiplicative constant. After simplification, identifying which function of the data is necessary and sufficient to calculate it, we obtain the result.

2.3.4 Additional notions

In the case of dependent data, it is necessary to adapt the likelihood formula to account for this. Denote by f all the densities of the random variables X_i, we would have

$$L(\theta; x) = f(x; \theta) = f(x_1; \theta) \prod_{i=2}^{n} f(x_i | \theta, x_1, ..., x_{i-1}).$$

Moreover, the likelihood is invariant under bijective transformations of the dataset. Indeed, if $z = \phi^{-1}(x)$ where ϕ is a diffeomorphism, then the densities have the following form:

$$f_Z(z; \theta) = f_X(\phi(z); \theta) \times \left| \frac{\partial \phi}{\partial z}(z) \right|,$$

with the last term being the Jacobian matrix (which is clearly a constant term).

We now say a few words about the concept of relative likelihood.

Definition 2.23 *When $\underset{\theta'}{\max} L(\theta') < \infty$, we can define the relative likelihood as*

$$RL(\theta) = \frac{L(\theta)}{\underset{\theta'}{\max} L(\theta')}.$$

In the scenario where we cannot seek to estimate θ through optimization, we can consider all values of θ that have a sufficiently large relative likelihood as plausible. Therefore, all the values of θ such that $RL(\theta) > c$, with c a given threshold.

2.4 Properties of estimators

The objective here is to state the properties of estimators obtained by the method of moments and those obtained by the maximum likelihood method.

2.4.1 Moment estimators

We first focus on the empirical mean $\bar{X}_n$. We have already seen that when $\theta = E[X]$, the estimator $\tilde{\theta}_n$ by the method of moments is $\tilde{\theta}_n = \bar{X}_n$. This result is justified by the law of large numbers. Therefore, if $\theta = E[X]$, $\bar{X}_n$ is a consistent estimator of θ as it converges almost surely to θ (asymptotically).

Furthermore, we can also observe that $\bar{X}_n$ is a good estimator of θ, without using the law of large numbers, because it is an unbiased estimator with an asymptotic null variance. Indeed,

$$\begin{array}{rcl} b(\bar{X}_n) & = & E[\bar{X}_n] - \theta = E[X_i] - \theta = 0 \\ Var(\bar{X}_n) & = & (1/n)Var(X_i) \underset{n\to\infty}{\longrightarrow} 0. \end{array}$$

$\bar{X}_n$ is therefore an unbiased estimator that converges in quadratic mean to the expectation θ.

When $\theta = \phi(E[X_i])$ with ϕ a continuous function and given that $\bar{X}_n$ is consistent to estimate $E[X_i]$, then $\phi(\bar{X}_n)$ is still a consistent estimator of θ by continuity.

Now, what about the empirical variance?
Let us consider the natural estimator of $Var(X_i)$, i.e.

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

First, we study the bias:

$$\begin{array}{rcl} E[S_n^2] & = & E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\bar{X}_n^2] = E[X_i^2] - E[\bar{X}_n^2] \\ & = & Var(X_i) + E[X_i]^2 - Var(\bar{X}_n) - E[\bar{X}_n]^2 \\ & = & Var(X_i) + E[X_i]^2 - (1/n)Var(X_i) - E[X_i]^2 \\ & = & \left(1 - \frac{1}{n}\right) Var(X_i) = \frac{n-1}{n}Var(X_i). \end{array}$$

It only remains to multiply by the inverse of this coefficient to obtain an unbiased estimator. Hence,

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

is an unbiased estimator of the variance of X_i.

Furthermore,

$$\begin{array}{rcl} Var(S_n'^2) & = & Var\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) \\ & = & \frac{1}{n(n-1)}\left((n-1)E[(X_i - E[X_i])^4] - (n-3)Var(X_i)^2\right) \underset{n\to\infty}{\longrightarrow} 0. \end{array}$$

The estimator $S_n'^2$ is thus an unbiased estimator of $Var(X_i)$, and converges in quadratic mean towards $Var(X_i)$.

Remark 2.24 *Additionally,*

- *Statistical softwares like* `R` *provides the unbiased version of the estimator of the variance,* $S_n'^2$ *;*
- *There is no result on the quality of S_n as an estimator of the standard deviation of X_i (remember that if a function applies to an estimator with properties like unbiasedness, that does imply the new estimator to keep such properties);*
- *An estimator obtained by the method of moments is asymptotically unbiased and almost surely consistent;*

Concerning the correlation between the two last estimators, $\bar{X}_n$ and $S_n'^2$, it is possible to prove that it equals

$$Cov(\bar{X}_n, S_n'^2) = (1/n)E[(X_i - E[X_i])^3].$$

This means that these random variables are not independent. However, they are asymptotically uncorrelated.

Remark 2.25 *In the case where $X_i \sim \mathcal{N}(\mu, \sigma^2)$, these two estimators are independent.*

2.4.2 Maximum likelihood estimator

A maximum likelihood estimator $\hat{\theta}_n$ is not necessarily unique. Indeed, there can be multiple maxima. It is not necessarily unbiased, nor does it have minimal variance or efficiency. However, it has **very nice asymptotic properties**, provided that the following assumptions are met.

Assumption 2.1 *The data come from i.i.d. replications of $(X_1, ..., X_n)$.*

Assumption 2.2 *Θ is a compact set, with dimension d.*

Assumption 2.3 *The real value θ^0 of the model ($X \sim P_{\theta^0}$) belongs to the interior of Θ (useful for Taylor development). We used the notation θ^0 to avoid confusion since in this course, θ refers both to the underlying true parameter of the distribution of X_i (parametric setup) and some variable of functions (like the likelihood function).*

Assumption 2.4 *There is identifiability issues. A model is identifiable if considering two different parameters leads to two different distributions (models).*

Assumption 2.5 *There exists a neighborhood $v(\theta^0) \in \Theta$ such that:*

- *the log-likelihood is C^3 on $v(\theta^0)$,*
- *$\forall \theta \in v(\theta^0)$, $I(\theta)$ is a positive definite matrix and*

$$\forall r, s, \qquad I_{r,s}(\theta) = E\left[\frac{\partial}{\partial r} \ln L(\theta; X) \frac{\partial}{\partial s} \ln L(\theta; X)\right].$$

- $\forall r, s, t, \qquad \theta \to \frac{1}{n} E\left[\frac{\partial^3}{\partial r \partial s \partial t} \ln L(\theta; X)\right]$ *is uniformly bounded on $v(\theta^0)$.*

Under these conditions, we have:

Proposition 2.26 *If the X_i are independent and identically distributed, depending on the real parameter θ, with the distribution of X_i satisfying the regularity conditions mentioned above, then*

- *$\hat{\theta}_n$ converges almost surely to θ,*
- $\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) \underset{n\to\infty}{\overset{\mathcal{L}}{\approx}} \mathcal{N}(0, 1)$.

The maximum likelihood estimator (MLE) $\hat{\theta}_n$ is thus asymptotically unbiased and efficient, as it reaches the Cramer-Rao bound! The very nice result is about the convergence in distribution, meaning that we know the behaviour of such an estimator.

In general, the MLE is therefore a better estimator than the estimator by the method of moments, since $Var(\hat{\theta}_n) \leq Var(\tilde{\theta}_n)$. At least asymptotically!

Remark 2.27 *Given that the MLE is asymptotically Gaussian, the Delta method can be used to know the behaviour of some transformation of $\hat{\theta}_n$.*
If $\hat{\theta}_n$ is the MLE of θ, then $\phi(\hat{\theta}_n)$ is also the MLE of $\phi(\theta)$. Moreover, if there exist a derivative for ϕ, we have

$$(\phi(\hat{\theta}_n) - \phi(\theta)) \underset{n\to\infty}{\overset{\mathcal{L}}{\approx}} \mathcal{N}\left(0, \frac{\phi^{'}(\theta)^2}{I_n(\theta)}\right)$$

Remark 2.28 *The distribution of the MLE is asymptotically Gaussian. It is centered on θ (meaning unbiasedness), with a variance that depends on θ! Therefore, in practice, one cannot use this result to get an idea about the variance of the MLE! However, using Slutsky's theorem (since $\hat{\theta}_n$ is consistent towards θ), we consider that the variance equals $(I_n(\hat{\theta}_n))^{-1}$ instead of $(I_n(\theta))^{-1}$.*

3 Confidence intervals

3.1 Framework

In the previous section, we estimated θ by a single value, which was a point estimate. We might expect this estimate to be close to θ, but there is little to no chance, especially in the continuous case, that it is exact. Indeed, if the estimator's distribution is continuous (as for the MLE), we know the probability of the estimator being exactly equal to the target point is zero. Therefore, it is natural to estimate θ by proposing a set of reasonable and plausible values, close to the estimator. This is known as a confidence region. We assume from now on that $\theta \in \mathbb{R}$, so the region will be an interval.

Definition 3.1 *A confidence interval with a level $\alpha \in [0, 1]$ for some parameter θ, is a random interval I such that $\mathbb{P}(\theta \in I) = 1 - \alpha$. We denote this confidence interval by $CI_{1-\alpha}(\theta)$.*

Remark 3.2 *Notice in this definition a few things:*

- *α is an error probability, so, if possible, it should be set small.*
- *the confidence interval has random bounds. Indeed, these bounds depend on some functions of the X_i's, which are themselves random variables.*

When writing $\mathbb{P}(\theta \in I)$, θ is an unknown parameter that is constant and $I = [Z_1, Z_2]$ is random (Z_1 and Z_2 are the random bounds). We denote by z_1 and z_2 the realizations of these bounds taking $X = x$ (remind that $X = (X_1, ..., X_n)$, and $x = (x_1, ..., x_n)$). This means that we focus on some particular experience (or simulation). In such a framework, it is wrong to say that θ has a $(1 - \alpha)$ probability to lie between z_1 and z_2. Actually, it is or it is not, and the probability is thus 1 or 0. Conversely, it is correct to say that θ has a $(1 - \alpha)$ probability to lie between Z_1 and Z_2. Indeed, if we simulate independently 100 times the same experience, each one leading to a different observed sample x and bounds z_1 and z_2, θ should lie between the bounds $100 \times (1 - \alpha)$ times approximately.

3.2 Strategy for building a confidence interval

First, it is important to make the difference between an asymptotic versus a non asymptotic CI. An asymptotic CI relies on using some asymptotic properties (like SLLN, CLT) in its building, which is not the case for an exact CI.

