A short course in Mathematical Statistics

Randal Douc

This course provides a concise introduction to mathematical statistics. We have chosen to focus on the main results, and proofs are provided when they are straightforward, informative, and not overly technical. For more in-depth details, explanations, and broader assumptions, you can refer for example to the following textbooks:

(a) "Mathematical Statistics" by P. Bickel and K. Docksum. Chapman and Hall/CRC.

(b) "Statistique et Apprentissage" by G. Fort, M. Lerasle, and E. Moulines. (in French.) Lecture notes of the course MAP433 given at Ecole Polytechnique.

The following notation is used throughout the document.

- i.i.d means independent and identically distributed.
- r.v. means random variables.
- For $r,s \in \mathbb{N}$ such that $r \leqslant s$, we write $[r:s] = \{r, r+1, \dots, s\}$.
- $X \perp\!\!\!\perp Y$ means X and Y are independent random variables.
- $X \overset{\mathcal{L}}{=} Y$ means X and Y have the same law.
- If Q is a probability measure, $X \sim Q$ means that the random variable X follows the distribution Q. By abuse of notation, if Q has a density q with respect to some dominating measure μ, then when no ambiguity occurs, we may write $X \sim q$ instead of $X \sim Q$.
- $\liminf_n a_n = \lim_{n\to\infty} (\inf_{k\geqslant n} a_k)$ and similarly, $\limsup_n a_n = \lim_{n\to\infty} \left(\sup_{k\geqslant n} a_k\right)$. Moreover, $\lim_n a_n$ exists if and only if $\liminf_n a_n = \limsup_n a_n$.
- For any $a \in \mathbb{R}$, $a^+ = \max(a,0)$ and $a^- = \max(-a,0) = -\min(a,0)$ and we have $|a| = a^+ + a^-$ and $a = a^+ - a^-$.

Other notation will be introduced progressively in the Lecture Notes.

Contents

Preliminary on measure theory and integration

In this introductory chapter, we offer a concise overview of measure and integration theory, with a specific emphasis on their application in probability theory.

0.1 Sigma-fields, Measures and Probability

Let us start with the most basic concept in probability: sigma-fields!

Definition 0.1. Let Ω be a given set. We say that a family of sets $\mathcal{F} \subset \mathcal{P}(\Omega)$ is a sigma-field on Ω if and only if the three following properties are satisfied

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

(ii) if $A \in \mathcal{F}$ then $\bar{A} = \Omega \setminus A \in \mathcal{F}$,

(iii) if for all $i \in \mathbb{N}$, $A_i \in \mathcal{F}$ then $\cap_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

We then say that $(\Omega, \mathcal{F})$ is a measurable space.

A sigma-field is stable by complementary sets, countable intersection, countable union and also by taking the "set difference" $\setminus$ in the sense that if $A, B \in \mathcal{F}$, then $A \setminus B \in \mathcal{F}$ (indeed, just write $A \setminus B = A \cap B^c$.)

►**Q-0.1.** Do those properties have special names?

The second property is often called *stability by complementary sets* and the last one *stability by countable intersection*. You may also find in the literature some other *equivalent* definitions:

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

(ii) if $A \in \mathcal{F}$ then $\bar{A} = \Omega \setminus A \in \mathcal{F}$,

(iii) if for all $i \in \mathbb{N}$, $A_i \in \mathcal{F}$ then $\cup_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

But I prefer the way it is expressed in Definition 0.1.

►**Q-0.2.** Why do you need these properties?

In the theory of probability, a set A will typically correspond to an event that may occur, it can be expressed as a constraint with respect to all the possibilities. The fact that we ask stability by complementary sets or by countable intersections corresponds to considering either the complementary event or the fact that all the events A_i are satisfied.

►**Q-0.3.** Do you have any examples in mind?

The smallest sigma-field is $\mathcal{F} = \{\Omega, \emptyset\}$ and the largest one is $\mathcal{P}(\Omega)$. Often, sigma-fields we are interested in, are generated by some family of sets...

►**Q-0.4.** What do you mean ?

Say that you are interested in a family of sets $\mathcal{C} \subset \mathcal{P}(\Omega)$ but unfortunately, some of the 3 properties that define sigma-fields are not satisfied for $\mathcal{C}$. In that case, we can still include $\mathcal{C}$ into a larger family so that it is a sigma-field. We can even consider the "smallest" one (in a sense to be defined), which contains all the sets in $\mathcal{C}$.

Definition 0.2. Let $\mathcal{C} \subset \mathcal{P}(\Omega)$. There exists a sigma-field, named $\sigma(\mathcal{C})$ which contains $\mathcal{C}$ and which is minimal for the inclusion, that is, any other sigma-field that contains $\mathcal{C}$ also contains $\sigma(\Omega)$. We then say that $\sigma(\mathcal{C})$ is the sigma-field generated by $\mathcal{C}$.

A valuable exercise is to prove the property stated in the above definition. This can be achieved by defining $\sigma(\mathcal{C})$ as the collection of all sets that belong to any sigma field containing $\mathcal{C}$. In other words, define

$$\mathcal{A} = \cap \{\mathcal{T} \ : \ \mathcal{C} \subset \mathcal{T} \text{ and } \mathcal{T} \text{ is a sigma-field}\} \ .$$

Although $\mathcal{A}$ is defined as an uncountable intersection, you can verify that $\mathcal{A}$ is a sigma-field. Furthermore, this sigma-field contains all the sets in $\mathcal{C}$, and any other sigma-field that contains all the sets in $\mathcal{C}$ must necessarily contain all the sets in $\mathcal{A}$. In conclusion, $\mathcal{A}$ is the smallest sigma-field satisfying this property, and we can refer to it as $\sigma(\mathcal{C})$.

►**Q-0.5.** What is your favorite example?

An important example is the case of open sets. If Ω is $\mathbb{R}^k$ and $\mathcal{C}$ is the family of open sets on Ω, then the sigma-field generated by open sets is called the *Borel sigma-field* and is noted $\mathcal{B}(\Omega)$ and any set $A \in \mathcal{B}(\Omega)$ is called a *borelian* set.

►**Q-0.6.** For different family of sets, if you consider the sigma-fields generated by each of them, do you systematically find different sigma-fields?

Of course not... In practice, if we have two family of sets $\mathcal{C} \subset \mathcal{P}(\Omega)$ and $\mathcal{D} \subset \mathcal{P}(\Omega)$ and if we want to check if $\sigma(\mathcal{C}) = \sigma(\mathcal{D})$ then a necessary and sufficient condition for getting that is to check successively that $\mathcal{D} \subset \sigma(\mathcal{C})$ and $\mathcal{C} \subset \sigma(\mathcal{D})$. You can use this property for checking that the sigma-field generated by open sets (i.e. the Borel sigma-field) is also the sigma-field generated by closed sets.

►**Q-0.7.** That's nice. You can prove it easily?

Yes, please do so. It's a good way to check that everything is understandable.

Definition 0.3. Let $(\Omega, \mathcal{F})$ be a measurable space. We say that a function $\mu : \mathcal{F} \to \bar{\mathbb{R}}^+ := \mathbb{R}^+ \cup \{\infty\}$ is a measure if it satisfies the *sigma-additivity property*, that is for any family of sets (A_i) such that $A_i \in \mathcal{F}$ for any $i \in \mathbb{N}$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

$$\mu(\cup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \mu(A_i) \ . \tag{1}$$

We then say that $(\Omega, \mathcal{F}, \mu)$ is a measured space. Moreover, if $\mu(\Omega) = 1$ then we say that μ is a probability measure.

Note that in the right hand side of (1), we sum quantities in $\bar{\mathbb{R}}^+$ that is, we use the convention that if $a \in \mathbb{R}^+$, $a + \infty = \infty$ and $\infty + \infty = \infty$.

►**Q-0.8.** The measure μ evaluated on a set A can be infinite?

Yes of course, $\mu(A)$ takes its values between 0 and $\mu(\Omega)$ actually... But if you consider a measure of probability (which is nothing but a particular measure), then the values of $\mu(A)$ are between 0 and 1.

►**Q-0.9.** Some useful properties?

(i) If $A \subset B$, then $\mu(A) \leqslant \mu(B)$.

(ii) $\mu(\cup_{i=1}^{\infty} A_i) \leqslant \sum_{i=1}^{\infty} \mu(A_i)$.

(iii) $\mu(\cup_{i=1}^{\infty} A_i) = \lim_{n\to\infty} \mu(\cup_{i=1}^{n} A_i)$.

(iv) If $\mu(A_1) < \infty$, then, $\mu(\cap_{i=1}^{\infty} A_i) = \lim_{n\to\infty} \mu(\cap_{i=1}^{n} A_i)$.

►**Q-0.10.** What are the typical measures I will deal with?

There are at least two fundamental examples of measures:

(i) The Dirac measure on a which is defined by $\delta_a(A) = 0$ if $a \notin A$ and 1 otherwise.

(ii) The Lebesgue measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which is defined with the following characterizing property: it is the only measure such that for any segment $A = [a,b]$, we have $\lambda(A) = b - a$. Then, you can show easily that the Lebesgue measure of any interval (the interval may be closed, open or none of them) is the length of the interval (which thus may be infinite, take $A = [1,\infty[$ for example).

From these measures, you can construct other measures, for example by multiplying them by some non negative measurable functions.

0.2 Integrals, random variables, expectation

►**Q-0.11.** You said measurable functions?

Yes. The definition is below.

Definition 0.4. If $(A, \mathcal{A})$ and $(B, \mathcal{B})$ are measurable sets. We say that $h : A \to B$ is a $\mathcal{A}/\mathcal{B}$ measurable function if and only if for any $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{A}$.

Of course, if $\mathcal{B} = \sigma(\mathcal{C})$, then instead of checking for any $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{A}$, we may only check for any $C \in \mathcal{C}$, $f^{-1}(C) \in \mathcal{A}$. For a given measurable function f, we may define $\sigma\left\{f^{-1}(B) \,:\, B \in \mathcal{B}\right\}$, it turns out that it is a sigma field, called $\sigma(f)$. Measurable functions are linked with random variables...

►**Q-0.12.** Can you be more precise?

Here is the definition of a random variable.

Definition 0.5. Let $(\Omega, \mathcal{F})$ be a measurable space and consider the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A random variable $X : \Omega \to \mathbb{R}$ is, by definition, a $\mathcal{F}/\mathcal{B}(\mathbb{R})$ measurable function.

►**Q-0.13.** You mean that a random variable is nothing more than a measurable function?

Yes, and it always takes real-valued outcomes. If X takes its values in $\mathbb{R}^k$, we refer to it not as a random variable but as a random vector, and in some books, it is also referred to as a random element. Now that we have defined general measures, including the important particular case of probability measures, we can define the *integral* associated with a measure μ. If μ is a probability measure, we will define the *expectation* of a random variable.

Recall that a measure μ associates a real number $\mu(A)$ to a set $A \in \mathcal{F}$. Now, we want μ to associate a real number, denoted $\mu(f)$ or $\int_\Omega f(w)\mu(\mathrm{d}w)$, to any real-valued $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable function f. Here, there is an abuse of notation. Stricto sensu: $\mu(A)$ is well-defined, but $\mu(f)$ is an abuse of notation because, within the brackets, we have a function f and not a set A. To avoid confusion, most of the time we write $\int f(w)\mu(\mathrm{d}w)$ or $\int f\,\mathrm{d}\mu$ (to avoid w) instead of $\mu(f)$. However, all these notations refer to the same object.

Actually, it will be not be possible to define $\mu(f)$ for any measurable function... Let us be more precise. The construction of the integral wrt μ will be done progressively. We start with the $\mu(\mathbf{1}_A)$. By definition, we set:

$$\mu(\mathbf{1}_A) = \mu(A) \ .$$

Then, we define

$$\mu(\sum_{i=1}^{n} \alpha_i \mathbf{1}_{A_i}) = \sum_{i=1}^{n} \alpha_i \mu(A_i) \ .$$

Then, for any measurable *non-negative* function f,

$$\mu(f) = \sup \left\{ \mu(\sum_{i=1}^{n} \alpha_i \mathbf{1}_{A_i}) \ : \ \sum_{i=1}^{n} \alpha_i \mathbf{1}_{A_i} \leqslant f \right\} \ .$$

Finally, for any measurable function such that $\mu(|f|) < \infty$, we set

$$\mu(f) = \mu(f^+) - \mu(f^-) \ .$$

►**Q-0.14.** You always integrate on the whole space?

Yes, but if you integrate a function f on a subset $\Omega_0 \in \mathcal{F}$ where $\Omega_0 \subset \Omega$, then by definition, it just means $\int f(w)\mathbf{1}_{\Omega_0}(w)\mu(\mathrm{d}w)$. That is, you integrate on the whole space but thanks to the indicator function $\mathbf{1}_{\Omega_0}$, only the values of f on Ω_0 are meaningful.

►**Q-0.15.** You told me that two examples of measures were important.

So what?

►**Q-0.16.** What are the integrals associated to those measures?

Our two important examples of integrals constructed from measures μ are

(i) Integrals associated to Dirac measures... We can show that for any $a \in \Omega$ and any measurable function f, $\int f(w)\delta_a(\mathrm{d}w) = f(a)$.