It seems logical to propose, as a confidence interval, a set of values centered around the maximum likelihood estimator $\hat{\theta}_n$, as it is asymptotically efficient. Therefore, the confidence interval would take the form $CI = [\hat{\theta}_n - \epsilon, \hat{\theta}_n + \epsilon]$. Then, it remains to determine ϵ such that

$$\mathbb{P}(\theta \in CI) = \mathbb{P}(\hat{\theta}_n - \epsilon \leq \theta \leq \hat{\theta}_n + \epsilon) = \mathbb{P}(|\hat{\theta}_n - \theta| \leq \epsilon) = 1 - \alpha.$$

Sometimes, this approach does not work, because ϵ and α must not depend on θ for the confidence interval to be usable. In fact, we can only determine such an ϵ if the probability distribution of $\hat{\theta}_n - \theta$ does not depend on θ.

More generally, the most effective strategy to find a confidence interval is to look for a pivotal function, or pivot. A pivot is a random variable that depends on θ and the observations X_i, but whose distribution does not depend on θ.

In terms of interpretation, if the confidence interval is small, the set of plausible values for θ is tightly clustered around the estimator. Otherwise, the opposite is true. Thus, a confidence interval constructed from an estimator allows us to assess the precision of that estimator.

3.3 Some examples

3.3.1 Gaussian distribution

Remind that if the X_i are i.i.d., with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, we know that

- the unbiased estimator with minimal variance of μ is $\bar{X}_n$;
- S_n^2 is an estimator of σ^2 which is asymptotically unbiased and that converges in quadratic mean to σ^2;
- $\bar{X}_n$ and S_n^2 are consistent;
- $\bar{X}_n$ and S_n^2 are independent;
- $S_n'^2$ is an unbiased estimator of the variance of X_i;
- $\bar{X}_n$ is Gaussian-distributed (not asymptotically, but "exactly"), and $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$;

We also need other important intermediate results in statistics concerning the link between some distributions based on the Gaussian one.

Proposition 3.3 *(The Chi-square and the Student distributions) Consider an i.i.d. sample* $(Y_1, ..., Y_n)$ *where* $Y_i \sim \mathcal{N}(0, 1)$. *Then,*

- *The random variable* $Z = \sum_{i=1}^n Y_i^2$ *follows a Chi-square distribution with n degrees of freedom, denoted by* $\chi^2(n)$. *The density of Z is given by*

$$f_Z(z) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-z/2} z^{(n/2)-1} 1_{\mathbb{R}^+}(z).$$

- *Consider two independent random variables Y and Z, such that* $Y \sim \mathcal{N}(0, 1)$ *and* $Z \sim \chi^2(n)$. *We call the Student distribution with n degrees of freedom the distribution of the random variable* $Y/\sqrt{Z/n}$, *and we denote it by* $St(n)$.

From these results, it is possible to determine the distribution of the estimator S_n^2 in the case of a Gaussian sample X. We have

$$\frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1).$$

3.3.1.1 CI on the mean parameter First, let us consider the case where the variance is known. The initial idea is to seek a confidence interval for μ of the form $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$. Thus, the problem reduces to finding ϵ, for a fixed α, such that

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon) = 1 - \alpha.$$

Given the previous results, we know that

$$U = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \quad \sim \quad \mathcal{N}(0,1).$$

We thus have found a pivotal function, i.e. a function that depends on the parameter of interest μ and the observations X_i, whose distribution is independent from the parameter μ. Therefore,

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon) = \mathbb{P}(|U| \leq \frac{\sqrt{n}}{\sigma}\epsilon) = 1 - \mathbb{P}(|U| > \frac{\sqrt{n}}{\sigma}\epsilon) = 1 - \alpha.$$

Finally, it leads to consider u_α, some quantile of the standard Gaussian distribution, such that $u_\alpha = \frac{\sqrt{n}}{\sigma}\epsilon$, leading to $\epsilon = \frac{\sigma}{\sqrt{n}}u_\alpha$. Of course, one must choose appropriately the quantile u_α to obtain the probability $(1-\alpha)$.

We choose to consider a symmetric error probability since we deal with gaussian distribution, meaning that we spread the error in two terms, each one being equal to an error of $\alpha/2$. This gives

$$u_\alpha = q_{1-\alpha/2}^{\mathcal{N}(0,1)}.$$

Proposition 3.4 *A confidence interval with level α on μ is given by*

$$CI_{1-\alpha}(\mu) \quad = \quad \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}q_{1-\alpha/2}^{\mathcal{N}(0,1)} \quad , \quad \bar{X}_n + \frac{\sigma}{\sqrt{n}}q_{1-\alpha/2}^{\mathcal{N}(0,1)}\right].$$

Remark 3.5 *Several interesting remarks can be formulated:*

- *it is not mandatory to consider symmetric confidence interval, and any interval such that $\mathbb{P}(\mu \in CI) = 1 - \alpha$ could have been chosen;*
- *the CI is centered around the empirical mean $\bar{X}_n$;*
- *the confidence bounds are random variables, they depend on $\bar{X}_n$;*
- *this CI is an exact CI (not asymptotic);*
- *it relies solely on observed or known quantities;*
- *the higher the variance of X_i, the larger the CI (which seems logical);*
- *the greater the sample size n, the thinner the CI.*

Now, let us move to the case when we do not know the variance parameter σ^2, but still want to provide a CI for μ. Indeed, the previous CI cannot be used because the parameter σ involved in the bounds is unknown. It is thus necessary to look for another pivot! Proposition 3.3 can help us in such a direction... Indeed, we have:

- $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \mathcal{N}(0,1)$;
- $\frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1)$ (known as Fisher's theorem);
- Up to some coefficient ($\sqrt{d}$, where d is the degree of freedom for the χ^2 distribution), the ratio of a standard Gaussian distribution with the square root of a Chi-square distribution is Student-distributed.

Taking into account this, we propose to study

$$\frac{\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2}/(n-1)}}.$$

This leads to

$$\begin{aligned}\frac{\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2}/(n-1)}} &= \sqrt{n}\frac{\bar{X}_n-\mu}{\sigma} \quad\times\quad \frac{\sqrt{n-1}}{\sqrt{\frac{nS_n^2}{\sigma^2}}}\\ &= \sqrt{n-1}\frac{\bar{X}_n-\mu}{\sqrt{S_n^2}} = \sqrt{n-1}\frac{\bar{X}_n-\mu}{\sqrt{\frac{n-1}{n}S_n'^2}}\\ &= \sqrt{n}\;\frac{\bar{X}_n-\mu}{S_n'}.\end{aligned}$$

Magic! We now have a new pivot, independent from σ^2. From this pivot, it is easy to propose the updated confidence interval when the variance of the Gaussian observations is unknown.

Proposition 3.6 *A confidence interval with level α on μ is given by*

$$CI_{1-\alpha}(\mu) \quad=\quad \left[\bar{X}_n - \frac{S_n'}{\sqrt{n}}q_{1-\alpha/2}^{St(n-1)} \quad,\quad \bar{X}_n + \frac{S_n'}{\sqrt{n}}q_{1-\alpha/2}^{St(n-1)}\right],$$

where $q_{1-\alpha/2}^{St(n-1)}$ is the $(1-\frac{\alpha}{2})$th-quantile of the Student distribution with $(n-1)$ degrees of freedom.

Note that the Student distribution also has a symmetric density, which enables us to spread the error α into two terms, leading to this symmetric CI. The quantile of interest can be found in tables dedicated to this task.

If we would have targeted a non-symmetric interval, this could have been for instance

$$CI_{1-\alpha}(\mu) = \quad=\quad \left[\bar{X}_n - \frac{S_n'}{\sqrt{n}}q_{1-\alpha}^{St(n-1)} \quad,\quad +\infty\right].$$

The latter CI provides a lower bound for the estimation of μ with level α.

3.3.1.2 CI on the variance parameter Given the Fisher's theorem, there is a natural pivot for the estimation of σ^2. Indeed, we know that

$$\frac{n\,S_n^2}{\sigma^2} \quad\sim\quad \chi^2(n-1).$$

Hence, $\forall(a,b)\in\mathbb{R}^2$ such that $0<a<b$,

$$\begin{aligned}\mathbb{P}\left(a\le\frac{n\,S_n^2}{\sigma^2}\le b\right) &= F_{\chi^2(n-1)}(b)-F_{\chi^2(n-1)}(a)\\ \mathbb{P}\left(a\le\frac{n\,S_n^2}{\sigma^2}\le b\right) &= \mathbb{P}\left(\frac{n\,S_n^2}{b}\le\sigma^2\le\frac{n\,S_n^2}{a}\right)=1-\alpha.\end{aligned}$$

There exist an infinite number of possibilities to choose a and b in order to obtain the probability $(1-\alpha)$. Usually, we balance the risk of error, meaning that we choose a and b such that

$$F_{\chi^2(n-1)}(b) = 1 - \frac{\alpha}{2} \qquad \text{and} \qquad F_{\chi^2(n-1)}(a) = \frac{\alpha}{2}.$$

We finally obtain:

Proposition 3.7 *A confidence interval with level α on σ^2 is given by*

$$CI_{1-\alpha}(\sigma^2) \quad = \quad \left[\frac{n\,S_n^2}{q_{\alpha/2}^{\chi^2(n-1)}} \quad , \quad \frac{n\,S_n^2}{q_{1-\alpha/2}^{\chi^2(n-1)}} \right],$$

where $q_\alpha^{\chi^2(n-1)}$ is the α-quantile of the distribution $\chi^2(n-1)$.

3.3.2 Bernoulli distribution

This concerns determining the confidence interval for a proportion. Specifically, this proportion is the parameter p of a Bernoulli distribution, given a sample $X = (X_1, ..., X_n)$ from this distribution. We have already seen that a consistent estimator with minimal variance for the proportion is given by $\hat{p}_n = \bar{X}_n$.

Returning to the example of defective items, suppose that among 800 tested items, 420 are defective. The items are independent, and the probability that a randomly chosen item is defective is p. We aim to propose a set estimation for p. Take $X_i \sim \mathcal{B}(p)$: we do not know the details of all the defective items, but we do know that $\hat{p}_n = \frac{420}{800} = 0.525$.