(ii) Integrals associated to the Lebesgue measure... This is a common case, and instead of writing $\int f(w)\lambda(\mathrm{d}w)$, we usually write $\int f(w)\mathrm{d}w$.

►**Q-0.17.** OK for the construction of the integral but what are the essential properties?

I guess the very essential ones allow to interchange limit and integral. Two of them are essential:

(i) **The monotone convergence theorem**: if $\{f_n, \ n \in \mathbb{N}\} n \in \mathbb{N}$ is a family of measurable *non-negative* functions and $f_n \leqslant f_{n+1}$ for all large enough n, then $\int \lim_{n\to\infty} f_n(w)\mu(\mathrm{d}w) = \lim_{n\to\infty} \int f_n(w)\mu(\mathrm{d}w)$

(ii) **The Lebesgue dominated convergence theorem:** if $\{f_n \ : \ n \in \mathbb{N}\}$ is a family of measurable functions such that $\lim_{n\to\infty} f_n(w)$ exists for μ-almost all $w \in \Omega$ and if $|f_n| \leqslant h$ where $\int h d\mu < \infty$, then $\int \lim_{n\to\infty} f_n(w)\mu(\mathrm{d}w) = \lim_{n\to\infty} \int f_n(w)\mu(\mathrm{d}w)$

But there are also essential properties: the linearity of the integral, or if $f \leqslant g$ then, $\mu(f) \leqslant \mu(g)$, or $|\mu(f)| \leqslant \mu(|f|)$. Or if $\mu(|f|) < \infty$ then, $|f(w)| < \infty$ for μ-almost all $w \in \Omega$.

►**Q-0.18.** Anything else?

Humm, let me think... Yes! **The Fatou lemma!** It can be quite useful sometimes, especially because the assumptions are very light. Let $\{f_n \ : \ n \in \mathbb{N}\}$ be a family of non-negative measurable functions. Then,

$$\liminf_n \int f_n(w)\mu(\mathrm{d}w) \geqslant \int \liminf_n f_n(w)\mu(\mathrm{d}w) \ .$$

►**Q-0.19.** You said that when you multiply a non-negative measurable function and a measure, it is a measure... What do you mean exactly?

If f is a non-negative measurable function, then, the measure $f\mathrm{d}\mu$ is defined by: $A \mapsto \int \mathbf{1}_A(w)f(w)\mu(\mathrm{d}w)$. Therefore, with the two typical measures (δ_a and λ), you can define so many different measures $f\mathrm{d}\delta_0$ or $f\mathrm{d}\lambda$ by using a non-negative measurable function f.

►**Q-0.20.** You told me there is some link between the expectation operator associated to a probability measure and the integral associated to a measure.

The expectation is defined in the same way as integrals associated to some measure: it is constructed from a measured space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X (that is a $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable function). Then, by definition, $\mathbb{E}[X]$ is just the integral associated to the measure $\mathbb{P}$ taken at the random variable X, that is, $\mathbb{E}[X] = \int_\Omega X(w)\mathbb{P}(\mathrm{d}w)$.

Definition 0.6. The law of a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is the measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by: $\mu : A \mapsto \mathbb{P}(X \in A)$.

In this definition, we use $\mathbb{P}(X \in A)$ which means $\mathbb{P}(\Omega_0)$ where $\Omega_0 = \{w \in \Omega \,:\, X(w) \in A\} := \{X \in A\}$. We also call this measure, the push-forward measure of $\mathbb{P}$ through the function X. It defines a measure on the arrival sigma field.

►**Q-0.21.** How do you check that two random vectors are independent?

Actually, two random vectors X, Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ but taking values in $\mathbb{R}^p$ and $\mathbb{R}^d$ are independent if and only one the equivalent properties are satisfied:

(i) For all $(A, B) \in \mathcal{B}(\mathbb{R}^p) \times \mathcal{B}(\mathbb{R}^d)$,

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

(ii) For all bounded or non-negative $\mathcal{B}(\mathbb{R}^p)/\mathcal{B}(\mathbb{R})$-measurable functions $f : \mathbb{R}^p \to \mathbb{R}$ and for all $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$-measurable bounded or non-negative functions $g : \mathbb{R}^d \to \mathbb{R}$, we have

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

(iii) For all $(u, v) \in \mathbb{R}^p \times \mathbb{R}^d$,

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}u^T X + \mathrm{i}v^T Y}] = \mathbb{E}[\mathrm{e}^{\mathrm{i}u^T X}]\mathbb{E}[\mathrm{e}^{\mathrm{i}v^T Y}]\ .$$

(iv) For all $(u, v) \in \mathbb{R}^p \times \mathbb{R}^d$ and all $(x, y) \in \mathbb{R}^2$,

$$\mathbb{P}(u^T X \leqslant x, v^T Y \leqslant y) = \mathbb{P}(u^T X \leqslant x)\mathbb{P}(v^T Y \leqslant y)\ .$$

Also, $\boxed{\text{if } X \perp\!\!\!\perp Y, \text{ then for any measurable functions } h_0 \text{ and } h_1, \text{ we have } h_0(X) \perp\!\!\!\perp h_1(Y)}$.

►**Q-0.22.** Ok, thanks for all these equivalent formulations. Now, in the same spirit, if I want to check that two random variables have the same law, what are the tools I can use?

We have $X \overset{\mathcal{L}}{=} Y$ (where X and Y are two random variables) if and only if any of the following conditions holds true:

(i) $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for any $A \in \mathcal{B}(\mathbb{R})$.

(ii) $\mathbb{P}(X \leqslant t) = \mathbb{P}(Y \leqslant t)$ for any $t \in \mathbb{R}$. Note that $t \mapsto \mathbb{P}(X \leqslant t)$ is the cumulative distribution function for the random variable X.

(iii) $\mathbb{E}[\mathrm{e}^{\mathrm{i}uX}] = \mathbb{E}[\mathrm{e}^{\mathrm{i}uY}]$ for any $u \in \mathbb{R}$. Note that $u \mapsto \mathbb{E}[\mathrm{e}^{\mathrm{i}uX}]$ is the characteristic function for the random variable X.

Chapter 1

Probability refresher

1.1 A recap on some classical laws.

Let us start with some classical distributions. We put a brief description of these distributions in one place for future reference.

1.1.1 Some discrete-valued distributions

The Bernoulli distribution

Y follows a Bernoulli distribution with parameter $\theta \in [0,1]$ (and we write $Y \sim \mathcal{B}(\theta)$) if and only if $\mathbb{P}_\theta(Y = 1) = \theta = 1 - \mathbb{P}_\theta(Y = 0)$. And we have $\mathbb{E}_\theta[Y] = \theta$ and $\mathbb{V}\mathrm{ar}_\theta(Y) = \theta(1-\theta)$.

The Binomial distribution

Y follows a Binomial distribution with parameter (n,p) (and we write $Y \sim Bin(n,p)$) if and only if $\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$. And we have $\mathbb{E}_p[Y] = np$ and $\mathbb{V}\mathrm{ar}_p(Y) = np(1-p)$.

The Geometric distribution

Y follows a Geometric distribution with parameter p (and we write $Y \sim Ge(p)$) if and only if $\mathbb{P}_p(Y = k) = p(1-p)^{k-1}$. And we have $\mathbb{E}_p[Y] = 1/p$ and $\mathbb{V}\mathrm{ar}_p(Y) = (1-p)/p^2$.

The Poisson distribution

Y follows a Poisson distribution with parameter λ (and we write $Y \sim \mathcal{P}(\lambda)$) if and only if $\mathbb{P}_\lambda(Y = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$. And we have $\mathbb{E}_\lambda[Y] = \lambda$ and $\mathbb{V}\mathrm{ar}_\lambda(Y) = \lambda$.

1.1.2 Some real-valued distributions

Name	Acronym	Parameter	density function: $f_X(x)$
Normal	$\mathcal{N}(\mu,\sigma^2)$	(μ,σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$
Exponential	$\mathbf{exp}(\lambda)$	$\lambda > 0$	$\lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}^+}(x)$
Gamma	$\Gamma(k,\theta)$	$(k,\theta) \in \left(\mathbb{R}_+^*\right)^2$	$\frac{x^{k-1}\theta^k e^{-\theta x}}{\Gamma(k)} \mathbf{1}_{\mathbb{R}^+}(x)$

In the above description,

(i) if $X_i \sim \Gamma(k_i,\theta)$ and (X_i) are independent, then $\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n k_i, \theta)$.

(ii) The density of the Gamma distribution involves the constant $\Gamma(k)$. Actually,

$$\Gamma(k) = \begin{cases} \int_0^\infty t^{k-1}\mathrm{e}^{-t}\mathrm{d}t & \text{if } k \in \mathbb{R}_+^* \\ (k-1)! & \text{if } k \in \mathbb{N}. \end{cases} \quad (\blacktriangleright\text{GAMMA FUNCTION})$$

Definition 1.1.

- **The Chi-square distribution**. Let (X_i) be i.i.d. with $X_i \sim \mathcal{N}(0,1)$. Then,

$$\sum_{i=1}^{p} X_i^2 \sim \chi_2(p) \quad (\blacktriangleright\text{THE } \chi_2\text{-DISTRIBUTION WITH } p \text{ DEGREES OF FREEDOM})$$

- **The Student distribution**. Let (U,V) be independent with $U \sim \mathcal{N}(0,1)$, $V \sim \chi_2(p)$. Then,

$$\frac{U}{\sqrt{\frac{V}{p}}} \sim \mathcal{T}(p) \quad (\blacktriangleright\text{THE } t\text{-DISTRIBUTION WITH } p \text{ DEGREES OF FREEDOM})$$

- **The Fisher distribution**. Let (U,V) be independent where $U \sim \chi_2(n_1)$, $V \sim \chi_2(n_2)$. Then,

$$\frac{U/n_1}{V/n_2} \sim \mathcal{F}(n_1,n_2) \quad (\blacktriangleright\text{THE FISHER-DISTRIBUTION WITH } (n_1,n_2) \text{ DEGREES OF FREEDOM})$$

1.2 A brief survey on limit theorems

Unless otherwise stated, all the random variables in this section will be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

In this course, several notions of convergence for random variables will be needed.

Definition 1.2. We say that a sequence of $\mathbb{P}$-a.s. finite random variables $\{X_n : n \in \mathbb{N}\}$ *converges in distribution* (or converges in law) to a $\mathbb{P}$-a.s. finite random variable X if and only if any of the following equivalent statements is satisfied.

(a) For all bounded continuous functions h, $\lim_n \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$.

(b) For all $A \in \mathcal{B}(\mathbb{R})$ such that $\mathbb{P}(X \in \partial A) = 0$, $\lim_n \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$.

(c) For all $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) = 0$, $\lim_n \mathbb{P}(X_n \leqslant x) = \mathbb{P}(X \leqslant x)$.

(d) For all $u \in \mathbb{R}$, $\lim_n \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX_n}\right] = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right]$.

► **Notation:** In this case, we write $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X$.

In (b), the notation ∂A means the frontier of A, that is, the set of points x such that any neighborhood of x contains at least an element of A and an element of A^c, which are different from x.

Remark 1.3. Note that (c) corresponds to the convergence of the cumulative distribution functions and (d) corresponds to the convergence of the characteristic functions.

Remark 1.4. If $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X$ and if X has a distribution, say $\mathcal{N}(0,1)$, we often write, for ease of reading, $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} \mathcal{N}(0,1)$ instead of $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X$ with $X \sim \mathcal{N}(0,1)$.

►**Q-1.1.** Which characterization do you prefer?

None of them. Nevertheless, often, we check (c) or (d) to obtain that $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X$ and conversely, once $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X$ is established, we often derive other properties by noting that (a) or (b) are then satisfied.

We now need two other notions of convergence. In the next two definitions, all the random variables, $(X_n)_{n\in\mathbb{N}}$ and X, are defined on the same probability space.

Definition 1.5. We say that a sequence of $\mathbb{P}$-a.s. finite random variables $\{X_n : n \in \mathbb{N}\}$ *converges in probability* to a $\mathbb{P}$-a.s. finite random variable X if and only if

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

► **Notation:** In this case, we write $X_n \overset{\mathbb{P}-prob}{\longrightarrow} X$.

Definition 1.6. We say that a sequence of $\mathbb{P}$-a.s. finite random variables $\{X_n : n \in \mathbb{N}\}$ *converges almost-surely* to X if and only if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1\ .$$

► **Notation:** In this case, we write $X_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} X$ or sometimes, $\lim_{n\to\infty} X_n = X$, $\mathbb{P}-a.s.$

►**Q-1.2.** Three different notions of convergence! Amazing! Are they completely disconnected?

No. Actually, these three notions of convergence are related to each other according to the following lemma.

Lemma 1.7. Let $\{X_n : n \in \mathbb{N}\}$ be a sequence of random variables, X be a random variable, all of them defined on the same probability space and being $\mathbb{P}$-a.s. finite. Let c be a constant. Then we have the following implications and equivalence:

$$\begin{aligned}
X_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} X \quad &\Longrightarrow \quad X_n \overset{\mathbb{P}-prob}{\longrightarrow} X\ .\\
X_n \overset{\mathbb{P}-prob}{\longrightarrow} X \quad &\Longrightarrow \quad X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X\ .\\
X_n \overset{\mathbb{P}-prob}{\longrightarrow} c \quad &\Longleftrightarrow \quad X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} c\ .
\end{aligned}$$

In words for the last equivalence, convergence in probability to a **constant** is equivalent to convergence in distribution to this **constant**.