We are interested in a confidence interval with a level of 5% for p; in other words, we are looking for a confidence interval (CI) such that

$$\mathbb{P}(p \in CI) = 1 - \alpha = 0.95.$$

In this example, it is not easy to find a pivotal function. Indeed, when considering the sum of X_i, it becomes difficult to manipulate the binomial distribution. However, there exists an exact confidence interval based on Fisher-Snedecor's distribution. In practice, the binomial distribution is more commonly approximated by the normal distribution. Indeed,

$$\frac{\sum_i X_i - n\mu}{\sqrt{n\sigma^2}} = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} = \sqrt{n}\,\frac{\bar{X}_n - E[X_i]}{\sqrt{Var(X_i)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\sim}} \mathcal{N}(0,1).$$

Hence, we get

$$\sqrt{n}\,\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\sim}} \mathcal{N}(0,1).$$

We now have a pivotal function. Denoting by $T = \sum_i X_i$, we have $\mathbb{P}\left(\left| \frac{T-np}{\sqrt{np(1-p)}} \right| \leq u_\alpha \right) \approx 1-\alpha.$

We thus try to write $\left| \frac{T-np}{\sqrt{np(1-p)}} \right| \leq u_\alpha$ as $Z_1 \leq p \leq Z_2$. In this spirit, we have

$$\left| \frac{T-np}{\sqrt{np(1-p)}} \right| \leq u_\alpha \quad \Leftrightarrow \quad \frac{(T-np)^2}{np(1-p)} \leq u_\alpha^2 \quad \Leftrightarrow \quad p^2(n+u_\alpha^2) - p(2T+u_\alpha^2) + \frac{T^2}{n} \leq 0.$$

This trinomial in p is always positive, except between its roots. Therefore, these two roots are the bounds of the sought interval. We obtain thus

$$\left[\frac{\frac{T}{n}+\frac{u_\alpha^2}{2n}-u_\alpha\sqrt{\frac{u_\alpha^2}{4n^2}+\frac{T(n-T)}{n^3}}}{1+\frac{u_\alpha^2}{n}} \quad ; \quad \frac{\frac{T}{n}+\frac{u_\alpha^2}{2n}+u_\alpha\sqrt{\frac{u_\alpha^2}{4n^2}+\frac{T(n-T)}{n^3}}}{1+\frac{u_\alpha^2}{n}}\right]$$

We can simplify a little bit this expression. Remind that $\hat{p}_n = (T/n)$, and consider that u_α^2 is negligible compared to n (which is true for usual values of α and n), we obtain an asymptotic CI.

Proposition 3.8 *An asymptotic CI on p with level α is given by*

$$CI_{1-\alpha}(p) = \left[\hat{p}_n - u_\alpha\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \quad ; \quad \hat{p}_n + u_\alpha\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right],$$

where $u_\alpha = q_{1-\alpha/2}^{\mathcal{N}(0,1)}$.

Returning to the example of defective items, we obtain an asymptotic confidence interval $CI_{0.95}(p) = [0.49, 0.56]$. For example, we can observe that it is possible for the proportion to be less than 50%, in which case the number of defective items is not a majority.

Remark 3.9 *Through the study of these confidence intervals, we observe that we can widen or narrow the interval primarily by adjusting two factors: the sample size n and the error level α.*

4 Link between maximum likelihood and statistical tests

Here, we take advantage of the properties of the maximum likelihood method to present two types of test statistics. The first family of these tests, the Wald tests, is commonly used for testing the significance of regression coefficients in a (generalized) linear model. The second type of test, the likelihood ratio test, is typically used to compare models with each other, thus facilitating model selection.

4.1 The family of Wald tests

We have here $\hat{\theta}_n$ an asymptotically normal estimator of θ. Hence,

$$Z = \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

where $\hat{\sigma}_n$ is an estimator of the standard deviation of $\hat{\theta}_n$.

Assume that we want to test

$$H_0 : \theta \le \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0,$$

with θ_0 a fixed constant.

We could consider a decision rule of the form:

- if $\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} > t$ (with $t > 0$), then we reject H_0;
- otherwise we do not reject H_0.

This makes sense since we consider that once the normalized distance between $\hat{\theta}_n$ and θ_0 becomes too large, there is little chance that θ is in fact lower than θ_0. To choose the threshold t, one could study

$$\begin{aligned}\alpha &= \mathbb{P}_{H_0}(\text{reject } H_0) = \sup_{\theta\leq\theta_0} \mathbb{P}_\theta\left(\frac{\hat\theta_n-\theta_0}{\hat\sigma_n} > t\right)\\ &= \sup_{\theta\leq\theta_0} \mathbb{P}_\theta\left(\frac{\hat\theta_n-\theta}{\hat\sigma_n} + \frac{\theta-\theta_0}{\hat\sigma_n} > t\right) = \sup_{\theta\leq\theta_0} \mathbb{P}_\theta\left(Z + \frac{\theta-\theta_0}{\hat\sigma_n} > t\right).\end{aligned}$$

The function $\theta \to \mathbb{P}_\theta\left(Z + \frac{\theta-\theta_0}{\hat\sigma_n} > t\right)$ is increasing, meaning that the supremum is reached in θ_0. Basically, we thus have

$$\alpha \approx \mathbb{P}(Z > t),$$

and t is therefore the $(1-\alpha)$-quantile of the standard Gaussian distribution.

4.2 Likelihood ratio test (LRT)

Recall that d is the dimension of the parameter θ. If the regularity constraints for the existence of the maximum likelihood estimator are fulfilled, then we can define the likelihood ratio test statistic for all θ (based on the relative likelihood concept, seen previously), denoted by $W(\theta)$:

$$W(\theta; X) = -2\ln RL(\theta; X) = 2[\ln L(\hat\theta_n; X) - \ln L(\theta; X)].$$

The question now is to understand the behaviour of this test statistic, that compares the value of the likelihood taken in some point θ to the maximum likelihood reachable. To this aim, we use a Taylor development of the log-likelihood. Consider that the MLE is unique, and that the log-likelihood function is twice differentiable with continuous derivatives, then $\forall\theta \in v(\hat\theta_n)$,

$$\begin{aligned}\ln L(\theta; X) &= \ln L(\hat\theta_n; X) + (\theta-\hat\theta_n)(\ln L(\hat\theta_n; X))' + \frac{1}{2}(\theta-\hat\theta_n)^2(\ln L(\hat\theta_n; X))'' + o(||(\theta-\hat\theta_n)^2||)\\ &\approx \ln L(\hat\theta_n; X) + (\theta-\hat\theta_n)\times 0 + \frac{1}{2}(\theta-\hat\theta_n)^2(\ln L(\hat\theta_n; X))''\\ &= \ln L(\hat\theta_n; X) + \frac{1}{2}(\theta-\hat\theta_n)^2(\ln L(\hat\theta_n; X))''.\end{aligned}$$

We thus obtain

$$\begin{aligned}\ln L(\theta; X) - \ln L(\hat\theta_n; X) &= \frac{1}{2}(\theta-\hat\theta_n)^2(\ln L(\hat\theta_n; X))''\\ 2\ \ln RL(\theta; X) &= (\theta-\hat\theta_n)^2\left(\sqrt{\ln L(\hat\theta_n; X)''}\right)^2.\end{aligned}$$

Now, we use the properties of the MLE, in particular that it is asymptotically Gaussian with a variance equal to the inverse of the Fisher information, where the Fisher information is the opposite of the second derivative of the log-likelihood.

$$W(\theta; X) = -2\ \ln RL(\theta; X) = \left((\hat\theta_n-\theta)\sqrt{I_n(\hat\theta_n)^{-1}}\right)^2.$$

We know the behaviour of the right-hand term: this a standard Gaussian distribution (remember that the MLE is unbiased, that is to say that θ is the target in the latter expression, say $\theta = \theta_0$ then), to the square. Hence this is a Chi-square distribution! We thus have that

$$W(\theta_0; X) = 2[\ln L(\hat\theta_n; X) - \ln L(\theta_0; X)] \quad \underset{n\to\infty}{\overset{\mathcal{L}}{\sim}} \quad \chi^2(d).$$

This likelihood ratio test (LRT) statistic is taken in θ_0 since θ_0 and $\hat{\theta}_n$ are neighbours, which must be the case when using the Taylor development!

The LRT makes sense: it compares the likelihood obtained in $\hat{\theta}_n$ (which is asymptotically a good estimator of θ) to the one in θ_0: if they are similar, $W(\theta_0)$ tends to small values, which implies that we cannot reject the null hypothesis H_0.

Now we have found a pivot. Indeed,

$$\mathbb{P}_{\theta_0}(W(\theta_0; X) \leq q_{1-\alpha}^{\chi^2(d)}) \quad \approx \quad 1 - \alpha.$$

This way,

$$W(\theta; X) = 2 \ln \left(\frac{L(\hat{\theta}_n; X)}{L(\theta; X)} \right) \leq q_{1-\alpha}^{\chi^2(d)} \quad \Leftrightarrow \quad \left\{ \theta \in \Theta : \ \ln L(\theta; X) \geq \ln L(\hat{\theta}_n; X) - \frac{1}{2} q_{1-\alpha}^{\chi^2(d)} \right\}.$$

The latter expression is called a confidence region with asymptotic level $(1 - \alpha)$.

Generalizing to the framework of statistical tests, Assume that we want to test

$$H_0 : \ \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \ \theta \in \Theta_1,$$

with $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \varnothing$ (partition).

It seems logical to reject H_0 if having observed X is much more likely to occur when θ varies across Θ_1, which gives

$$\sup_{\theta \in \Theta_1} L(\theta; X) \quad >> \quad \sup_{\theta \in \Theta_0} L(\theta; X).$$

We then use the test statistic

$$T(X) = 2 \ln \left(\frac{\sup_{\theta \in \Theta_1} L(\theta; X)}{\sup_{\theta \in \Theta_0} L(\theta; X)} \right).$$

The statistic $T(X)$ depends on the observations X, but not on the parameter θ. A rejection region that is natural is given by

$$R = \{X : \ T(X) > t\},$$

with the threshold t to be chosen depending on the test level α. To choose an appropriate t, we have to know the distribution of $T(X)$ under H_0 (or it asymptotic/limiting distribution).

5 Exercices

See the correction on the manuscript pages.

Exercise 1

Calculate the likelihood and log-likelihood in the cases below. Specify what the parameter space is and whether the classical assumptions are satisfied.

1. **Poisson Distribution:** x is a single observation, modeled by a Poisson distribution with parameter θ.

2. **Exponential Distribution:** x is a sample of size n, $x = (x_1, \ldots, x_n)$, and each observation is modeled by an exponential distribution with mean θ.

3. **Centered Cauchy Distribution:** x is a sample of size n, and each coordinate is modeled by the centered Cauchy distribution with parameter θ.