►**Q-1.3.** Ok, thanks for all these implications. Now, assuming that you have one of the previous convergence results, what can we say next? What does it imply?

It may be useful to apply to a function f to the previous convergence results. More precisely, we already know that if two random variables X, Y share the same law, i.e. $X \overset{\mathcal{L}}{=} Y$, then for any measurable function $f : \mathbb{R} \to \mathbb{R}$, $f(X) \overset{\mathcal{L}}{=} f(Y)$. If instead of having the equality in law "$X \overset{\mathcal{L}}{=} Y$", we have the limiting result "X_n converges to X" according to one of the types of convergence just seen before, then we may wonder whether $f(X_n)$ actually converges to $f(X)$ according to the corresponding type of convergence. This motivates the following lemma.

Theorem 1.8 (The continuous mapping theorem). Let $\{X_n : n \in \mathbb{N}\}$ be a sequence of random variables, X be a random variable , all of them defined on the same probability space and being $\mathbb{P}$-a.s. finite. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then we have the following implications:

$$X_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} X \implies f(X_n) \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} f(X)\,. \tag{1.1}$$

$$X_n \overset{\mathbb{P}-prob}{\longrightarrow} X \implies f(X_n) \overset{\mathbb{P}-prob}{\longrightarrow} f(X)\,. \tag{1.2}$$

$$X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X \implies f(X_n) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} f(X)\,. \tag{1.3}$$

A practical consequence of this lemma is that from any convergence result in this course, we may obtain many others by just applying any continuous function (note that this continuous function is not necessarily bounded). We now move on the Slutsky theorem which be used repeatedly in the sequel.

Theorem 1.9 (The Slutsky Theorem). If $X_n \overset{\mathbb{P}-prob}{\longrightarrow} c$ where c is a constant and if $Z_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} Z$, then $(X_n, Z_n) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} (c, Z)$, that is, for any real-valued continuous function f, we have $f(X_n, Z_n) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} f(c, Z)$.

PROOF. Under the assumptions of the Lemma, we will show that the random vector (X_n, Z_n) converges in distribution to (c, Z) and the last statement of the lemma would then derive from (1.3) with X_n replaced by (X_n, Z_n). To establish the convergence in law of (X_n, Z_n), we will show the convergence of the associated bi-dimensional characteristic function:

$$\lim_{n\to\infty} \mathbb{E}\left[e^{i(uX_n+vZ_n)}\right] = \mathbb{E}\left[e^{i(uc+vZ)}\right]\,. \tag{1.4}$$

Indeed,

$$\mathbb{E}\left[e^{i(uX_n+vZ_n)}\right] - \mathbb{E}\left[e^{i(uc+vZ)}\right] = \mathbb{E}[e^{i(uX_n+vZ_n)} - e^{iuc}e^{ivZ}] = \mathbb{E}[B_n + C_n]\,. \tag{1.5}$$

where we have set

$$B_n = e^{i(uX_n+vZ_n)} - e^{i(uc+vZ_n)} = e^{i(uc+vZ_n)}\left(e^{iu(X_n-c)} - 1\right)$$
$$C_n = e^{i(uc+vZ_n)} - e^{i(uc+vZ)} = e^{iuc}\left(e^{ivZ_n} - e^{ivZ}\right)\,.$$

Now, define $\varphi_0 : \mathbb{R} \to \mathbb{R}$ and $\varphi_1 : \mathbb{R} \to \mathbb{C}$ where $\varphi_0(x) = |e^{iu(x-c)} - 1|$ and $\varphi_1(x) = e^{ivx}$. These functions are continuous and bounded and $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} c$ and $Z_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} Z$. Then, by (a) in Definition 1.2,

$$\mathbb{E}[|B_n|] \leqslant \mathbb{E}\left[|e^{iu(X_n-c)} - 1|\right] = \mathbb{E}[\varphi_0(X_n)] \to_{n\to\infty} \mathbb{E}[\varphi_0(c)] = \mathbb{E}\left[|e^{iu(c-c)} - 1|\right] = 0$$
$$|\mathbb{E}[C_n]| = |\mathbb{E}[e^{ivZ_n}] - \mathbb{E}[e^{ivZ}]| = |\mathbb{E}[\varphi_1(Z_n)] - \mathbb{E}[\varphi_1(Z)]| \to_{n\to\infty} |\mathbb{E}[\varphi_1(Z)] - \mathbb{E}[\varphi_1(Z)]| = 0\,.$$

Hence $\lim_n \mathbb{E}[B_n + C_n] = 0$. Combining with (1.5) shows (1.4) and the proof is completed. ■

►**Q-1.4.** You said that Slutsky's theorem is an important tool. Could you please provide a simple illustration?

Sure, below is a simple illustration. In this example, we will use the fact that any converging sequence of real numbers is also a sequence of random variables that converges in probability or almost surely. This is because a constant is just a particular random variable.

►**Q-1.5.** Great! I am looking forward to seeing that.

I am at your service.

Example 1.10 (Illustration of the Slutsky theorem)**.** Assume that there exists a sequence of real numbers $\{r_n : n \in \mathbb{N}\}$ such that $\lim_n r_n = \infty$ and $Z_n := r_n(X_n - a) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} Z$. Then, we will show that $X_n \overset{\mathbb{P}-prob}{\longrightarrow} a$.

Note that by assumption, $U_n := 1/r_n \to 0$. In other words, $\{U_n \,:\, n \in \mathbb{N}\}$ is a sequence of real numbers that converges to 0. Since a constant is a particular random variable, we can always consider $\{U_n \,:\, n \in \mathbb{N}\}$ as a sequence of random variables and we have $U_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$.

Moreover, by assumption, $Z_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} Z$. Applying Slutsky's theorem (Theorem 1.9) to the continuous function $f : (u,z) \mapsto uz + a$ is continuous yields

$$X_n = U_n Z_n + a = f(U_n, Z) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} f(0,Z) = 0 \times Z + a = a\,.$$

Hence $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} a$, and according to Theorem 1.8, this is equivalent to $X_n \overset{\mathbb{P}-prob}{\longrightarrow} a$.

►**Q-1.6.** Before going further, say that you have a sequence of non-negative integrable random variables (X_n) that converges to a random variable X according to one of the types of convergence you recalled. Can you say something about $\lim_n \mathbb{E}[X_n]$? Does it exist and if so, is it equal to $\mathbb{E}[X]$?

Excellent question! If X_n converges $\mathbb{P}-a.s.$ to X and if there exists an integrable random variable Y such that $0 \leqslant X_n \leqslant Y$, then, the dominated convergence theorem shows that $\lim_n \mathbb{E}[X_n] = \mathbb{E}[\lim_n X_n] = \mathbb{E}[X]$. If it is not almost-sure convergence but convergence in law, you can use the following lemma, where (1.6) below introduces the notion of *uniform integrability*.

Lemma 1.11. Assume that $\{X_n \,:\, n \in \mathbb{N}\}$ is a sequence of non-negative and integrable random variables such that

$$\lim_{M\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}\left[X_n \mathbf{1}_{X_n \geqslant M}\right] = 0\,. \quad (\blacktriangleright \text{UNIFORM INTEGRABILITY}) \tag{1.6}$$

Then, $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} X$ implies that $\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$.

►**Q-1.7.** I wonder if uniform integrability is easy to check?

It is not always straightforward, but if you are aiming for uniform integrability, the following approach can be occasionally fruitful. Since for any $\alpha > 0$, $\mathbf{1}_{\{X_n \geqslant M\}} \leqslant X_n^\alpha / M^\alpha$, we deduce that

$$\mathbb{E}\left[X_n \mathbf{1}_{\{X_n \geqslant M\}}\right] \leqslant \mathbb{E}[X_n^{1+\alpha}]/M^\alpha\,.$$

Hence, if $\sup_{n\in\mathbb{N}} \mathbb{E}[X_n^{1+\alpha}] < \infty$, then (1.6) is satisfied.

We end up this section by two fundamental results that we state and prove.

Theorem 1.12 (The Strong Law of Large Numbers). Assume that $(X_i)_{i\in\mathbb{N}}$ are iid random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[|X_1|] < \infty$. Then, setting $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$, we have

$$\lim_{n\to\infty} \bar{X}_n = \mathbb{E}[X_1]\,, \quad \mathbb{P}-a.s.$$

► **Notation:** In what follows, SLLN stands for the Strong Law of Large Numbers.

PROOF. We start with a preliminary result underpinning the proof.

Lemma 1.13. Let (Y_i) be iid random variables such that $\mathbb{E}[|Y_1|] < \infty$ and $\mathbb{E}[Y_1] > 0$, then $\mathbb{P}$-a.s.,

$$\liminf_n S_n/n \geqslant 0\,,$$

where $S_n = \sum_{i=1}^n Y_i$.

► *(Proof of the lemma.)* Set $L_n = \inf(S_k, k \in [1:n])$, $L_\infty = \inf(S_k, k \in \mathbb{N}^*)$, $A = \{L_\infty = -\infty\}$. Let $\theta(y_1, y_2, \ldots,) = (y_2, y_3, \ldots)$ be the shift operator. Then, $\mathbb{P}$-a.s.,

$$\begin{aligned} L_n &= S_1 + \inf(0, S_2 - S_1, \ldots, S_n - S_1) = Y_1 + \inf(0, L_{n-1}\circ\theta) \\ &\geqslant Y_1 + \inf(0, L_n \circ \theta) = Y_1 - L_n^- \circ \theta. \end{aligned}$$

where the inequality follows from the fact that $n \mapsto L_n$ is non-increasing. This implies $\mathbb{P}$-a.s. (since $L_n^- \circ \theta$ is a.s. finite)

$$\mathbf{1}_A Y_1 \leqslant \mathbf{1}_A L_n + \mathbf{1}_A L_n^- \circ \theta \,.$$

Taking the expectation on both sides and then, using $\mathbb{P}(\mathbf{1}_A = \mathbf{1}_A \circ \theta) = 1$, and the strong stationarity of the sequence:

$$\mathbb{E}[\mathbf{1}_A Y_1] \leqslant \mathbb{E}[\mathbf{1}_A L_n] + \mathbb{E}[\mathbf{1}_A \circ \theta \; L_n^- \circ \theta] = \mathbb{E}[\mathbf{1}_A L_n] + \mathbb{E}[\mathbf{1}_A L_n^-] = \mathbb{E}[\mathbf{1}_A L_n^+] \to 0 \,,$$

where the right-hand side tends to 0 by the dominated convergence theorem since a.s. $\lim_n \mathbf{1}_A L_n^+ = \mathbf{1}_A L_\infty^+ = 0$ and $0 \leqslant \mathbf{1}_A L_n^+ \leqslant Y_1^+$. Finally $\mathbb{E}[\mathbf{1}_A Y_1] \leqslant 0$. Therefore, noting that $\mathbf{1}_A \circ \theta$ is independent from Y_1,

$$0 \geqslant \mathbb{E}[\mathbf{1}_A Y_1] = \mathbb{E}[\mathbf{1}_A \circ \theta \; Y_1] = \mathbb{E}[\mathbf{1}_A \circ \theta]\mathbb{E}[Y_1] = \underbrace{\mathbb{E}[\mathbf{1}_A]}_{\geqslant 0} \underbrace{\mathbb{E}[Y_1]}_{>0} \,.$$

This implies $\mathbb{P}(A) = 0$ and the lemma is proved.◀

(Proof of the Theorem.) We now turn to the proof of the Theorem. Without loss of generality, we assume that $\mathbb{E}[X_1] = 0$. Applying Lemma 1.13 with $Y_i = X_i + \varepsilon$ (where $\varepsilon > 0$), we get $\liminf_n n^{-1} \sum_{i=1}^n X_i \geqslant -\varepsilon$, $\mathbb{P}$-a.s. And applying again Lemma 1.13 with $Y_i = -X_i + \varepsilon$, we get $\mathbb{P}$-a.s., $\limsup_n n^{-1} \sum_{i=1}^n X_i \leqslant \varepsilon$ which finishes the proof since ε is arbitrary. ■

►**Q-1.8.** This is very elegant. Seeing this proof, it seems that you don't use much tools.

I only use one: the dominated convergence theorem. And the strong law of large numbers can be proved! Now let us turn to another very surprising result, the central limit theorem.

►**Q-1.9.** Surprising?

Yes. As you will see below, there is no assumption on the law of the random variables X_i, it can be discrete-valued or continuous-valued random variables. Still, if you consider the empirical mean $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$ of these iid random variables, and if you conveniently recenter and renormalize, the resulting random variable will always converge in law to the same normal distribution. This is why I find it very surprising.

Theorem 1.14 (The central limit theorem). Assume that $(X_i)_{i \in \mathbb{N}}$ are iid random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[X_1^2] < \infty$. Then, setting $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1)$,

$$Z_n = \frac{\bar{X}_n - \mathbb{E}[X_1]}{\sqrt{\sigma^2/n}} \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} \mathcal{N}(0,1) \,,$$

Or equivalently,

$$\sqrt{n}\,(\bar{X}_n - \mathbb{E}[X_1]) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} \mathcal{N}(0, \sigma^2) \,.$$

► **Notation:** In what follows, CLT stands for the Central Limit Theorem.