4. **Geometric Distribution:** x is a sample modeled by a geometric distrib. with param. π.

Exercise 2

Using a software tool, plot the log-likelihood for exponential distributions or Cauchy distributions for the sample $225, 171, 198, 189, 189, 135, 162, 135, 117, 162$, which represents the means (in thousands of load cycles) of several groups of springs before they wear out.

Depending on the value of c, what is the shape of the set of plausible values for the parameter, given by $\{\theta : l(\theta) > c\}$?

Exercise 3

Calculate the maximum likelihood estimator for θ for a sample of size n, where each observation is modeled by the distribution below.

1. $f(x|\theta) = \theta x^{\theta-1} 1_{\{0<x<1\}}$, where $\theta > 0$,

2. $f(x|\theta) = \theta^2 x e^{-\theta x} 1_{\{x>0\}}$, where $\theta > 0$,

3. $f(x|\theta) = (\theta + 1) x^{-\theta-2} 1_{\{x>1\}}$, where $\theta > 0$.

Exercise 4

Calculate the Fisher information for a sample of size n:

1. From a Bernoulli distribution $B(\pi)$,

2. From a normal distribution $N(\mu, \sigma^2)$.

Exercise 5

Consider a sample $x = (x_1, \dots, x_n)$, where each coordinate is modeled by a Bernoulli distribution with parameter π. Show that $S = X_1 + \cdots + X_n$ is an exhaustive statistic. Is it minimal?

Repeat the same question when each coordinate is modeled by a Gaussian distribution centered at θ with variance 1.

Repeat the same question when each coordinate is modeled by a Poisson distribution with parameter θ.

Exercise 6

Let $X = (X_1, \dots, X_n)$ be a sample of size n from a uniform distribution on $[0, \theta]$. Let $\hat{\theta}_n = \max(X_1, \dots, X_n)$, and $\tilde{\theta}_n = 2\bar{X}_n$.

1. Show that $\tilde{\theta}_n$ is an unbiased and consistent estimator of θ, and that $n(\tilde{\theta}_n - \theta)$ converges in distribution to $N(0, \theta^2/3)$.

2. Construct an asymptotic confidence interval of security level $1 - \alpha$, based on $\tilde{\theta}_n$.

3. Show that $\hat{\theta}_n/\theta$ is a pivot for the estimation of θ.

4. Construct a confidence interval for θ of security level $1 - \alpha$, based on $\hat{\theta}_n$.

5. Compare the two intervals.

Exercise 7

A sample of size 25 from a normal population with variance $\sigma^2 = 81$ gave an empirical mean of 81.2. Find a confidence interval with a security level of 0.95 for the mean μ.

Exercise 8

Let $\bar{X}_n$ be the empirical mean of a sample of size n from the distribution $N(\mu, \sigma = 4)$. Find the minimum value of n such that $[\bar{X}_n - 1, \bar{X}_n + 1]$ is a confidence interval for μ with a security level of 90%.

Chapter 4

Statistical Tests

1 Framework

1.1 Example

Imagine that we observe a 10-sample $(X_1, \ldots, X_{10})$ from the $\mathcal{N}(\mu, \sigma^2)$ distribution, and we want to know whether $\mu \leq 99$ kg or $\mu > 99$ kg. To decide between these two hypotheses, we can base our decision on the estimator $\hat{\mu}_n = \bar{X}_{10}$ of μ. For example, we can adopt the following decision rule:

- if $\hat{\mu}_n > 99$, we decide that $\mu > 99$;
- if $\hat{\mu}_n \leq 99$, we decide that $\mu \leq 99$.

Let's calculate the probabilities of making an error if we adopt this rule. One way to make an error is to decide that $\mu > 99$ when this is not actually the case. The associated error probability, in the worst-case scenario, is:

$$\alpha_1 = \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\hat{\mu}_n > 99\right) = \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} > \frac{99 - \mu}{\sigma/\sqrt{n}}\right).$$

Let F denote the cumulative distribution function of $\mathcal{N}(0,1)$. Since $\hat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, we obtain

$$\alpha_1 = \sup_{\mu \leq 99, \sigma > 0} 1 - F\left(\frac{99 - \mu}{\sigma/\sqrt{n}}\right) = \sup_{x > 0} 1 - F(x) = 1 - F(0) = 0.5$$

We thus have, in the worst-case scenario (which corresponds to $\mu = 99$), a 50% chance of making an error by deciding that $\mu > 99$ if we adopt this decision strategy. The other possible error is deciding that $\mu \leq 99$ when this is not the case:

$$\begin{aligned}
\alpha_2 &= \sup_{\mu > 99, \sigma > 0} P_{\mu,\sigma}\left(\hat{\mu}_n \leq 99\right) = \sup_{\mu > 99, \sigma > 0} P_{\mu,\sigma}\left(\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{99 - \mu}{\sigma/\sqrt{n}}\right) \\
&= \sup_{\mu > 99, \sigma > 0} F\left(\frac{99 - \mu}{\sigma/\sqrt{n}}\right) \\
&= \sup_{x < 0} F(x) = F(0) = 0.5
\end{aligned}$$

In the worst case, we also have a 50% chance of making an error by deciding that $\mu \leq 99$. Thus, with the decision rule we have chosen, we have a 50% chance of being wrong regardless of the conclusion we reach. We might as well flip a coin! How can we improve our decision rule? Instead of comparing $\hat{\mu}_n$ to 99, we could compare it to another value t, attempting to adjust t in order to reduce the previous error probabilities. If we revisit the previous expressions, we now have:

$$\alpha_1 = \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\hat{\mu}_n > t\right) \qquad \text{and} \qquad \alpha_2 = \sup_{\mu > 99, \sigma > 0} P_{\mu,\sigma}\left(\hat{\mu}_n \leq t\right)$$

It is clear from these expressions that as t increases, α_1 becomes smaller. Conversely, as t increases, α_2 becomes larger. **In general, both errors cannot be small at the same time**. The method used is then:

- to choose among the two types of possible errors the error that one wants to control absolutely (because, for example, it has more costly consequences than the other);

- to assign it a small level α;

- to choose t accordingly.

For example, one might want to avoid concluding that the new process is better ($\mu > 99$) when it is not. The error that we want to control is therefore

$$\alpha_1 = \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\hat{\mu}_n > t\right).$$

In practice, this means prioritizing one hypothesis over the other.
We will test $H_0 : \mu \leq 99$ kg against $H_1 : \mu > 99$ kg. By the definition of H_0 (and thus H_1), the error we are controlling is

$$\alpha = P_{H_0}\left(\text{reject}(H_0)\right).$$

α is called the significance level of the test or the type-I error. In the chosen example, to find a decision rule of level α, we will need to change the statistic used. Indeed, for all $t \in \mathbb{R}$,

$$\begin{aligned} \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\hat{\mu}_n > t\right) &= \sup_{\mu \leq 99, \sigma > 0} 1 - F\left(\frac{t-\mu}{\sigma/\sqrt{n}}\right) = 1 - \inf_{\mu \leq 99, \sigma > 0} F\left(\frac{t-\mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \inf_{\sigma > 0} F\left(\frac{t-99}{\sigma/\sqrt{n}}\right) \geq 1 - F(0) = 0.5 \end{aligned}$$

taking $\sigma^2 \to \infty$.

However, considering the case $\sigma \to +\infty$ is not necessarily relevant, since we have an estimate of σ given by S'_n, where S'^2_n is the unbiased estimator of the variance. Therefore, we will instead use the following test statistic:

$$\frac{\hat{\mu}_n - 99}{S'_n/\sqrt{n}}.$$

The decision rule will be as follows:

- If $\frac{\hat{\mu}_n - 99}{S'_n/\sqrt{n}} > t$, we reject H_0;

- If $\frac{\hat{\mu}_n - 99}{S'_n/\sqrt{n}} \leq t$, we do not reject H_0.

The value of t is chosen such that

$$\alpha = 5\% = \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\frac{\hat{\mu}_n - 99}{S'_n/\sqrt{n}} > t\right) = \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\frac{Z + \frac{\mu - 99}{\sigma/\sqrt{n}}}{\sqrt{Y/(n-1)}} > t\right),$$

where $Z = \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$ and $Y = \frac{(n-1)S'^2_n}{\sigma^2}$. It can be shown that Z and Y are independent, with $Z \sim \mathcal{N}(0,1)$

and $Y \sim \chi^2(n-1)$. Therefore, the supremum in μ is reached at $\mu = 99$, and we have:

$$\alpha = 5\% = \sup_{\mu \leq 99, \sigma > 0} P_{\mu,\sigma}\left(\frac{Z}{\sqrt{\frac{Y}{n-1}}} > t\right) = P\left(\frac{Z}{\sqrt{\frac{Y}{n-1}}} > t\right) = P(V > t),$$

where we have already seen that V follows a Student distribution, i.e. $V \sim St(n-1)$.

Thus, t is the 95% quantile of the $St(n-1)$ distribution:

$$t = q_{0.95}^{St(n-1)}.$$

Once the significance level α is set, t is fixed, the decision rule is established, and we have no way to control the second type of error, namely

$$\text{type-II error} = P_{H_1}(\text{accepter } H_0).$$

In the previous example, this error is given by the function

$$(\mu,\sigma) \in]99;+\infty[\times\mathbb{R}^+ \to P_{\mu,\sigma}\left(\frac{\hat{\mu}_n - 99}{\frac{S_n'}{\sqrt{n}}} \leq q_{0.95}^{St(n-1)}\right) = P_{\mu,\sigma}\left(\frac{Z + \frac{\mu-99}{\sigma/\sqrt{n}}}{\sqrt{Y/(n-1)}} \leq q_{0.95}^{St(n-1)}\right).$$

This function is decreasing in μ, and as $\mu \to 99$, the value of this error approaches

$$P\left(\frac{Z}{\sqrt{Y/(n-1)}} \leq q_{0.95}^{St(n-1)}\right) = 95\%.$$

Thus, the type II error can be high depending on the values of the parameter. This should not be surprising! The closer μ is to 99 kg (from above), the harder it is to determine from a sample whether μ is greater than or less than 99.

The lack of control over the type II error has the following consequence: the conclusion of a test only holds value as evidence when that conclusion is the rejection of H_0.
Indeed, if at the end of the test we conclude to reject H_0, we know there is a probability α of being wrong. On the other hand, if we conclude to accept H_0, we may have a high chance of being incorrect. Thus, it is more appropriate to view the acceptance of H_0 as a failure to reject H_0. The approach here is very empirical: based on my experience, nothing suggests that H_0 is not true.

As stated in Wasserman's book (All of Statistics: a concise course in Statistical Inference, Springer texts in statistics): *"Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that he is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0."*

1.2 Definition

More generally, we can formulate the following definitions.

Definition 1.1 *Let $\mathcal{T}_0 \cup \mathcal{T}_1$ be a partition of the parameter space $\mathcal{T}$. A test of level α for the hypothesis $H_0 : \theta \in \mathcal{T}_0$ against the hypothesis $H_1 : \theta \in \mathcal{T}_1$ is the specification of a subset $\mathcal{R}$ of the observation set X (where $\mathcal{R}$ does not depend on the parameter θ) such that*

$$\sup_{\theta \in \mathcal{T}_0} P_\theta(X \in \mathcal{R}) \leq \alpha. \tag{4.1}$$

In this definition,

- $\mathcal{R}$ is called the rejection region of H_0.

- H_0 is called the null hypothesis. It is the hypothesis favored by the tester, in the sense that they do not want to make an error by rejecting it. H_1 is the alternative hypothesis.
- The decision rule associated with a test with rejection region $\mathcal{R}$ consists of deciding that $\theta \notin \mathcal{T}_0$ (rejecting the null hypothesis H_0) whenever $X \in \mathcal{R}$ (the observation is in the rejection region). In other words, denoting 1 as the decision to reject H_0 and 0 as the contrary decision, the decision rule is the function $d : \mathcal{X} \to \{0, 1\}$ given by:

$$d(X) = \begin{cases} 1 & \text{if } X \in \mathcal{R} \\ 0 & \text{otherwise} \end{cases}$$

- The size of the test is given by $\sup_{\theta \in \mathcal{T}_0} P_\theta(X \in \mathcal{R}) = \sup_{\theta \in \mathcal{T}_0} P_\theta(d(X) = 1)$.
- When this maximum is achieved at a single value θ^* in $\mathcal{T}$, this value corresponds to the most difficult case. It is the value of θ for which it is most challenging to make a decision based on the data.
- The power function of the test is the function $\beta : \theta \in \mathcal{T}_1 \mapsto P_\theta(X \in \mathcal{R}) = P_\theta(d(X) = 1)$ (the probability of correctly rejecting H_0).
- The second type of error of the test is given by the function $\theta \in \mathcal{T}_1 \mapsto 1 - \beta(\theta)$.
- Typically, the supremum of the probability is taken in unilateral (one-sided) tests: the null hypothesis is a whole set of potential values for the parameter to test, and we consider the worst-case scenario. In the classical bilateral test, a single value has to be tested and there is no supremum.

Remark 1.2 *If we have the choice between several tests of the same level, we will obviously choose the one with the highest power function, assuming that these functions can be compared.*

Remark 1.3 *When the level α changes, the rejection region, as well as the decision rule, change. When studying a whole collection of tests with different levels α, it is common to index the confidence region R_α, as well as the decision rule d_α.*

1.3 How to build such a test

The construction of a test follows the following steps:

1. Choice of the preferred hypothesis H_0. H_0 should be chosen based on the error that one wants to control, $\alpha = P_{H_0}(\text{reject } H_0)$.
2. Choice of the level α (small). The traditional value is $\alpha = 5\%$.
3. Choice of a test statistic. One chooses a statistic $T = t(X)$ whose behavior is as different as possible between H_0 and H_1. For example, an estimator $\hat{\theta}_n(X)$ of the parameter θ.
4. Determination of the rejection region $\mathcal{R}$ based on the behavior of the statistic T when θ is far from $\mathcal{T}_0$: $\mathcal{R} = \{x \in \mathcal{X} : t(x) \in \ldots\}$
5. Determination of the bounds of the rejection region at level α, based on the distribution of the statistic under H_0, so that equation (4.1) is satisfied.
6. Study of the power function (or equivalently the second type error) of the constructed test.

If multiple tests are available (with same level), a final step is to choose, among these different tests, the one with the highest power function (if one exists).

Remark 1.4 *We must therefore know the distribution of the test statistic under H_0 in order to construct the rejection region. This constraint often imposes the choice of the hypothesis H_0.*

Remark 1.5 *The notions of confidence region and rejection zone are dual. We will show in an exercise that for any confidence region $CI(X)$ with a security coefficient $1-\alpha$, one can define a test of level α for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ by choosing the rejection region as $\mathcal{R} = \{x \in \mathcal{X}$ such that $\theta_0 \notin CI(x)\}$.*

1.4 The notion of p-value

We are given a family of decision rules $d_\alpha : \mathcal{X} \to \{0; 1\}$ for every level α in $[0, 1]$. We denote by $\mathcal{R}_\alpha = \{x \in \mathcal{X} : d_\alpha(x) = 1\}$ the associated rejection zones. We make the following assumptions:

Assumption 1.1

$$\forall \alpha, \qquad \sup_{\theta \in \mathcal{T}_0} P_\theta(X \in \mathcal{R}_\alpha) = \alpha. \tag{4.2}$$

Assumption 1.2 *Pour tout $0 \leq \alpha \leq \alpha' \leq 1$, on a $\mathcal{R}_\alpha \subset \mathcal{R}_{\alpha'}$.*

Remark 1.6 *Hypothesis 2 is equivalent to saying that, for all $x \in \mathcal{X}, \alpha \mapsto d_\alpha(x)$ is increasing.*

Definition 1.7 *The p-value is the smallest value of α for which we reject H_0 based on the observed data. Formally, for all $x \in \mathcal{X}$,*

$$p(x) = \inf\{\alpha : x \in \mathcal{R}_\alpha\}.$$

The p-value is a statistic. It is a function of the data set that can be calculated without knowing the unknown parameter θ. The p-value is not the probability of the alternative hypothesis! It is the value of α for which the decision changes based on the data.

Proposition 1.8 *Suppose that the supremum in equation (4.2) is attained at a unique θ_0 that does not depend on α. If $X \sim P_{\theta_0}$, then the p-value statistic $p(X)$ follows a continuous uniform distribution on the interval $[0, 1]$.*

2 Gaussian sample

2.1 Preliminary results

The useful results were given in Section 3.3.1 of Chapter 3. Coming back to this section and how to build a confidence interval, it is then easy to deduce the statistical tests below. For instance, we have

- $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$,
- $\frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1)$.

Our framework is the observation of an i.i.d. sample $X = (X_1, ..., X_n)$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. We remind that we had found the confidence intervals for the mean parameter as well as the variance parameter.

For the mean μ when the variance σ^2 is known, we had obtained:

$$CI_{1-\alpha}(\mu) \quad = \quad \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}^{\mathcal{N}(0,1)} \quad , \quad \bar{X}_n + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}^{\mathcal{N}(0,1)}\right],$$

where $q_\alpha^{\mathcal{N}(0,1)}$ denotes the α-quantile of the standard Gaussian distribution.

When the variance parameter is unknown, we had obtained the following:

$$CI_{1-\alpha}(\mu) \quad = \quad \left[\bar{X}_n - \frac{S_n^{'}}{\sqrt{n}} q_{1-\alpha/2}^{St(n-1)} \quad , \quad \bar{X}_n + \frac{S_n^{'}}{\sqrt{n}} q_{1-\alpha/2}^{St(n-1)} \right],$$

where $q_{1-\alpha/2}^{St(n-1)}$ is the $(1-\frac{\alpha}{2})$th-quantile of the Student distribution with $(n-1)$ degrees of freedom.

For the variance parameter, we had obtained:

$$CI_{1-\alpha}(\sigma^2) \quad = \quad \left[\frac{n\,S_n^2}{q_{\alpha/2}^{\chi^2(n-1)}} \quad , \quad \frac{n\,S_n^2}{q_{1-\alpha/2}^{\chi^2(n-1)}} \right],$$

where $q_\alpha^{\chi^2(n-1)}$ is the α-quantile of the distribution $\chi^2(n-1)$.

2.2 Test on the mean parameter

Say that we are interested in the following bilateral statistical test:

$$H_0 : \ \mu = \mu_0 \qquad \text{against} \qquad H_1 : \ \mu \neq \mu_0.$$

2.2.1 With known variance

Proposition 2.1 *(Bilateral test on μ with σ^2 unknown) The set*

$$\mathcal{R} \quad = \quad \left\{ \frac{|\bar{X}_n - \mu_0|}{\sigma/\sqrt{n}} \geq q_{1-\alpha/2}^{\mathcal{N}(0,1)} \right\}$$

is the rejection region with level α for this test.

Proof: exercice. Start by the definition of α, the first type error...

2.2.2 With unknown variance

Proposition 2.2 *(Bilateral test on μ with σ^2 unknown) The set*

$$\mathcal{R} \quad = \quad \left\{ \frac{|\bar{X}_n - \mu_0|}{S_n'/\sqrt{n}} \geq q_{1-\alpha/2}^{St(n-1)} \right\}$$

is the rejection region with level α for this test.

Proof: exercice.

Remark 2.3 *It is also possible to set up unilateral tests. For instance, one could consider the following hypotheses:*

$$H_0 : \ \mu \leq \mu_0 \qquad \textit{against} \qquad H_1 : \ \mu > \mu_0,$$

If we want to perform such a unilateral test, we would use the following result:

Proposition 2.4 *(unilateral test on μ with σ^2 unknown) The set*

$$\mathcal{R} \quad = \quad \left\{ \bar{X}_n \geq \mu_0 + q_{1-\alpha}^{St(n-1)} \frac{S_n'}{\sqrt{n}} \right\}$$

is the rejection region with level α for this test.

Proof: exercice.

2.3 Test on the variance parameter

Say that we are interested in the following unilateral statistical test:

$$H_0 : \sigma^2 \geq \sigma_0^2 \qquad \text{against} \qquad H_1 : \sigma^2 < \sigma_0^2.$$

Then, we have

Proposition 2.5 *(unilateral test on σ^2) The set*

$$\mathcal{R} \quad = \quad \left\{ \frac{nS_n^2}{\sigma_0^2} \leq q_{1-\alpha}^{\chi(n-1)} \right\}$$

is the rejection region with level α for this test.

Proof: exercice.

3 Back to LRT tests

We assume that we want to test

$$H_0 : \theta \in \mathcal{T}_0 \qquad \text{against} \qquad H_1 : \theta \in \mathcal{T}_1,$$

where $\mathcal{T}_0 \cup \mathcal{T}_1$ is a partition of $\mathcal{T}$.