PROOF. In this proof, for ease of notation, we consider a random variable X such that $X \overset{\mathcal{L}}{=} X_1$. Replacing if necessary X_i by $(X_i - \mathbb{E}[X])/\sigma$, we can assume that $\mathbb{E}[X] = 0$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] = 1$. In such a case, we only need to prove that

$$Z_N := \sum_{i=1}^N \frac{X_i}{\sqrt{N}} \Rightarrow \mathcal{N}(0,1) \,.$$

To this aim, we will show that the characteristic function of Z_N, $u \mapsto \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uZ_N}\right]$ tends to the one of $\mathcal{N}(0,1)$ that is $u \mapsto \mathrm{e}^{-u^2/2}$. Writing $\varphi(v) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}vX}\right]$, we have

$$\varphi(0) = 1 \,, \quad \varphi'(0) = \mathrm{i}\,\mathbb{E}\left[X\mathrm{e}^{\mathrm{i}vX}\right]\Big|_{v=0} = \mathrm{i}\mathbb{E}[X] = 0 \,, \quad \varphi''(0) = -\,\mathbb{E}\left[X^2\mathrm{e}^{\mathrm{i}vX}\right]\Big|_{v=0} = -\mathbb{E}[X^2] = -1 \,.$$

Then, a second-order Taylor expansion of φ yields

$$\varphi(v) = \varphi(0) + v\varphi'(0) + \frac{v^2}{2}\varphi''(0) + o(v^2) = 1 - \frac{v^2}{2} + o(v^2) \,.$$

Using first that (X_i) are iid and then the above Taylor expansion, we get

$$\mathbb{E}\left[e^{iuZ_N}\right] = \mathbb{E}\left[e^{iu\Sigma_{i=1}^N X_i/\sqrt{N}}\right] = \left\{\mathbb{E}\left[e^{iuX_1/\sqrt{N}}\right]\right\}^N = \left[\varphi(u/\sqrt{N})\right]^N = \left(1 - \frac{u^2}{2N} + o\left(\frac{1}{N}\right)\right)^N \to e^{-u^2/2} = \mathbb{E}[e^{iuZ}],$$

where $Z \sim \mathcal{N}(0,1)$. This completes the proof. ■

To be more familiar with the law of large numbers, the central limit theorem and the Slutsky theorem, consider the following example.

Example 1.15. Assume that $(X_i)_{i\in\mathbb{N}}$ iid with $\mathbb{E}[X_1^2] < \infty$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1) \in (0,\infty)$. Set

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 .$$

Then, we will show that

$$\tilde{Z}_n = \frac{\bar{X}_n - \mathbb{E}[X_1]}{\sqrt{\hat{\sigma}_n^2/n}} \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} \mathcal{N}(0,1) .$$

First note that by the strong law of large numbers, $\mathbb{P}$-a.s.,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i^2 = \mathbb{E}[X_1^2], \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i = \mathbb{E}[X_1] .$$

Hence, $\hat{\sigma}_n^2 \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} \sigma^2$, which in turn implies that $U_n = \sqrt{\frac{\sigma^2}{\hat{\sigma}_n^2}} \overset{\mathbb{P}-prob}{\longrightarrow} 1$. Moreover, according to the central limit theorem,

$$Z_n := \frac{\bar{X}_n - \mathbb{E}[X_1]}{\sqrt{\sigma^2/n}} \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} Z, \quad \text{where} \quad Z \sim \mathcal{N}(0,1) .$$

Applying the Slutsky theorem to the continuous function $(u,z) \mapsto uz$, we finally get:

$$\tilde{Z}_n = U_n Z_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} 1 \times Z = Z ,$$

and the proof is completed.

1.3 Gaussian vectors

In what follows the notation $X \sim \mathcal{N}(\mu,\sigma^2)$ means that X follows the normal distribution with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$. It is equivalent to the fact that the random variable X has the density

$$x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} ,$$

which in turn is equivalent to: for all $u \in \mathbb{R}$,

$$\mathbb{E}[e^{iuX}] = e^{-u^2\sigma^2/2 + iu\mu} .$$

If $\mu = 0$ and $\sigma = 1$, we then say that X follows a standard normal distribution.

Definition 1.16 (Gaussian vector). We say that X *is a d-dimensional Gaussian vector* if any of the two following equivalent conditions is satisfied:

(a) For any $u \in \mathbb{R}^d$, $u^T X$ follows a normal distribution.

(b) There exist a $d \times p$ matrix A and a vector $\mu \in \mathbb{R}^d$ s. t.

$$X = \mu + A \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix} ,$$

where $(\varepsilon_i)_{1 \leqslant i \leqslant p}$ are iid with $\varepsilon_1 \sim \mathcal{N}(0,1)$.

► **Notation:** In this case, we write $X \sim \mathcal{N}(\mu, \Gamma)$ where $\mu = \mathbb{E}[X]$ and

$$\begin{aligned} \Gamma = \mathbb{V}\mathrm{ar}(X) &= \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T \\ &= \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right] . \end{aligned}$$

►**Q- 1.10.** One thing puzzles me. The notation that you use for the distribution of Gaussian random vectors is $\mathcal{N}(\mu, \Gamma)$. It is the same as for the normal distribution ?

Yes, of course! I use the same notation because it is coherent. A Gaussian vector in dimension 1 is just a random variable with normal distribution.

Now, let us move on to a fundamental property with Gaussian vectors: they remain Gaussian through any linear mapping. To be specific, let X be a d-dimensional Gaussian random vector, $X \sim \mathcal{N}(\mu, \Gamma)$. Then, for any $B \in \mathbb{R}^{r \times d}$, BX is a Gaussian random vector and

$$BX \sim \mathcal{N}(B\mu, B\Gamma B^T) . \tag{1.7}$$

Indeed according to (b) in Definition 1.16, there exist a $d \times p$ matrix A and a vector $\mu \in \mathbb{R}^d$ s. t.

$$X = \mu + A \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix} ,$$

where $(\varepsilon_i)_{1 \leqslant i \leqslant p}$ are iid with $\varepsilon_1 \sim \mathcal{N}(0,1)$. Hence,

$$BX = B\mu + BA \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix} .$$

Therefore, BX satisfies (b) in Definition 1.16 and we conclude that BX is a Gaussian vector so that $BX \sim \mathcal{N}(\tilde{\mu}, \tilde{\Gamma})$ where

$$\begin{aligned} \tilde{\mu} &= \mathbb{E}[BX] = B\mathbb{E}[X] = B\mu \\ \tilde{\Gamma} &= \mathbb{V}\mathrm{ar}(BX) = B\mathbb{V}\mathrm{ar}(X)B^T = B\Gamma B^T \end{aligned}$$

This concludes the proof of (1.7).

In practice, if we want to check that a given random vector X is a Gaussian vector, we can either show that it is obtained from another Gaussian vector through a linear mapping as we have just seen here, or we can also show that the characteristic function is the one of a Gaussian vector or calculate the density and check that this density is the one of a Gaussian vector. To recognize the characteristic function and density of a Gaussian vector, recall that if $X \sim \mathcal{N}(\mu, \Gamma)$, then

- for any $u \in \mathbb{R}^d$,

$$\mathbb{E}[e^{iu^T X}] = e^{iu^T \mu - \frac{u^T \Gamma u}{2}} ,$$

- provided that Γ is definite, X has the density

$$x \mapsto f_{\mu,\Gamma}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Gamma}} \mathrm{e}^{-\frac{(X-\mu)^T \Gamma^{-1}(X-\mu)}{2}} ,$$

  wrt the Lebesgue measure on $\mathbb{R}^d$.

At this point, there is a classical confusion on which we should draw the attention of the reader.

A random vector such that *each component follows a normal distribution* is *not necessarily* a Gaussian vector. To see this, consider the following example.

Example 1.17. Draw independently $X \sim \mathcal{N}(0,1)$ and $Z \sim \mathcal{B}(1/2)$. If $Z = 1$, set $Y = X$. Otherwise, set $Y = -X$. We will show that the random vector $\begin{pmatrix} X \\ Y \end{pmatrix}$ is not a Gaussian random vector whereas each of its components follows a normal distribution.

Indeed, note that for any measurable set A,

$$\begin{aligned} \mathbb{P}(Y \in A) &= \mathbb{P}(Y \in A,\, Z = 1) + \mathbb{P}(Y \in A,\, Z = 0) \\ &= \mathbb{P}(X \in A,\, Z = 1) + \mathbb{P}(-X \in A,\, Z = 0) \\ &= \mathbb{P}(X \in A)/2 + \mathbb{P}(-X \in A)/2 = \mathbb{P}(X \in A) . \end{aligned}$$

Therefore, $Y \sim \mathcal{N}(0,1)$, but

$$\mathbb{P}(X - Y = 0) = \mathbb{P}(Z = 1) = 1/2 ,$$

so that $X - Y$ does not follow a normal distribution. We have thus found a particular linear combination $X - Y$ of the random vector $\begin{pmatrix} X \\ Y \end{pmatrix}$ which does not follow a normal distribution and we can conclude that this random vector is not a Gaussian random vector.

Proposition 1.18 (Gaussian vector and independence). Assume that $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is a d-dimensional Gaussian vector, then we have the equivalence

$$(X_1, X_2) \text{ are independent} \iff \quad \mathbb{C}\mathrm{ov}(X_1, X_2) = 0 . \tag{1.8}$$

In the above proposition, the covariance matrix, denoted $\mathbb{C}\mathrm{ov}(X_1, X_2)$, between two random vectors X_1 and X_2 (of possibly different dimensions) is defined by the rectangular matrix

$$\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{E}[X_1 X_2^T] - \mathbb{E}[X_1]\mathbb{E}[X_2]^T = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1]) (X_2 - \mathbb{E}[X_2])^T \right] .$$

And we can note that $\mathbb{C}\mathrm{ov}(X_2, X_1) = \mathbb{C}\mathrm{ov}(X_1, X_2)^T$.

PROOF. [of Proposition 1.18]

$\Longrightarrow$ If (X_1, X_2) are independent, then

$$\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1]) (X_2 - \mathbb{E}[X_2])^T \right] = \underbrace{\mathbb{E}[X_1 - \mathbb{E}[X_1]]}_{0} \mathbb{E}\left[(X_2 - \mathbb{E}[X_2])^T \right] = 0 .$$

$\Longleftarrow$ Assume that $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Gamma)$. Replacing if necessary X by $X - \mu$, we may assume without loss of generality that $\mu = 0$. For $i, j \in \{0, 1\}$, define $\Gamma_{i,j} = \mathbb{C}\mathrm{ov}(X_i, X_j)$. Then,

$$\Gamma := \mathbb{V}\mathrm{ar}(X) = \begin{pmatrix} \Gamma_{1,1} & \Gamma_{1,2} \\ \Gamma_{2,1} & \Gamma_{2,2} \end{pmatrix} = \begin{pmatrix} \Gamma_{1,1} & 0 \\ 0 & \Gamma_{2,2} \end{pmatrix} \quad \text{and hence,} \quad \Gamma^{-1} = \begin{pmatrix} \Gamma_{1,1}^{-1} & 0 \\ 0 & \Gamma_{2,2}^{-1} \end{pmatrix} .$$

Decomposing $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ such that $u^T X = u_1^T X_1 + u_2^T X_2$, the characteristic function of X then writes:

$$\mathbb{E}\left[e^{iu^T X}\right] = \mathbb{E}\left[e^{i\left(u_1^T X_1 + u_2^T X_2\right)}\right] = e^{-\frac{1}{2}u^T \Gamma^{-1} u} = e^{-\frac{1}{2}u_1^T \Gamma_{1,1}^{-1} u_1} e^{-\frac{1}{2}u_2^T \Gamma_{2,2}^{-1} u_2} = \mathbb{E}\left[e^{iu_1^T X_1}\right]\mathbb{E}\left[e^{iu_2^T X_2}\right],$$

where we have used that $X_1 \sim \mathcal{N}(0, \Gamma_{1,1})$ and $X_2 \sim \mathcal{N}(0, \Gamma_{2,2})$. Hence, Q-0.21-(iii) shows that $X_1 \perp\!\!\!\perp X_2$ and the proof is concluded.

■

Theorem 1.19 (The Cochran theorem). Assume that

(a) $X \sim \mathcal{N}(0, I_d)$

(b) P is a $d \times d$ matrix such that $P^2 = P = P^T$

Then, setting

$$Y = PX, \quad \text{and} \quad Z = QX$$

where $Q = I_d - P$, we have

(i) $\begin{cases} Y \sim \mathcal{N}(0, P) \\ Z \sim \mathcal{N}(0, Q) \end{cases}$ and (Y, Z) are **independent**

(ii) $\|Y\|^2 \sim \chi^2_r$ and $\|Z\|^2 \sim \chi^2_{d-r}$ where $r = \operatorname{rank}(P)$.

Remark 1.20. In Theorem 1.19-(a), in the notation $\mathcal{N}(0, I_d)$, by abuse of notation, 0 is the d dimensional null vector and I_d is the $d \times d$ identity matrix.