It is reasonable to reject H_0 if the observed value X is much more probable when the parameter θ varies in $\mathcal{T}_1$ than when it varies in $\mathcal{T}_0$, i.e., if

$$\sup_{\theta \in \mathcal{T}_1} L(\theta; X) \quad >> \quad \sup_{\theta \in \mathcal{T}_0} L(\theta; X).$$

Thus, the likelihood ratio test should rely on the statistic

$$T(X) = 2 \ln \left(\frac{\sup_{\theta \in \mathcal{T}_1} L(\theta, X)}{\sup_{\theta \in \mathcal{T}_0} L(\theta, X)} \right),$$

(where T depends only on X and not on the parameter θ). A reasonable rejection region is therefore given by

$$R = \{X \text{ such that } T(X) \geq t\},$$

where t is to be chosen based on the level of the test.

To choose t, we must therefore know the distribution of $T(X)$ under H_0, or at least its asymptotic distribution as n approaches infinity.

When testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we can note that as long as $L(\theta, X)$ is continuous in θ,

$$\sup_{\theta \in \mathcal{T}_0} L(\theta, X) = L(\theta_0, X),$$

and

$$\sup_{\theta \in \mathcal{T}_1} L(\theta, X) = \sup_{\theta \neq \theta_0} L(\theta, X) = \sup_{\theta \in \mathcal{T}} L(\theta, X) = L(\hat{\theta}_n; X),$$

with $\hat{\theta}_n$ the MLE. We therefore consider

$$T(X) = 2 \left[l(\hat{\theta}_n; X) - l(\theta_0; X) \right],$$

where l is the log-likelihood.

At the end of Chapter 3, we found the distribution of this test statistic.

Proposition 3.1

$$2\left(l(\hat{\theta}_n, X) - l(\theta, X)\right) \xrightarrow{\mathcal{L}} \chi^2(d),$$

Thus, under $H_0 : \theta = \theta_0$, $T(X)$ converges in distribution to the $\chi^2(d)$ distribution. We can therefore complete the construction of the test and determine the value of t:

$$\alpha = P_{H_0}\,(\text{reject } H_0) = P_{\theta_0}\,(T(X) \geq t) \approx P(Z \geq t), \quad \text{where } Z \sim \chi^2(d).$$

t is thus the $(1-\alpha)$-quantile of the $\chi^2(d)$ distribution, and we now have the rejection region.

Remark 3.2 *If we test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are two different fixed values of θ, we have*

$$T(X) = 2\left(\log(L(\theta_1, X)) - \log(L(\theta_0, X))\right).$$

Assume that there exists a unique k_α such that $P_{\theta_0}(T(X) \geq k_\alpha) = \alpha$, hence the region

$$R_0 = \{X; T(X) \geq k_\alpha\}$$

is the rejection region of the LRT with level α. We then have the following seminal lemma:

Lemma 3.3 *(Lemma by Neyman-Pearson) Let $\mathcal{R}$ be the rejection region of the statistical test with level α, such that*

$$H_0 :\ \theta = \theta_0 \quad \textit{against} \quad H_1 :\ \theta = \theta_1.$$

Then, $P_{\theta_1}(\mathcal{R}_0) \geq P_{\theta_1}(\mathcal{R})$. In other words, the LRT is the most powerful test among all tests with level α.

4 Exercices

Exercise 1

Recall that a confidence region $I_n(\alpha)$ with confidence level $1-\alpha$ is defined by:

$$P_\mu[\mu \in I_n(\alpha)] \geq 1 - \alpha, \quad \forall \mu \in \Theta.$$

Show that for any confidence region $I_n(\alpha)$, there exists a test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, such that:

$$I_n(\alpha) = \{\mu_0 \,|\, [\text{we accept } H_0(\mu_0)](\alpha)\}.$$

Exercise 2

Let X be a sample of size n from the distribution $\mathcal{N}(\mu, 1)$. Assume that $\mu \in \Theta = \{\mu_0, \mu_1\}$, where $\mu_0 < \mu_1$. We want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$.

Consider the test with the rejection region:

$$R(X) = \{\bar{X}_n \geq \frac{\mu_0 + \mu_1}{2}\}.$$

Compute the level of the test. Let α denote the level, which value of n do you consider to guarantee a test with level α?
Give the power of the test.

Exercise 3

Let X be a sample of size n from the Poisson distribution with parameter μ, where $\mu \in \Theta = \{1, 2\}$.

We consider the test for $H_0 : \mu = 1$ against $H_1 : \mu = 2$, with the rejection region

$$R(X) = \{\bar{X}_n > 3\}.$$

If we want a test with level $\alpha = 0.05$, which value of n should we consider? Compute the power of the test with this value of n.

Exercise 4

Let $X = (X_1, \dots, X_n)$ be a sample of size n from the uniform distribution on $[0, \mu]$. Define $\hat{\mu}_n = \max(X_1, \dots, X_n)$.

1. Show that $\frac{\hat{\mu}_n}{\mu}$ is a pivotal quantity for estimating μ.

2. Construct a test for $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ with level α.

3. Determine the power curve of this test.

4. Is the constructed test the likelihood ratio test?

5. Construct the Wald test based on $\hat{\mu}_n$.

Exercise 5

A product sold is presented in boxes with the label: "Contents 500 grams." One may wonder what this means, as the quantity contained in a randomly chosen box is a random variable (denoted by X) with a probability density. Let μ be the mean of the variable X. The information gathered from the manufacturer asserts that the value of μ is supposed to be equal to 500 grams.

1. Provide a 95% confidence interval for μ.

2. Test whether the manufacturer's specification is met.

Given data (in grams): 490; 490; 490; 492; 492; 495; 497; 497; 502; 505.

Exercise 6

A factory manufactures cables whose breaking strength follows a normal distribution $N(\mu_0, \sigma^2)$ with $\mu_0 = 99$ kg. To test a new manufacturing process for the cables, the breaking strengths of ten cables were observed with the following values (in kilograms): 101; 102; 100; 104; 105; 99; 103; 100; 101; 105.

Is the new manufacturing process better than the previous one? Given that the standard deviation for the old manufacturing process was $\sigma = 1$ kg, is the new process more precise than the old one?

Exercise 7

Consider an i.i.d. sample $(X_i)_{1 \leq i \leq n}$ where X_1 has the probability density function:

$$f_\theta(x) = \theta \exp(-\theta x) \mathbb{I}_{[0,+\infty[}(x),$$

where $\theta > 0$ is the unknown parameter.

1. Estimation of θ using $Y_n = n \,/\, \sum_{i=1}^n X_i$

(a) Show that the random variable Y_n is well-defined.

(b) Explain why it makes sense to choose Y_n as an estimator for θ.

(c) Determine the limiting distribution of $\sqrt{n}(Y_n - \theta)$.

(d) Find the distribution of $\sum_{i=1}^n X_i$. From this, deduce the value of $\mathbb{E}[(Y_n - \theta)^2]$.

2. Estimation of θ using $Z_n = \dfrac{n-1}{n} Y_n$

(a) Does Z_n satisfy similar convergence properties as Y_n?

(b) Which estimator would you choose, Y_n or Z_n, for estimating θ?

3. Confidence Interval

Provide a two-sided asymptotic confidence interval of level $1 - \alpha$ for θ.

4. Hypothesis Test for θ

Let $\alpha \in]0, 1[$. Propose an asymptotic test at level α for testing $H_0 : \theta \geq 1$ against $H_1 : \theta < 1$.

5. Hypothesis Test for θ

Propose an asymptotic test at level α for testing $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$.

Chapter 5

Bayesian Statistics

1 Introduction to bayesian concepts

One the seminal books on bayesian statistics is : C.P. Robert, The Bayesian choice: a decision-theoretic motivation. Springer.

We consider a parametric statistical model $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$, where frequentist statistics assume the existence of a true parameter θ_0 that generates the data, i.e. $(X_1, \cdots, X_n) \sim g(.; \theta_0)$ where g is the probability density function (pdf) here.

Then we look for an estimator $\hat{\theta}_n$ of θ_0 with nice properties, like the MLE:

$$\hat{\theta}_n \quad \underset{n\to\infty}{\overset{d}{\sim}} \quad \mathcal{N}(\theta, I_n^{-1}(\theta)).$$

The estimator provides us with a point estimate for the parameter, which can be complemented by an interval estimate using its asymptotic results. For instance,

$$CI_{1-\alpha}(\theta) = [\hat{\theta}_n - q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \hat{\sigma}_n \ ; \ \hat{\theta}_n + q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \hat{\sigma}_n],$$

with $\hat{\sigma}_n$ an estimator of the standard deviation de $\hat{\theta}_n$.

In summary, the estimator $\hat{\theta}_n$ is a random variable for which we often only know the asymptotic distribution. In the basic framework of bayesian statistics, the parameter θ is no longer considered as deterministic and unknown. Instead, this parameter is itself considered a random variable.

2 Milestones in bayesian statistics

In Bayesian statistics, we also model the certainty we have about the parameter through a probability distribution on this parameter. We use the accumulated experience from collected observations $x_1, ..., x_n$ to update this uncertainty.

2.1 a priori distribution

It is the probability distribution chosen to model the uncertainty about the parameter ; this prior distribution allows us to incorporate, for example, expert opinion to encode our prior knowledge and ignorance about the true parameter θ_0 before observing data.

Let us denote it by Π, meaning that

$$\theta \quad \sim \quad \Pi.$$

Example 2.1 *In a Gaussian model, i.e. $X_i \sim \mathcal{N}(\theta)$ where $\theta = (\mu, \sigma^2)$, we could for instance choose* *as an a priori distribution:*

$$\Pi(\theta) = \Pi(\mu) \times \Pi(\sigma^2) \quad \sim \quad \mathcal{N}(\mu_0, \Gamma) \times \mathcal{IG}(\alpha, \beta).$$

We thus consider that the a priori distributions of μ and σ^2 are independent.

Definition 2.1 *The parameters of the prior distribution are called hyperparameters. These hyperparameters have values set by the statistician or expert. They are not estimated.*

In the previous example, the hyperparameters are $(\mu_0, \Gamma, \alpha, \beta)$.

Definition 2.2 *A Bayesian model provides, for a random variable or a sequence of random variables, a conditional distribution and a prior distribution.*

$$\begin{aligned} X|\Theta = \theta \quad &\sim \quad f(x|\theta) \\ \Theta \quad &\sim \quad \pi(\alpha) \end{aligned}$$

where f is the pdf of X, and the hyperparameter α as to be given by the user.