Remark 1.21. Note that since $P^2 = P = P^T$, we obtain that P is the orthogonal projection matrix on $F = \operatorname{Span}(P_1, \ldots, P_d)$ where (P_i) are the column vectors of P, that is: $P = [P_1, \ldots, P_d]$.

Similarly, $Q = I_d - P$ is an orthogonal projection matrix on $F^\perp$ and it can be readily checked that $Q^2 = Q^T = Q$.

Proof. We start with the proof of (i). Since X is a Gaussian random vector, so is PX by (1.7). Moreover,

$$\mathbb{E}[PX] = P\mathbb{E}[X] = 0 \quad \text{and} \quad \mathbb{V}\text{ar}(PX) = P\mathbb{V}\text{ar}(X)P^T = PP^T = P^2 = P\,.$$

This implies that $Y := PX \sim \mathcal{N}(0, P)$. Similarly, $Z := QX$ is a Gaussian vector such that

$$\mathbb{E}[QX] = Q\mathbb{E}[X] = 0 \quad \text{and} \quad \mathbb{V}\text{ar}(QX) = Q\mathbb{V}\text{ar}(X)Q^T = QQ^T = Q^2 = Q\,.$$

Hence, $Z \sim \mathcal{N}(0, Q)$. To complete the proof of (i), it remains to show that Y and Z are independent. Since

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} P \\ Q \end{pmatrix} X\,, \quad \text{where} \quad X \sim \mathcal{N}(0, I_d)\,,$$

we get that $\begin{pmatrix} Y \\ Z \end{pmatrix}$ is a Gaussian vector. Moreover, $\mathbb{C}\text{ov}(Y, Z) = \mathbb{C}\text{ov}(PX, QX) = P\mathbb{V}\text{ar}(X)Q^T = P(I_d - P)^T = P(I_d - P) = P - P^2 = 0$. Then Proposition 1.18 shows that (Y, Z) are independent and (i) is proved.

We now turn to (ii). We will only show that $\|Y\|^2 \sim \chi^2_r$, where $r = \operatorname{rank}(P)$. The case $\|Z\|^2 \sim \chi^2_{d-r}$ is similar and is omitted for brevity. Recall that P is the orthogonal projection matrix on a space of dimension r. Hence, there exists a square matrix U satisfying $UU^T = U^T U = I_d$ such that $P = U \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U^T$. Setting $\tilde{X} = U^T X$, we have

$$\|Y\|^2 = X^T \underbrace{P^T P}_{P} X = X^T P X = X^T U \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U^T X = \tilde{X}^T \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \tilde{X} = \sum_{i=1}^{r} \tilde{X}_i^2 \sim \chi^2_r\,,$$

since $\tilde{X} \sim \mathcal{N}(0, U^T \underbrace{\mathbb{V}\text{ar}X}_{I_d} U) \sim \mathcal{N}(0, \underbrace{U^T U}_{I_d})$

■

1.4 After studying this chapter...

a) I know the law of large numbers and the central limit theorem and importantly, I remember under which assumptions these theorems hold.

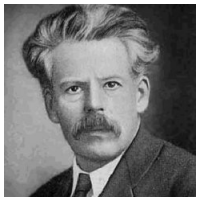
b) I have a grasp on the Slutsky theorem and understand how to use it.

c) I memorize the definition of a Gaussian vector and I can check independence between sub-vectors of Gaussian vectors by computing their covariance matrix.

d) I understand the Cochran theorem and understand the interpretation of P satisfying $P = P^2 = P^T$ as an orthogonal projection matrix.

1.5 Highlights

Evgeny Slutsky (source: Wikipedia)



Evgeny "Eugen" Evgenievich Slutsky (1880 – 1948) was a Russian and Soviet mathematical statistician, economist and political economist.

Slutsky is principally known for work in deriving the relationships embodied in the very well known Slutsky equation which is widely used in microeconomic consumer theory for separating the substitution effect and the income effect of a price change on the total quantity of a good demanded following a price change in that good, or in a related good that may have a cross-price effect on the original good quantity. There are many Slutsky analogs in producer theory.

He is less well known by Western economists than some of his contemporaries, due to his own changing intellectual interests as well as external factors forced upon him after the Bolshevik Revolution in 1917. His seminal paper in Economics, and some argue his last paper in Economics rather than probability theory, was published in 1915 (Sulla teoria del bilancio del consumatore). Paul Samuelson noted that until 1936, he had been entirely unaware of Slutsky's 1915 "masterpiece" due to World War I and the paper's Italian language publication. R. G. D. Allen did the most to propagate Slutsky's work on consumer theory in published papers in 1936 and 1950.

Vincent Barnett argues:

"A good case can be made for the notion that Slutsky is the most famous of all Russian economists, even more well-known [than] N. D. Kondratiev, L. V. Kantorovich, or Mikhail Tugan-Baranovsky. There are eponymous concepts such as the Slutsky equation, the Slutsky diamond, the Slutsky matrix, and the Slutsky-Yule effect, and a journals-literature search conducted on his name for the years 1980-1995 yielded seventy-nine articles directly using some aspect of Slutsky's work... Moreover, many microeconomics textbooks contain prominent mention of Slutsky's contribution to the theory of consumer behavior, most notably the Slutsky equation, christened by John Hicks as the 'Fundamental Equation of Value Theory'. Slutsky's work is thus an integral part of contemporary mainstream economics and econometrics, a claim that cannot really be made by any other Soviet economist, perhaps even by any other Russian economist."

The Slutsky Effect. In the 1920s, Slutsky turned to working on probability theory and stochastic processes, but in 1927 he published his second famous article on economic theory, "The Summation of Random Causes as a Source of Cyclical Processes". This showed that it was possible for apparently cyclic behaviour to emerge as the result of random shocks to the economy if the latter were modelled using a stable stochastic difference equation with certain technical properties. This opened up a new approach to business cycle theory by hypothesising that the interaction of chance events could generate periodicity when none existed initially.

Mathematical statistics work. Slutsky's later work was principally in probability theory and the theory of stochastic processes. He is generally credited for the result known as Slutsky's theorem. In 1928 he was an Invited Speaker of the ICM in Bologna.

Chapter 2

Estimation

2.1 Main ideas

On one hand, the statistician has access to some data (or observations) $(Y_1, \dots, Y_n)$ and on the other hand, he has a family of possible distributions $\mathcal{Q} = (\mathbb{P})_{\mathbb{P}\in\mathcal{Q}}$. The observations have been produced according to some unknown distribution $\mathbb{P}_\star$. When comparing the observations and the family of possible distributions, the statistician aims to answer questions such as:

- How can $\mathbb{P}_\star$ be approximated using a function of the observations? $\rightarrow$ estimator.
- Based on the observations, what could be the range of the probabilities $\mathbb{P}$ which may have produced these observations? $\rightarrow$ confidence intervals/regions.
- Do the observations align with other sources of information that suggest $\mathbb{P}_\star$ is equal to something or belongs to some subset? $\rightarrow$ statistical test.

2.2 Statistical model and characteristics

Definition 2.1. A statistical model is defined by a family of probability measures $\mathcal{Q} := (\mathbb{P})_{\mathbb{P}\in\mathcal{Q}}$ on the space of the observations $(\mathbb{Y}, \mathcal{F})$.

- $\mathbb{Y}$ is the state space (or the space of the observations).
- $\mathcal{F}$ is a σ-field on $\mathbb{Y}$.
- $\mathcal{Q}$ is a family of probability measures.

Typically, if the observations are $\mathsf{Y} = (Y_1, \dots, Y_n)$ and Y_i are p-dimensional random vectors, then $\mathbb{Y}$ may be chosen as $\mathbb{Y} = \mathbb{R}^p \times \dots \times \mathbb{R}^p$ and $\mathcal{F}$ as the associated Borel σ-field, $\mathcal{F} = \sigma(\mathbb{Y})$.

In statistics, the probability $\mathbb{P}_\star$ associated to the observations is not known a priori and one has to get as many information on $\mathbb{P}_\star$ as possible given that $\mathsf{Y} = (Y_1, \dots, Y_n) \in \mathbb{Y}$ have been observed. Note that $\mathbb{P}_\star$ may belong to the family $\mathcal{Q}$ or not.

2.2.1 Parametrisation of the model

- If $\mathcal{Q} = (\mathbb{P}_\theta)_{\theta\in\Theta}$ where Θ is a subset of $\mathbb{R}^d$, then the statistical model is *parametric*.
- If $\mathcal{Q} = (\mathbb{P}_\theta)_{\theta\in\Theta_1\times\Theta_2}$ where Θ_1 is a subset of $\mathbb{R}^k$, then the statistical model is *semi-parametric*.

- Otherwise, the statistical model is *non-parametric*.

In what follows, if we consider a parametric statistical model $(\mathbb{Y}, \mathcal{F}, Q)$, then, we implicitly use the notation $Q = (\mathbb{P}_\theta)_{\theta\in\Theta}$ where Θ is a subset of $\mathbb{R}^d$.

Definition 2.2. A parametric statistical model is identifiable if and only if $\theta \neq \theta'$ implies that $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$.

2.2.2 Dominated parametric model

Definition 2.3. A parametric statistical model $(\mathbb{Y}, \mathcal{F}, Q)$ is dominated by a measure μ if and only if $\mathbb{P}_\theta(\mathrm{d}x) = \ell_\theta(x)\mu(\mathrm{d}x)$ for any $\theta \in \Theta$.

Remark 2.4. The dominating measure μ does not depend on θ. In many examples, μ is the Lebesgue measure or a linear combination of Dirac measures.

In this course, almost all real-valued random variables have a density with respect to the Lebesgue measure. However, for an integer-valued random variable, the dominating measure consists of the counting measure $\sum_{j=0}^{\infty}\delta_j$. The following lemma provides the exact expression of the density for discrete-valued random variables with respect to the counting measure.

Lemma 2.5. Let Y be a random variable which takes values in $\mathbb{N}$. Then, Y follows a distribution having a density f with respect to $\nu = \sum_{j=0}^{\infty}\delta_j$ defined by

$$\begin{cases} f: & \mathbb{N} \to [0,\infty) \\ & i \mapsto f(i) = \mathbb{P}(Y = i) \end{cases}$$

Proof. Denote f such that for all $i \in \mathbb{N}$, $f(i) = \mathbb{P}(Y = i)$. Then, for any integrable function h,

$$\int h(y)f(y)\nu(\mathrm{d}y) = \int h(y)f(y)\left[\sum_{j=0}^{\infty}\delta_j(\mathrm{d}y)\right] = \sum_{j=0}^{\infty}\int h(y)f(y)\delta_j(\mathrm{d}y) = \sum_{j=0}^{\infty}h(j)f(j) = \sum_{j=0}^{\infty}h(j)\mathbb{P}(Y=j) = \mathbb{E}[h(Y)]\,.$$

Thus, $\mathbb{E}[h(Y)] = \int h(y)f(y)\nu(\mathrm{d}y)$. So, f is the density of Y with respect to ν. ■

Definition 2.6. An iid parametric statistical model $(\mathbb{Y}, \mathcal{F}, Q)$, where $Q = (\mathbb{P}_\theta)_{\theta\in\Theta}$, is defined as follows: For all $\theta \in \Theta$, $\mathsf{Y} = (Y_1, \ldots, Y_n)$ is, under $\mathbb{P}_\theta$, an n-tuple of iid random vectors $Y_1, \ldots, Y_n$.

In other words, in more mathematical terms (with heavier notation but possibly less ambiguity), if (Y_i) are random vectors taking values on the measurable space $(\mathbb{Y}_1, \mathcal{F}_1)$, then we define $(\mathbb{Y}, \mathcal{F}) = (\mathbb{Y}_1^n, \mathcal{F}_1^{\otimes n})$. If we consider that the statistical model is iid, it means that for all $\theta \in \Theta$, there exists a probability measure $\mathbb{P}_{\theta,1}$ on $(\mathbb{Y}_1, \mathcal{F}_1)$ such that $\mathbb{P}_\theta = \mathbb{P}_{\theta,1}^{\otimes n}$.

In what follows, we will typically consider an iid dominated parametric statistical model where for all $\theta \in \Theta$, $\mathbb{P}_\theta(\mathrm{d}y) = \prod_{i=1}^n \ell_\theta(y_i)\mu_1(\mathrm{d}y_i)$ with $y = (y_1, \ldots, y_n)$. By abuse of notation, we still denote by ℓ_θ the density of the n-tuple wrt $\mu = \mu_1^{\otimes n}$, i.e., we write

$$\ell_\theta(y_1, \ldots, y_n) = \ell_\theta(y_1)\cdots\ell_\theta(y_n). \tag{2.1}$$

►**Q-2.1.** Sorry but where exactly is your abuse of notation?

In (2.1), it would be more rigorous to write $\ell_{\theta,n}(y_1, \ldots, y_n) = \ell_{\theta,1}(y_1)\cdots\ell_{\theta,1}(y_n)$, but I find it cumbersome and somewhat distracting. In what follows, I will only keep the parameter θ in the subscript when no ambiguity occurs.

2.3 Extracting information from the observations

In all this section, the observations are denoted as $\mathsf{Y} = (Y_1, \dots, Y_n)$.