2.2 Joint and a posteriori distributions

Given that the parameter θ has a prior distribution, we can estimate the joint distribution for a given statistical model, namely

$$\pi_{X,\theta}(x, \theta) = f(x|\theta)\pi(\theta).$$

In practice, the joint distribution has little interest in Bayesian statistics; we are more concerned with the posterior distribution, usually denoted by $\pi(\theta|X = x)$.

Definition 2.3 *One can differentiate four different cases to give the expression of the a posteriori distribution:*

- *X and Θ have discrete distributions:*

$$\pi(\Theta = \theta_i|X = x) = \frac{P(\Theta = \theta_i, X = x)}{P(X = x)} = \frac{P(X = x|\Theta = \theta_i)\pi(\Theta = \theta_i)}{\sum_k P(X = x|\Theta = \theta_k)\pi(\Theta = \theta_k)}$$

- *X has a discrete distribution and Θ has a continuous one:*

$$\pi(\theta|X = x) = \frac{P(X = x|\theta)\pi(\theta)}{\int_{u\in\Theta} P(X = x|u)\pi(u)du}$$

- *X has a continuous distribution and Θ has a discrete one:*

$$\pi(\Theta = \theta_i|x) = \frac{f(x|\Theta = \theta_i)\pi(\Theta = \theta_i)}{\sum_k f(x|\Theta = \theta_k)\pi(\Theta = \theta_k)}$$

- *X and Θ have continuous distributions:*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{u\in\Theta} f(x|u)\pi(u)du}$$

We often use the last case as a general expression to give the posterior distribution of the parameter θ.

Remark 2.4 • *The denominator corresponds to the marginal distribution of X, which is often denoted as $m(x)$. This denominator also serves as a normalization constant for the posterior distribution. This constant ensures that we obtain a valid probability density. Therefore, this denominator is independent of the parameter θ.*

- *To determine the behavior of the posterior distribution, we often work up to this constant, i.e.*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{u\in\Theta} f(x|u)\pi(u)du} \propto f(x|\theta)\pi(\theta).$$

  The distribution of the right-hand side term is equivalent to that of the posterior distribution up to a constant.

- *If we seek the maximum of the posterior distribution with respect to θ, it is unnecessary to calculate the marginal distribution. Indeed, the marginal distribution of X does not depend on θ, so the denominator is irrelevant.*

In practice, x is the vector of observed values that allows us to incorporate experience and update the prior distribution, which then becomes the posterior distribution. We refer to Bayesian statistics because the connection with Bayes' theorem is evident.

Example 2.2 *Let $P \in (0,1)$ be the probability of heads for a biased coin, and let $X_1, \dots, X_n$ represent the outcomes of n tosses of this coin. If we have no prior information about P, we may choose the prior distribution Uniform$(0,1)$, with the PDF $f_P(p) = 1$ for all $p \in (0,1)$. Given $P = p$, we model $X_1, \dots, X_n \sim$ Bernoulli(p). Then the joint distribution of $P, X_1, \dots, X_n$ is given by*

$$f_{X,P}(x_1, \dots, x_n, p) = f_{X|P}(x_1, \dots, x_n|p) f_P(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \times 1 = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

Let $s = x_1 + \cdots + x_n$. The marginal distribution of $X_1, \dots, X_n$ is obtained by integrating $f_{X,P}(x_1, \dots, x_n, p)$ over p:

$$f_X(x_1, \dots, x_n) = \int_0^1 p^s(1-p)^{n-s}\, dp = B(s+1, n-s+1),$$

where $B(x,y)$ is the Beta function defined as $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.
Thus, the posterior distribution of P given $X_1 = x_1, \dots, X_n = x_n$ has PDF

$$f_{P|X}(p|x_1, \dots, x_n) = \frac{f_{X,P}(x_1, \dots, x_n, p)}{f_X(x_1, \dots, x_n)} = \frac{1}{B(s+1, n-s+1)} p^s(1-p)^{n-s}.$$

This is the PDF of the Beta$(s+1, n-s+1)$ distribution, so the posterior distribution of P given $X_1 = x_1, \dots, X_n = x_n$ is Beta$(s+1, n-s+1)$, where $s = x_1 + \cdots + x_n$.

3 Digging into the a priori distribution

3.1 Choice of this distribution

This is often the step criticized in Bayesian statistics.
However, it is a fundamental step and represents the most significant difference from frequentist statistics. In general, this choice can be motivated by several types of considerations:

- based on similar past experiences,
- expert opinion or intuition,

- computational feasibility,
- or the desire not to introduce new information that could bias the estimation.

Computational feasibility is a point to be nuanced, as today, Monte Carlo Markov Chain methods enable to recover the a posteriori distribution thanks to simulations.

3.2 Conjugate prior

Definition 3.1 *A family $\mathcal{F}$ of probability distributions on Θ is said a conjugate prior to the statistical model $\{f(x|\theta) : x \in \mathcal{X}, \theta \in \Theta\}$ if, for all prior $\Pi \in \mathcal{F}$, the a posteriori distribution $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ also belongs to the family $\mathcal{F}$.*

There are basically two reasons why models with conjugate priors are popular:

1. they usually allow us to derive a closed-form expression for the posterior distribution;
2. they are easy to interpret, as we can easily see how the parameters of the prior change after the Bayesian update.

Example 3.1 *Here are several famous conjugate priors to classical statistical models:*

- *with discrete likelihood:*
  - *the Poisson-Gamma bayesian model, i.e.* $X|\Theta = \theta \sim \mathcal{P}(\theta), \quad \Theta \sim \mathcal{G}(\alpha, \lambda)$*;*
  - *the Normal-Normal bayesian model,*
  - *the Multinomial-Dirichlet bayesian model,*
  - *the Bernoulli-Beta model,*
  - *the Binomial-Beta model,*
  - *the Geometric-Beta model...*
- *with continuous likelihood:*
  - *the Normal-Normal model,*
  - *the Normal-Gamma model,*
  - *the Uniform-Pareto model,*
  - *the Pareto-Gamma model,*
  - *the Exponential-Gamma model,*
  - *the Gamma-Gamma model...*

The list is obviously not exhaustive!

Remark 3.2 *A conjugate prior can be determined simply looking at the likelihood and choosing a prior distribution with the same form. This is easily seen for instance in the Poisson-Gamma model, where the Poisson likelihood looks like something similar to the Gamma density. Indeed,*

$$L(\theta|x) = \prod_i e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

and

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\beta\theta} \beta^{\alpha}}{\Gamma(\alpha)}$$

We see immediately that some terms will simplify when considering the product of these two!

Finally, the impact of the prior distribution diminishes as the number of observations tends to infinity. This is logical since that means that we do not pay much attention to the expert opinion once we have enough information.

3.3 Non informative a priori distribution

In cases where we have little or no information about the parameter θ, we aim to let the data "speak for themselves" as much as possible and avoid influencing the posterior distribution. Non-informative prior distributions are often used in such cases, as they minimize any prior bias about θ.

Two common types of non-informative priors in this context are:

- **Laplace's prior**: This prior is based on the idea of assigning a uniform probability to all possible values of θ, thereby reflecting an absence of specific information. In the case of a parameter within a continuous interval, the uniform distribution is a natural choice that does not favor any particular value of θ.
- **Jeffreys' prior**: This is a more sophisticated approach that incorporates the structure of the statistical model through Fisher information. Jeffreys' prior is defined to be invariant under transformation, meaning it is minimally affected by changes in scale or reparameterization of the problem. Jeffreys' prior is especially useful in cases where a uniform prior might introduce bias due to the nature of the metric in the parameter space.

Using non-informative priors thus allows for constructing a posterior distribution that is primarily shaped by the data, ideally minimizing the influence of non-empirical information.

3.3.1 Laplace prior

Definition 3.3 *The Laplace distribution as a prior consists in considering*

$$\pi(\theta) \quad \infty \quad 1_{\theta\in\Theta}.$$

The symbol ∞ means up to some proportional coefficient, thus with same behaviour (distribution).

The Laplace prior is thus the Uniform distribution (discrete or continuous, depending on Θ). It can also be the Lebesgue measure in case of an improper distribution.

However, this prior has some drawbacks, as it can lead to an improper prior distribution and is not invariant under reparameterization.

Example 3.2 *Consider the exponential model, i.e. the a priori distribution follows an exponential distribution with parameter λ, with $\lambda > 0$. Consider now the reparameterization such that*

$$\lambda = e^{\theta}, \quad \theta \in \mathbb{R}.$$

- *Without reparameterization, we have $\pi_1(\lambda)\infty 1_{\lambda>0}$;*
- *With reparameterization, we have*

$$\pi_2(\theta) \quad \infty \quad 1_{\theta\in\mathbb{R}} \quad \Rightarrow \quad \frac{1}{\lambda}\pi_2(\ln(\lambda))1_{\lambda>0} \quad \infty \quad \frac{1}{\lambda}1_{\lambda>0}.$$

We can see that the distribution is very informative in the second case.

3.3.2 Jeffreys prior

We have seen that a good notion of a non-informative prior law is one that is invariant under reparameterization. This is not necessarily the case for a uniform law depending on the parameterization of the problem. Jeffreys' prior is based on Fisher's information.

Definition 3.4 *The Jeffreys' prior is given by*

$$\pi(\theta) \quad \infty \quad |I(\theta)|^{\frac{1}{2}},$$

where $|A|$ stands for the determinant of A.

Remark 3.5 *A few interesting remarks can be formulated in this case:*

- *This distribution is invariant by reparametrization.*
- *That can also lead to an improper distribution.*
- *It is not recommended to use this prior when $dim(\theta) > 1$.*
- *$I(\theta)$ indicates the quantity of information brought by the statistical model $f(x|\theta)$: it is therefore intuitive to overweight a priori the values of θ with higher information $I(\theta)$.*

About the reparameterization: let $\phi = h(\theta)$ where h is a C^1-diffeomorphism. Denoting by π the prior distribution on θ, then ϕ has distribution $\tilde{\pi}$ such that

$$\tilde{\pi}(\phi) = \pi(\phi)|(h^{-1})'(\phi)|.$$

Moreover, we have $\tilde{I}(\phi) = I(\phi)|(h^{-1})'(\phi)|^2$. Thus we get

$$\tilde{\pi}(\phi) \quad \infty \quad \sqrt{\tilde{I}(\phi)}.$$

This shows why we consider a square-root in the definition.