2.3.1 Statistic

Definition 2.7. A statistic S is a measurable function from $\mathbb{Y}$ to $\mathbb{R}^k$.

A statistic is a "known" function of the observations.

Example 2.8. If $(Y_i)_{1 \leqslant i \leqslant n}$ are iid random variables and $Y_i \sim \mathcal{B}(p)$, where p is unknown, recalling that $\mathsf{Y} = (Y_1, \dots, Y_n)$,

- $\mathsf{Y} \mapsto \sum_{i=1}^n Y_i/n$ is a statistic.
- $\mathsf{Y} \mapsto Y_1$ is a statistic.
- $\mathsf{Y} \mapsto p$ is not a statistic since p is unknown.

Definition 2.9 (Sufficient statistic). A statistic S is sufficient with respect to the parametric family $\mathcal{Q} = (\mathbb{P}_\theta)_{\theta \in \Theta}$ if and only if the law of Y conditionally to S does not depend on θ.

In some sense, the fact that this conditional law does not depend on θ means that when S is known, the knowledge of Y does not provide any additional information about θ.

In other words, $S(\mathsf{Y})$ contains all the information about θ.

Proposition 2.10 (Factorization Theorem). For a dominated parametric statistical model, a statistic S is sufficient if and only if the density of the observations Y, denoted as $y \mapsto \ell_\theta(y)$, admits a decomposition of the form $\ell_\theta(y) = \psi_\theta(S(y))\phi(y)$, where ϕ does not depend on θ.

PROOF. We only provide the proof in the discrete case.

- *Necessary condition:*
$$\mathbb{P}_\theta(\mathsf{Y} = y) = \mathbb{P}_\theta(S(\mathsf{Y}) = S(y))\mathbb{P}(\mathsf{Y} = y | S(\mathsf{Y}) = S(y)) \,,$$
where the last term does not depend on θ. Thus, the decomposition follows.
- *Sufficient condition*:
$$\mathbb{P}_\theta(\mathsf{Y} = y | S(\mathsf{Y}) = s) = \frac{\mathbb{P}_\theta(\mathsf{Y} = y \quad \text{and} \quad S(\mathsf{Y}) = s)}{\mathbb{P}_\theta(S(\mathsf{Y}) = s)} = \frac{\phi(y)}{\sum_{y : S(y) = s} \phi(y)} \,,$$
which does not depend on θ. Thus, S is a sufficient statistic.

■

Definition 2.11 (Exponential model). The parametric statistical model $(\mathbb{Y}, \mathcal{F}, \mathcal{Q})$ is a *k-parameter exponential model* if and only if for any $\theta \in \Theta$,

$$\ell_\theta(y) = \phi_0(y)\phi_1(\theta)e^{\eta(\theta)^T S(y)} \quad \text{where} \quad \eta = \begin{pmatrix} \eta_1 \\ \dots \\ \eta_k \end{pmatrix} . \tag{2.2}$$

In that case,

(i) the (η_i) are called the *natural parameters*.

(ii) S is a k-dimensional *sufficient statistic*.

(iii) Q is called a k-parameter *exponential family*.

Example 2.12 (The binomial family)**.** Suppose that $Y \sim \mathrm{Bin}(n,\theta)$, where n is known but $\theta \in (0,1)$ is not known. We have

$$\ell_\theta(y) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{n}{y} \times \mathrm{e}^{n\log(1-\theta)} \times \mathrm{e}^{y\log\left(\frac{\theta}{1-\theta}\right)}.$$

▷ One-parameter exponential family with $S(y) = y$ and $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$.

Of course, if we can also consider the above family as a *degenerate* two-parameters family by setting

$$S(y) = \begin{pmatrix}1\\ y\end{pmatrix}, \quad \text{and} \quad \eta(\theta) = \begin{pmatrix} n\log(1-\theta)\\ \log\left(\frac{\theta}{1-\theta}\right)\end{pmatrix}.$$

We call it degenerate because the first component of S does not actually depend on the observations. Therefore, it is less informative than the one-parameter formulation, and hence, in practice, it is better to stick with the one-parameter formulation.

►**Q-2.2.** In the above example, you consider only one observation?

Exactly, if we consider an iid model with k observations, then,

$$\ell_\theta(y_1,\dots,y_k) = \prod_{i=1}^k \ell_\theta(y_i) = \prod_{i=1}^k \left(\binom{n}{y_i}\theta^y(1-\theta)^{n-y_i}\right) = \left(\prod_{i=1}^k \binom{n}{y_i}\right) \times \mathrm{e}^{nk\log(1-\theta)} \times \mathrm{e}^{(\sum_{i=1}^k y_i)\log\left(\frac{\theta}{1-\theta}\right)}$$

It is thus a one-parameter exponential family with sufficient statistic $S(y_1,\dots,y_k) = \sum_{i=1}^k y_i$ and natural parameter $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$.

Example 2.13 (The Gaussian family)**.** Suppose that $Y \sim \mathcal{N}(\mu,\sigma^2)$ where $\theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_*^+$. Then,

$$\ell_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(y-\mu)^2/(2\sigma^2)} = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\mu^2/(2\sigma^2)} \times \mathrm{e}^{-\frac{y^2}{2\sigma^2}+\frac{y\mu}{\sigma^2}}.$$

▷ Two-parameters exponential family with $S(y) = \begin{pmatrix}-y^2/2\\ y\end{pmatrix}$ and $\eta(\theta) = \begin{pmatrix}1/\sigma^2\\ \mu/\sigma^2\end{pmatrix}$.

►**Q-2.3.** May I ask the same question as for the other example? What would it be if we consider an iid model with k observations?

Similarly to the previous example, we still keep a two-parameters exponential family but with sufficient statistics $S(y) = \begin{pmatrix}\sum_{i=1}^k -y_i^2/2\\ \sum_{i=1}^k y_i\end{pmatrix}$ and natural parameters $\eta(\theta) = \begin{pmatrix}1/\sigma^2\\ \mu/\sigma^2\end{pmatrix}$.

2.3.2 Likelihood, Score function and Information matrix

Definition 2.14 (Log-Likelihood). If $(\mathbb{Y},\mathcal{F},Q)$ is a dominated parametric model, then $\ell_\theta(\mathsf{Y})$, resp. $\log\ell_\theta(\mathsf{Y})$, is the likelihood, resp. the log-likelihood, associated to θ and to the observation Y.

►**Q-2.4.** What is the difference between the likelihood and a density function?

A likelihood $\ell_\theta(Y)$ is precisely the value of the density ℓ_θ evaluated at the random element Y. Hence, a likelihood is actually a random variable.

Example 2.15. The likelihood for an integer-valued observation writes as

$$\ell_\theta(Y) = \prod_{i=0}^{\infty} [\mathbb{P}_\theta(Y=i)]^{\mathbf{1}_i(Y)} \ .$$

where $\mathbf{1}_i(Y)$ is the indicator function that takes the value 1 when $Y=i$ and 0 otherwise.

Definition 2.16 (Score function). If $(\mathbb{Y}, \mathcal{F}, \mathbb{P}_\theta)$ is a dominated parametric model, then $\xi_\theta(Y) = \frac{\partial \log \ell_\theta(Y)}{\partial \theta}$ (when it is well-defined) is the score function.

►**Q-2.5.** Can you explain the notation $\frac{\partial \log \ell_\theta(Y)}{\partial \theta}$?

Here, the notation $\frac{\partial \log \ell_\theta(Y)}{\partial \theta}$ stands for the gradient (with respect to θ) of the log-likelihood. A score function is thus *a d-dimensional random vector*. Actually, in what follows, we use the formal notation $\frac{\partial}{\partial \theta}$ and $\frac{\partial}{\partial \theta^T}$ as follows

$$\frac{\partial}{\partial \theta} = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_d} \end{pmatrix} \quad \text{and} \quad \frac{\partial}{\partial \theta^T} = \left(\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_d} \right) = \left(\frac{\partial}{\partial \theta} \right)^T ,$$

and to anticipate, we will use also the notation:

$$\frac{\partial^2}{\partial \theta \partial \theta^T} = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_d} \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial \theta_d \partial \theta_1} & \cdots & \frac{\partial^2}{\partial \theta_d^2} \end{pmatrix} .$$

Now, we move on to the main properties of the score function. Introduce the following assumption.

(A1) For μ-almost all y, $\theta \mapsto \ell_\theta(y)$ is continuously differentiable on Θ and

$$\int \sup_{\theta \in \Theta} \left\| \frac{\partial \ell_\theta(y)}{\partial \theta} \right\| \mu(\mathrm{d}y) < \infty. \tag{2.3}$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^d$.

Proposition 2.17. Assume (A1). Then, for any $\theta \in \Theta$, the score function $\xi_\theta(Y)$ is centered under $\mathbb{P}_\theta$.

PROOF. For any $\theta \in \Theta$,

$$\mathbb{E}_\theta[\xi_\theta(Y)] = \int \frac{\partial \log \ell_\theta(y)}{\partial \theta} \ell_\theta(y) \mu(\mathrm{d}y) = \int \frac{\partial \ell_\theta(y)}{\partial \theta} \mu(\mathrm{d}y) = \frac{\partial}{\partial \theta} \underbrace{\int \ell_\theta(y) \mu(\mathrm{d}y)}_{1} = 0 \, ,$$

where the condition (2.3) allows to interchange integral and derivation wrt θ. ■

Before defining Fisher information matrix, we introduce two assumptions:

(A2) For μ-almost all y, $\theta \mapsto \ell_\theta(y)$ is differentiable on Θ and for any $\theta \in \Theta$,

$$\mathbb{E}_\theta\left[\left\|\frac{\partial \log \ell_\theta(Y)}{\partial \theta}\right\|^2\right] < \infty\,.$$

(A3) For μ-almost all y, $\theta \mapsto \ell_\theta(y)$ is continuously twice-differentiable on Θ and for any $\theta \in \Theta$,

$$\int \sup_{\theta\in\Theta}\left\|\frac{\partial^2 \ell_\theta(y)}{\partial\theta\partial\theta^T}\right\| \mu(\mathrm{d}y) < \infty\,, \tag{2.4}$$

where by abuse of notation, we again denote by $\|\cdot\|$ any norm on $\mathbb{R}^{d\times d}$.

Definition 2.18 (Fisher Information Matrix). Let $(\mathbb{Y}, \mathcal{F}, Q)$ be a dominated parametric model.

(i) Assume (A1), (A2). Then the Fisher information matrix is defined by

$$I_F^\theta(Y) := \mathbb{V}\mathrm{ar}_\theta(\xi_\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \log \ell_\theta(Y)}{\partial\theta}\right)\left(\frac{\partial \log \ell_\theta(Y)}{\partial\theta}\right)^T\right]\,, \tag{2.5}$$

where by definition $\mathbb{V}\mathrm{ar}_\theta(U) := \mathbb{E}_\theta[UU^T] - \mathbb{E}_\theta[U]\mathbb{E}_\theta[U]^T$ is the covariance matrix of the random vector U.

(ii) Assume (A1), (A2) and (A3). Then,

$$I_F^\theta(Y) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log \ell_\theta(Y)}{\partial\theta\partial\theta^T}\right]\,. \tag{2.6}$$

Proof. The second equality in (2.5) comes from the fact that ξ_θ is centered under $\mathbb{P}_\theta$. To prove (2.6), write

$$-\frac{\partial^2 \log \ell_\theta(Y)}{\partial\theta\partial\theta^T} = \frac{1}{\ell_\theta(Y)}\frac{\partial^2 \ell_\theta(Y)}{\partial\theta\partial\theta^T} + \left(\frac{\partial \log \ell_\theta(Y)}{\partial\theta}\right)\left(\frac{\partial \log \ell_\theta(Y)}{\partial\theta}\right)^T\,,$$

and take the expectation under the parameter θ. The proof is completed noting that

$$\mathbb{E}_\theta\left[\frac{1}{\ell_\theta(Y)}\frac{\partial^2 \ell_\theta(Y)}{\partial\theta\partial\theta^T}\right] = \int \frac{\partial^2}{\partial\theta\partial\theta^T}\ell_\theta(y)\nu(\mathrm{d}y) = \frac{\partial^2}{\partial\theta\partial\theta^T}\underbrace{\int \ell_\theta(y)\nu(\mathrm{d}y)}_{1} = 0\,,$$

where (2.4) allows to interchange integral and second derivatives wrt θ. ∎

Remark 2.19. The previous definition involves three different expressions (in (2.5) and (2.6)) of the Fisher Information matrix. In some statistical models, the second derivatives of the log-likelihood may not be defined, whereas the first derivatives exist. In such cases, the Fisher Information matrix is defined as the covariance matrix of the score function (i.e., we use only (2.5) in the previous definition).

Lemma 2.20. For an i.i.d dominated parametric statistical model and under the assumptions of Definition 2.18, $I_F^\theta(Y_1,\dots,Y_n) = nI_F^\theta(Y_1)$ for any $\theta \in \Theta$.

PROOF. We have for an i.i.d. model, $\ell_\theta(Y_1,\dots,Y_n) = \prod_{i=1}^n \ell_\theta(Y_i)$ and thus,

$$I_F^\theta(Y_1,\dots,Y_n) = \mathbb{V}\mathrm{ar}_\theta\left[\frac{\partial \log \ell_\theta(Y_1,\dots,Y_n)}{\partial\theta}\right] = \mathbb{V}\mathrm{ar}_\theta\left[\sum_{i=1}^n \frac{\partial \log \ell_\theta(Y_i)}{\partial\theta}\right] = \sum_{i=1}^n \mathbb{V}\mathrm{ar}_\theta\left[\frac{\partial \log \ell_\theta(Y_i)}{\partial\theta}\right] = nI_F^\theta(Y_1)\,.$$

■

Before going further, let us recall some expressions of expectation and variance using conditional expectations or conditional variances.

Lemma 2.21. If U and V are random vectors on the same probability space, then

- if $\mathbb{E}[\|U\|] < \infty$, then $\mathbb{E}[U] = \mathbb{E}[\mathbb{E}[U|V]]$,
- if $\mathbb{E}[\|U\|^2] < \infty$, then $\mathbb{V}\mathrm{ar}(U) = \mathbb{V}\mathrm{ar}(\mathbb{E}[U|V]) + \mathbb{E}[\mathbb{V}\mathrm{ar}(U|V)]$.

We now need to use the observations to select a statistical model which may have produced the observations. We thus consider a function of the observations which approximates a parameter of the model. This is the notion of estimator. After defining the estimator, we will introduce some characteristics of the estimator which quantifies how the estimator is close to the parameter.

2.4 Estimator

Recall the notation $\mathsf{Y} = (Y_1,\dots,Y_n)$.

Definition 2.22 (Estimator). In a parametric statistical model, an estimator $\mathsf{Y} \mapsto \delta(\mathsf{Y})$ of $g(\theta)$ is a statistic that is "supposed" to approximate $g(\theta)$.

Example 2.23. If (Y_i) are i.i.d and $Y_1 \sim \mathcal{N}(\mu,\sigma^2)$, then

- $\mathsf{Y} = (Y_1,\dots,Y_n) \mapsto \frac{\sum_{i=1}^n Y_i}{n} := \bar{Y}_n$ is an estimator of $\mathbb{E}[Y_1] = \mu$.
- If μ is known, then $\mathsf{Y} = (Y_1,\dots,Y_n) \mapsto \frac{\sum_{i=1}^n (Y_i-\mu)^2}{n}$ is an estimator of $\mathbb{V}\mathrm{ar}(Y_1) = \sigma^2$.

Note that even when (Y_i) do not follow the normal distribution, these two statistics are estimators of the mean and the variance of Y_1.

An estimator is thus a "known" function of the observations (a function which does not depend on θ) that approximates either the parameter or a function of the parameter. Most of the time, $g(\theta) = \theta$ but it may happen that the statistician wants to approximate just one component of θ or more generally a function of θ that we write $g(\theta)$.

Of course, this notion needs to be refined by some criterion which states that the approximation is close enough to the quantity to be approximated. We now introduce the bias, which will quantify in the "mean" sense the proximity of the estimator to the function of interest $g(\theta)$.

Definition 2.24 (Biased and unbiased estimator).

- An estimator $\delta(\mathsf{Y})$ of $g(\theta)$ is *unbiased* if and only if for all $\theta \in \Theta$, $\mathbb{E}_\theta[\delta(\mathsf{Y})] = g(\theta)$.
- More generally, the *bias* is defined for any possibly biased estimator by

$$b_\theta(\delta, g) = \mathbb{E}_\theta[\delta(\mathsf{Y})] - g(\theta)\,.$$

Example 2.25. If (Y_i) are i.i.d and $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. Assume that μ is not known then it can be shown that

$$\mathsf{Y} \mapsto \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}{n-1}$$

is an unbiased estimator of σ^2.

More generally, for a possibly biased estimator, the mean-squared error (MSE) is more suitable for measuring how the estimator is close to the parameter:

$$\mathrm{MSE}^2 = \mathbb{E}_\theta[(\delta(\mathsf{Y}) - g(\theta))^2] = b_\theta(\delta, g)^2 + \mathbb{V}\mathrm{ar}_\theta(\delta(\mathsf{Y}))\ .$$

To diminish the MSE, there is therefore a compromise between the bias and the variance of an estimator.

2.4.1 Improving estimation with sufficient statistics

In this short section, we are provided with an unbiased estimator $\delta(\mathsf{Y})$ and by considering some transformation of this estimator using sufficient statistics, we can find an unbiased estimator with a smaller variance.

Theorem 2.26 (The Rao-Blackwell theorem). If $\delta(\mathsf{Y})$ is an unbiased estimator of $g(\theta)$ and if S a sufficient statistic, then $\mathbb{E}[\delta(\mathsf{Y})|S]$ is an unbiased estimator of $g(\theta)$ with smaller variance than the one of $\delta(\mathsf{Y})$.

Proof. First, since S is sufficient, $\mathbb{E}_\theta[\delta(\mathsf{Y})|S]$ does not depend on θ, it is thus an estimator and we may thus drop the dependence on θ, i.e. we only write $\mathbb{E}[\delta(\mathsf{Y})|S]$. Moreover, the estimator is unbiased since

$$\mathbb{E}_\theta[\mathbb{E}_\theta[\delta(\mathsf{Y}))|S]] = \mathbb{E}_\theta[\delta(\mathsf{Y})] = g(\theta)\ ,$$

where we have used Lemma 2.21. The proof is concluded by noting that

$$\mathbb{V}\mathrm{ar}_\theta[\mathbb{E}[\delta(\mathsf{Y}))|S]] \leqslant \mathbb{V}\mathrm{ar}_\theta[\mathbb{E}[\delta(\mathsf{Y}))|S]] + \mathbb{E}_\theta[\mathbb{V}\mathrm{ar}_\theta[\delta(\mathsf{Y}))|S]] = \mathbb{V}\mathrm{ar}_\theta[\delta(\mathsf{Y})]\ ,$$

where we used again Lemma 2.21. ■

2.4.2 The Cramér-Rao Bound

Definition 2.27 (MVUB). An estimator $\delta(\mathsf{Y})$ of $g(\theta) \in \mathbb{R}^k$ is MVUB (Minimum Variance UnBiased estimator) if and only if

- $\delta(\mathsf{Y})$ is unbiased,
- for any other unbiased estimator $\tilde{\delta}(\mathsf{Y})$, we have

$$\mathbb{V}\mathrm{ar}_\theta(\delta(\mathsf{Y})) \leqslant \mathbb{V}\mathrm{ar}_\theta(\tilde{\delta}(\mathsf{Y})) \quad \forall \theta \in \Theta\ .$$

In this definition, for two $k \times k$ symmetric matrices A, B with real entries, $A \leqslant B$ means that for all $u \in \mathbb{R}^k$, $u^T A u \leqslant u^T B u$.

The uniqueness of such an estimator is given by the following theorem.

Theorem 2.28. If $\delta(\mathsf{Y}), \delta'(\mathsf{Y})$ are two MVUB estimators, then for any $\theta \in \Theta$, $\delta(\mathsf{Y}) = \delta'(\mathsf{Y})$, $\mathbb{P}_\theta$-a.s.

PROOF. If δ and δ' are unbiased estimators of minimal variance, then $\mathbb{V}\mathrm{ar}_\theta(\delta(Y)) = \mathbb{V}\mathrm{ar}_\theta(\delta'(Y))$. Now, consider $\delta'' = (\delta+\delta')/2$. We have by definition of δ and δ', for any $\theta \in \Theta$,

$$\mathbb{V}\mathrm{ar}_\theta(\delta(Y)) \leqslant \mathbb{V}\mathrm{ar}_\theta(\delta''(Y)) = \frac{1}{4}(\mathbb{V}\mathrm{ar}_\theta(\delta(Y)) + \mathbb{V}\mathrm{ar}_\theta(\delta'(Y)) + 2\mathbb{C}\mathrm{ov}_\theta(\delta(Y),\delta'(Y))$$
$$= \frac{1}{2}\left(\mathbb{V}\mathrm{ar}_\theta(\delta(Y)) + \mathbb{C}\mathrm{ov}_\theta(\delta(Y),\delta'(Y))\right),$$

which is equivalent to $0 \leqslant \mathbb{V}\mathrm{ar}_\theta(\delta(Y) - \delta'(Y)) \leqslant 0$. Using that these two estimators are unbiased, we finally have $\delta(Y) = \delta'(Y)$, $\mathbb{P}_\theta$-a.s. ■

In the previous section, we reduced the variance of an unbiased estimator by utilizing sufficient statistics. However, for any dominated parametric model, sufficient statistics may not always exist. Even in cases where they do, the computation of $\mathbb{E}[\delta(Y)|S]$ may not be explicit. In this section, we will demonstrate the existence of an explicit lower bound for the variance of an unbiased estimator.

Theorem 2.29 (The Cramér-Rao Bound). Assume that (A1), (A2) hold and $I_F^\theta(Y)$ is invertible. For any unbiased "regular" estimator $\delta(Y)$ of $g(\theta)$, we have

$$\mathbb{V}\mathrm{ar}_\theta[\delta(Y)] \geqslant \frac{\partial g(\theta)}{\partial \theta^T}\left[I_F^\theta(Y)\right]^{-1}\frac{\partial g^T(\theta)}{\partial \theta}, \tag{2.7}$$

where we recall that Y is the generic notation of the observations, ie $Y = (Y_1,\dots,Y_n)$.

As a byproduct, any unbiased estimator of $g(\theta)$ cannot be too close to $g(\theta)$ due to the Cramér-Rao bound.

PROOF. We first rewrite the covariance matrix between $\delta(Y)$ and the score function $\xi_\theta(Y) = \frac{\partial \log \ell_\theta(Y)}{\partial \theta}$, using that the score is centered under $\mathbb{P}_\theta$ and that δ is an unbiased estimator of $g(\theta)$,

$$\mathbb{C}\mathrm{ov}_\theta\left(\delta(Y),\xi_\theta(Y)\right) = \mathbb{E}_\theta\left[\delta(Y)\frac{\partial \log \ell_\theta(Y)}{\partial \theta^T}\right] - 0 = \int \delta(y)\frac{\partial \ell_\theta(y)}{\partial \theta^T}\mu(\mathrm{d}y) = \frac{\partial}{\partial \theta^T}\underbrace{\int \delta(y)\ell_\theta(y)\mu(\mathrm{d}y)}_{g(\theta)} = \frac{\partial g(\theta)}{\partial \theta^T}. \tag{2.8}$$

Hence, denoting by M_θ, the rectangular matrix with deterministic entries, $M_\theta := \frac{\partial g(\theta)}{\partial \theta^T}\left[I_F^\theta(Y)\right]^{-1}$, we have

$$0 \leqslant \mathbb{V}\mathrm{ar}_\theta\left[\delta(Y) - M_\theta\xi_\theta(Y)\right] = \mathbb{V}\mathrm{ar}_\theta[\delta(Y)] + M_\theta\underbrace{\mathbb{V}\mathrm{ar}_\theta[\xi_\theta(Y)]}_{I_F^\theta(Y)}M_\theta^T - 2\mathbb{C}\mathrm{ov}_\theta(\delta(Y),M_\theta\xi_\theta(Y))$$
$$= \mathbb{V}\mathrm{ar}_\theta[\delta(Y)] + \frac{\partial g(\theta)}{\partial \theta^T}\left[I_F^\theta(Y)\right]^{-1}\frac{\partial g(\theta)^T}{\partial \theta} - 2\mathbb{C}\mathrm{ov}_\theta\left(\delta(Y),\xi_\theta(Y)\right)M_\theta^T$$
$$= \mathbb{V}\mathrm{ar}_\theta[\delta(Y)] - \frac{\partial g(\theta)}{\partial \theta^T}\left[I_F^\theta(Y)\right]^{-1}\frac{\partial g(\theta)^T}{\partial \theta},$$

where the last equality follows from (2.8) and the definition of M_θ. The proof is concluded. ■

►**Q-2.6.** Amazing! I only have a question: in the statement of the Cramér-Rao bound, you said that $\delta(Y)$ is an unbiased "regular" estimator of $g(\theta)$. If I am not mistaken, I have not seen any definition of "regular" estimators.

You have a keen eye! Actually, I wanted to keep the Cramér-Rao theorem simple, without being too technical... However, a "regular" estimator is such that it has a second-order moment under $\mathbb{P}_\theta$ so that its covariance matrix under $\mathbb{P}_\theta$ is well-defined and such that we can interchange integration wrt μ and derivatives wrt θ in the penultimate equality in (2.8). For this property to hold, we can, for example, assume that

$$\int \|\delta(y)\| \sup_{\theta\in\Theta}\left\|\frac{\partial \ell_\theta(y)}{\partial \theta}\right\|\mu(\mathrm{d}y) < \infty.$$

The Cramér-Rao bound leads to the definition of *efficiency*. An unbiased estimator which satisfies equality in (2.9) is said to be **efficient**. As an immediate consequence of the Cramér-Rao bound, an efficient estimator is MVUB.

►**Q-2.7.** So, efficiency closes the case of estimation? All we have to do is to find efficient estimators?"