4 Bayesian inference

The result of the bayesian analysis leads to the posterior distribution $\pi(\theta|x)$, where x is the observed sample. The latter distribution embeds much more information on θ than in the frequentist framework, and we are used to summarizing it with much simpler indicators such as:

- the maximum a posteriori (MAP),
- the median a posteriori,
- the posterior mean (or mean a posteriori),
- a posterior quantile of interest.

4.1 Bayesian estimator

Definition 4.1 *Let $d()$ be a given loss function. The bayesian risk of the estimator $\hat{\theta} = t(X)$ of the parameter θ for the loss d is given by*

$$R^B(\hat{\theta}) = \mathbb{E}_{(\Theta,X)}[d(\hat{\theta},\theta)] = \mathbb{E}_{(\Theta,X)}[d(t(X),\theta)].$$

Note that this definition can be generalized to some transformation $\psi(\theta)$, if we have an estimator of $\psi(\theta)$. A crucial difference with the classical frequentist risk is that we consider the expectation of the loss under the joint distribution (Θ, X).

Proposition 4.2 *The bayesian risk is the mean of the frequentist risk considering the prior distribution π on θ. Hence,*

$$R^B(\hat{\theta}) = \mathbb{E}_{\Theta}[R(\hat{\theta},\theta)] = \int_{\Theta} R(\hat{\theta},\theta)\pi(\theta)d\theta,$$

where $R(\hat{\theta},\theta)$ is the classical frequentist risk of the estimator $\hat{\theta}$.

We now study what would be an optimal estimator in the bayesian risk framework.

Definition 4.3 *Given some loss function $d()$, the bayesian estimate is defined for all $x \in \mathcal{X}$ by*

$$t^B(x) = \arg\min_t \mathbb{E}[d(t,\theta)|X=x].$$

The bayesian estimator is thus given by $\hat{\theta}^B = t^B(X)$.

Theorem 4.4 *The bayesian estimator is the estimator minimizing the bayesian risk.*

Proposition 4.5 *The bayesian estimator associated to the quadratic risk (quadratic loss function) is given by the expectation a posteriori:*

$$\hat{\theta}^B = \mathbb{E}_\Theta[\Theta|X].$$

Remark 4.6 *In this bayesian framework,*

- *Minimizing a least square criterion still leads to the expectation as the best estimate.*
- *If the loss function would have been the L^1-risk (i.e. the loss function is the difference in absolute value), the bayesian estimator would have been the median of the posterior distribution $\Theta|X$.*
- *All these results can also be generalized to any function $\psi(\theta)$, replacing θ by $\psi(\theta)$ everywhere.*

4.2 Extension of confidence interval to the bayesian framework

Confidence intervals in the bayesian framework are also called credibility intervals. The concept is similar to that of a confidence interval but is different. In frequentist statistics, it is recalled that a confidence interval has random bounds. The level α of this interval corresponds to the proportion $(1-\alpha)$ of n realizations of this confidence interval that will contain the true parameter as n approaches infinity.

4.2.1 α-credible region

Definition 4.7 *For a given prior π, a set $C_x \subset Theta$ is a α-credible set if*

$$\mathbb{P}_\pi(\theta \in C_x \,|\, x) \geq 1-\alpha.$$

For the interpretation, one can write the latter definition with another formula:

$$\mathbb{P}_\pi(\theta \in C_x \,|\, x) = \mathbb{E}[1_{\theta\in C_x} \,|\, x] = \int_\Theta 1_{\theta \in C_x} \pi(\theta \,|\, x) d\theta.$$

We thus consider the posterior distribution, integrating the information of the past experience x into the definition of this interval.

4.2.2 Credibility interval

Let us simplify the notion of credible region by considering that θ is a scalar. In this context, $dim(\theta)=1$, and we can talk about credibility intervals.

Definition 4.8 *For a given prior π, an interval $CI_x \subset \mathbb{R}$ is a credibility interval with security coefficient $(1-\alpha)$ if*

$$\mathbb{P}_\pi(\theta \in C_x \,|\, x) = 1-\alpha.$$

We often use symmetric credibility intervals, i.e.

$$CI_x(\theta) = \left[q_\pi\left(\frac{\alpha}{2}, x\right)\ ,\ q_\pi\left(1-\frac{\alpha}{2}, x\right)\right],$$

where

$$q_\pi\,(p, x) = \inf\{u \in \mathbb{R} :\ \mathbb{P}_\pi(\theta < u\,|\,x) \geq 1-\alpha\}.$$

The main differences between a credibility interval and a confidence interval lie in that:

- a confidence interval CI has random bounds, whereas the parameter θ itself is random in credibility intervals;
- θ_0 has a probability $(1-\alpha)$ to belong to the realization of CI, whereas we have the same result but given x observed in the bayesian framework.

5 Posterior predictive distribution

Consider that $x = (x_1, ..., x_n)$ has been observed. The goal is to make an efficient prediction of the next observation x_{n+1}.

In a classical (frequentist) framework, we would use the predictor given by the expectation $\mathbb{E}(X)$, with an estimated distribution such that $X \sim f(., \hat{\theta})$ and $\hat{\theta}$ an estimator of θ. We therefore do not take into account the uncertainty underlying the estimation of θ !

Definition 5.1 *The predictive distribution is the distribution with density*

$$\pi(x_{n+1}\,|\,x) = \int f(x_{n+1}\,|\,\theta)\ \pi(\theta\,|\,x)\ d\theta.$$

Once calculated the posterior predictive density, one proceeds in the same way as in the frequentist framework, taking the expectation. The bayesian predictor thus follows

$$\hat{x}_{n+1} = \int x_{n+1}\ \pi(x_{n+1}\,|\,x)\ dx_{n+1}.$$

6 Simulations

Finally, to conclude this chapter, it is essential to mention that efficient simulation methods exist, though they are computationally expensive. These methods make it possible to approximate the posterior distribution of the parameter theta ($\pi(\theta\,|\,x)$) in a general framework that is less restrictive than the framework of conjugate prior distributions.

These techniques are out of the scope of this class, but the interested reader can go beyond the notions already seen and have a look to:

- the methods based on Monte Carlo simulations,
- the algorithms around the Monte Carlo Markov Chain (MCMC) technique such as:
  - the Gibbs sampler (when known univariate conditional distributions),
  - the Metropolis-Hastings algorithm in a more general setting,
- the hierarchical bayesian models.

7 Exercices

Exercise 1

Let $\alpha, \beta > 0$. The Beta distribution with parameters (α, β), denoted $\mathrm{Beta}(\alpha, \beta)$, is the distribution on $[0, 1]$ with density

$$\theta \mapsto \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)},$$

where the normalization constant is given by $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Its expectation is $\mathbb{E}[\Theta] = \frac{\alpha}{\alpha+\beta}$, and its variance is

$$\mathrm{Var}(\Theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

In particular, $\mathrm{Beta}(1, 1)$ is the uniform distribution on $[0, 1]$.

We consider a Bernoulli sample $X_{1:n} \sim \mathrm{B}(\theta)^{\otimes n}$, and we are interested in the parameter $\theta \in [0, 1]$. Define $S_n = X_1 + \cdots + X_n$.

1. Using a Gaussian approximation of the distribution of $\hat{\theta}_n = \frac{S_n}{n}$, statistics textbooks propose an approximate $(1 - \alpha)$ confidence interval for θ with the formulas:

   $$\left[\hat{\theta}_n - q_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}, \hat{\theta}_n + q_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}\right],$$

   where $q_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.

   What do you think of this approximate confidence interval when $x_{\mathrm{obs}} = (1, 1, \ldots, 1)$ of size n? And when $x_{\mathrm{obs}} = (0, 0, \ldots, 0, 1)$?

2. What is the posterior distribution of Θ when the prior distribution is $\mathrm{Beta}(\alpha, \beta)$?

3. What are the expectation and variance of this posterior distribution?

4. Is the posterior expectation, viewed as a function of X, a consistent estimator of θ?

5. Do credibility intervals for θ, when x_{obs} is the same as in Question 1, have the same flaws as the asymptotic confidence intervals provided in the answer to that question?

Exercise 2 (Geometric Distribution)

We consider X, a sample $X_{1:n} \,|\, \Theta = \theta \sim G(\theta)^{\otimes n}$, and the parameter θ of this distribution. Assume a prior distribution $\Theta \sim \mathrm{Beta}(\alpha, \beta)$.

1. Compute the posterior distribution of Θ and the posterior expectation.

2. Is the posterior expectation, viewed as a function of X, a consistent estimator of the parameter of interest?

Exercise 3 (Gaussian Model with Fixed Variance)

We consider the Gaussian model for a sample $X = X_{1:n} \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$, where the variance σ^2 is assumed known. The parameter of interest is $\theta = \mu$. Let $x_{1:n} = (x_1, \ldots, x_n)$ denote the observations.

1. Compute the bias and the mean squared error (MSE) of the two estimators X_1 and $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ of μ. Which has the lower error?

2. Assume prior information that $\mu \sim \mathcal{N}(m, \tau^2)$. Compute the posterior distribution.

3. Compute the Bayes estimator under the quadratic loss. What is its risk?

Exercise 4

We consider the parameter θ of a sample $X_{1:n} \sim B(\theta)^{\otimes n}$. Using different prior distributions, propose a family of admissible estimators for θ. What are their mean squared risks?

Exercise 5

We partition the parameter space into two disjoint subsets: $T = T_0 \cup T_1$, with $T_0 \cap T_1 = \emptyset$. We are interested in $\Psi(\Theta) \in \{0, 1\}$ defined as $\Psi(\theta) = 1_{\{\theta \in T_1\}}$. The loss function is $d(\hat{\psi}, \psi) = 1_{\{\hat{\psi} \neq \psi\}}$ (thus we aim to maximize $\mathbb{E}_\theta[d(\hat{\psi}, \psi)]$).

1. Compute the risk of an estimator $s(X)$ taking values in $\{0, 1\}$.

2. Given a prior distribution on T, what is the Bayesian estimator associated with this cost?

Exercise 6

We consider the case where T is a finite set and aim to estimate $\theta \in T$ from an observation x taking values in a finite set $\mathcal{X}$. Given a prior distribution, we introduce the cost $d(\hat{\theta}, \theta) = 1_{\{\hat{\theta} \neq \theta\}}$. Show that the Bayes estimator is the maximum a posteriori (MAP) estimator:

$$s(X) = \arg\max_\theta \pi(\theta \,|\, X),$$

where $\pi(\theta \,|\, x)$ is the posterior mass function, i.e., $P(\Theta = \theta \,|\, X = x)$.