Easier said than done... Although finding efficient estimators would be an ultimate goal, there is no result that guarantees their universal existence. Moreover, upon inspection of the proof, an efficient estimator $\delta(Y)$ would satisfy: $\mathbb{V}\mathrm{ar}_\theta[\delta(Y) - M_\theta\xi_\theta(Y)] = 0$, indicating that $\delta(Y) = M_\theta\xi_\theta(Y) + g(\theta)$, $\mathbb{P}_\theta$-a.s. Unfortunately, there is no reason to believe that this latter quantity solely depends on the observations Y and not on the parameter θ (which is essential for estimators).

Nonetheless, you must not be too disappointed. At the very least, it provides a lower bound for the variance of any unbiased estimator, which could prove to be useful. Furthermore, as we will see later, we will be able to find some estimators that will be "asymptotically efficient," in a sense to be defined.

A particular case for the Cramér-Rao bound where $g(\theta) = \theta$ implies that the bound is equal to the inverse of the Fisher Information matrix.

Corollary 2.30. For a regular model, if δ is an unbiased regular estimator of $g(\theta) = \theta$ then

$$\mathbb{V}\mathrm{ar}_\theta[\delta(Y_1,\dots,Y_n)] \geqslant \left[I_F^\theta(Y_1,\dots,Y_n)\right]^{-1}. \tag{2.9}$$

Hence, if in addition the model is i.i.d.,

$$\mathbb{V}\mathrm{ar}_\theta[\delta(Y_1,\dots,Y_n)] \geqslant \left[nI_F^\theta(Y_1)\right]^{-1}.$$

Remark 2.31. (About the notation $\frac{\partial g(\theta)}{\partial \theta^T}$ and $\frac{\partial g^T(\theta)}{\partial \theta}$). We have

- if $g(\theta) = \begin{pmatrix} g_1(\theta) \\ \vdots \\ g_p(\theta) \end{pmatrix}$ then $g^T(\theta) := (g_1(\theta),\dots,g_p(\theta))$.
- if $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$ then $\frac{\partial}{\partial\theta} := \begin{pmatrix} \frac{\partial}{\partial\theta_1} \\ \vdots \\ \frac{\partial}{\partial\theta_d} \end{pmatrix}$ and $\frac{\partial}{\partial\theta^T} := (\frac{\partial}{\partial\theta_1},\dots,\frac{\partial}{\partial\theta_d})$ and thus

$$\frac{\partial g^T(\theta)}{\partial\theta} := \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial\theta_1} & \cdots & \frac{\partial g_p(\theta)}{\partial\theta_1} \\ \vdots & & \vdots \\ \frac{\partial g_1(\theta)}{\partial\theta_d} & \cdots & \frac{\partial g_p(\theta)}{\partial\theta_d} \end{pmatrix} \quad \text{and} \quad \frac{\partial g(\theta)}{\partial\theta^T} := \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial\theta_1} & \cdots & \frac{\partial g_1(\theta)}{\partial\theta_d} \\ \vdots & & \vdots \\ \frac{\partial g_p(\theta)}{\partial\theta_1} & \cdots & \frac{\partial g_p(\theta)}{\partial\theta_d} \end{pmatrix}.$$

2.5 Methods of estimation

In this section, given a parametric statistical model and the observations $(Y_i)_{1\leqslant i\leqslant n}$, we provide several methods for obtaining an approximation $\hat{\theta}_n$ of $\theta \in \mathbb{R}^d$.

2.5.1 Method of Moments (MOM)

The Method of Moments (MOM) can be split into several steps:

(i) Choose d functions $(T_1,\dots,T_d)$ and set $e_j(\theta)=\mathbb{E}_\theta[T_j(Y_1)]$.

(ii) Solve with respect to $\theta=(\theta_1,\dots,\theta_d)$ the d equations

$$\frac{1}{n}\sum_{i=1}^{n} T_j(Y_i) = e_j(\theta) \quad \text{for } j=1,\dots,d\ ,$$

and call $\hat{\theta}_n$ this approximation.

To fully understand the methods of moments, three ingredients should be gathered. First, the mapping $\psi:(\theta_1,\dots,\theta_d)\mapsto(\mathbb{E}_\theta[T_1(Y_1)],\dots,\mathbb{E}_\theta[T_d(Y_1)])$ should be one-to-one. Second, for $j\in[1:d]$, we replace $\mathbb{E}_\theta[T_j(Y_1)]$ by $n^{-1}\sum_{i=1}^n T_j(Y_i)$ for sufficiently large n (this approximation is asymptotically justified by using the SLLN). Third, we apply ψ^{-1} to get $\hat{\theta}_n$.

►**Q-2.8.** Really? How can you apply ψ^{-1}?

This is exactly Step (ii) in the Method of Moments. Solving the d equations in Step (ii) boils down to applying ψ^{-1} to the approximations $n^{-1}\sum_{i=1}^n T_j(Y_i)$. We now invite you to practice the MOM with two examples.

►**Q-2.9.** Oh, with pleasure... thank you so much for your invitation!

You are welcome. Let us start right away.

Example 2.32 (Exponential distribution)**.** In this first example, we consider a one-dimensional method of moments. If $Y_1\sim\mathbf{exp}(\theta)$, then,

$$\mathbb{E}_\theta[Y_1]=\frac{1}{\theta} \quad \text{and} \quad \mathbb{E}_\theta[Y_1^2]=\frac{2}{\theta^2}\ .$$

Here, we provide two ways for applying the method of moments:

1. Choosing $T(y)=y$,
   - Solve $n^{-1}\sum_{i=1}^n Y_i = 1/\theta$.
   - We get $\hat{\theta}_n=\frac{1}{n^{-1}\sum_{i=1}^n Y_i}$.
2. Choosing $T(y)=y^2$,
   - Solve $n^{-1}\sum_{i=1}^n Y_i^2 = 2/\theta^2$.
   - We get $\hat{\theta}_n=\left(\frac{2}{n^{-1}\sum_{i=1}^n Y_i^2}\right)^{1/2}$.

Example 2.33 (Normal distribution)**.** If $Y_1\sim\mathcal{N}(m,\sigma^2)$ then setting $\theta=(m,\sigma^2)$, we have

$$\mathbb{E}_\theta[Y_1]=m \quad \text{and} \quad \mathbb{E}_\theta[Y_1^2]=m^2+\sigma^2\ .$$

Since θ is two-dimensional, we will consider here a two-dimensional method of moments. Choosing $T_1(y)=y$ and $T_2(y)=y^2$

- we solve in $\theta=(m,\sigma^2)$, the equations

$$\begin{cases} n^{-1}\sum_{i=1}^n Y_i &= m \\ n^{-1}\sum_{i=1}^n Y_i^2 &= m^2+\sigma^2 \end{cases}$$

- we obtain

$$\begin{cases} \hat{m}_n = n^{-1}\sum_{i=1}^n Y_i \\ \hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n Y_i^2 - \hat{m}_n^2 \end{cases}$$

and we set $\hat{\theta}_n=(\hat{m}_n,\hat{\sigma}_n^2)$.

2.5.2 Maximum likelihood estimator

Recall that $\mathsf{Y} = (Y_1, \dots, Y_n)$ are observed.

Definition 2.34 (Maximum likelihood estimator). We say that $\hat{\theta}_n^{\text{ML}}$ is the Maximum Likelihood Estimator (MLE) if the following condition is satisfied:

$$\ell_{\hat{\theta}_n^{\text{ML}}}(Y_1, \dots, Y_n) = \max_{\theta \in \Theta} \ell_\theta(Y_1, \dots, Y_n) \, ,$$

or equivalently,

$$\hat{\theta}_n^{\text{ML}} \in \text{argmax}_{\theta \in \Theta} \ell_\theta(Y_1, \dots, Y_n) \, .$$

As a fundamental particular case, consider iid dominated parametric models. Then,

$$\ell_\theta(Y_1, \dots, Y_n) = \prod_{i=1}^n \ell_\theta(Y_i) \, ,$$

and thus,

$$\hat{\theta}_n^{\text{ML}} \in \text{argmax}_{\theta \in \Theta} \, n^{-1} \sum_{i=1}^n \log \ell_\theta(Y_i) \, .$$

Asymptotic properties

So far, we have considered a fixed number of observations n. To study the asymptotic properties of the Maximum Likelihood Estimator (MLE) as n tends to infinity, we need to consider parametric statistical models of the form $(\mathbb{Y}, \mathcal{F}, Q) = (\mathbb{Y}_1^{\mathbb{N}}, \mathcal{F}_1^{\otimes \mathbb{N}}, (\mathbb{P}_\theta)_{\theta \in \Theta})$. Here, we define an iid dominated parametric statistical model $(\mathbb{Y}, \mathcal{F}, Q)$ by the property that, for any $\theta \in \Theta$, under $\mathbb{P}_\theta$, the random sequence $\mathsf{Y} = (Y_1, \dots, Y_n, \dots)$ is composed of iid random variables $\{Y_i \, : \, i \in \mathbb{N}\}$ with $Y_i \sim \ell_\theta(y_1)\mu_1(\mathrm{d}y_1)$.

We say that an iid parametric model associated to the observations (Y_i) is *well-specified* if there exists a parameter $\theta_\star \in \Theta$ such that the observations (Y_i) are iid according to $\mathbb{P}_{\theta_\star}$. In what follows, we assume that the model is well-specified.

Proposition 2.35 (Strong Consistency of the MLE). Assume (A4). Then,

$$\lim_{n \to \infty} \hat{\theta}_n^{\text{ML}} = \theta_\star \, , \quad \mathbb{P}_{\theta_\star} - a.s.$$

This property shows that the MLE is *asymptotically unbiased.* Indeed, under the assumptions of Proposition 2.35, the set Θ is compact. Hence, $\left\| \hat{\theta}_n^{\text{ML}} \right\| \leqslant \sup_{\theta \in \Theta} \|\theta\|$ and the dominated convergence theorem combined with Proposition 2.35 yields $\lim_n \mathbb{E}_{\theta_\star}[\hat{\theta}_n^{\text{ML}}] = \theta_\star$. Hence,

$$\mathbb{E}_{\theta_\star}[\hat{\theta}_n^{\text{ML}}] \approx \theta_\star \text{ for large } n.$$

Proposition 2.36 (Asymptotic normality of the MLE). Assume (A5). Then,

$$\sqrt{n}(\hat{\theta}_n^{\text{ML}} - \theta_\star) \overset{\mathcal{L}_{\mathbb{P}_{\theta_\star}}}{\rightsquigarrow} \mathcal{N}\left(0, \left[I_F^{\theta_\star}(Y_1)\right]^{-1}\right) \, .$$

This property shows that the MLE is *asymptotically efficient*. Indeed, under mild additional assumptions, the above theorem shows that

$$n\mathbb{E}_{\theta_\star}\left[\left(\hat{\theta}_n^{\text{ML}}-\theta_\star\right)^2\right] \approx \left(I_F^{\theta_\star}(Y_1)\right)^{-1} .$$

Hence, $\mathbb{E}_{\theta_\star}\left[\left(\hat{\theta}_n^{\text{ML}}-\theta_\star\right)^2\right] \approx \left(nI_F^{\theta_\star}(Y_1)\right)^{-1}$ and thus,

$$\mathbb{E}_{\theta_\star}\left[\left(\hat{\theta}_n^{\text{ML}}-\theta_\star\right)^2\right] \approx \left(I_F^{\theta_\star}(Y_1,\ldots,Y_n)\right)^{-1} .$$

This asymptotic efficiency is a strong argument for using MLE but of course, we should be careful: when we say that the MLE is *asymptotically efficient*, it means that n goes to infinity, we don't say that it is efficient for a given finite-sample.

►**Q-2.10.** Ok, for the asymptotic efficiency, but you did not define (A4) or (A5) for the strong consistency or the asymptotic normality properties.

Yes, I chose to include the statements of these assumptions in the appendix of this chapter, along with the complete proofs of Proposition 2.35 and Proposition 2.36. By doing so, I hope that the pace of the reading is not slowed down, and you are not distracted by technicalities, so you can focus more quickly on the asymptotic efficiency of the MLE.

Still, there is one point I would like to draw your attention to. It is the Kullback-Leibler divergence.

►**Q-2.11.** What is it? I've never heard of it before.

Let us recall it.

Kullback-Leibler divergence and links with the MLE.

Definition 2.37. Let $\mathrm{d}P = f\mathrm{d}\mu$ and $\mathrm{d}Q = g\mathrm{d}\mu$ be two probability measures on $(\mathbb{Y},\mathcal{F})$. Then, the Kullback-Leibler divergence between P and Q is noted $KL(P||Q)$ and is defined by

$$KL(P||Q) = \int f(y)\log\frac{f(y)}{g(y)}\mu(\mathrm{d}y) .$$

Two properties are often used:

- $KL(P||Q) \geqslant 0$ with equality if and only if $P = Q$.
- $KL(P||Q) \neq KL(Q||P)$.

We now delve into the connections between the Kullback-Leibler divergence and the Maximum Likelihood Estimator (MLE). The proof of the strong consistency of the MLE follows a general scheme: $\hat{\theta}_n^{\text{ML}}$ is defined as the argmax of $n^{-1}\sum_{i=1}^n \log\ell_\theta(Y_i)$, which is shown to converge almost surely for a fixed θ to the limiting function $\mathbb{E}_{\theta_\star}[\log\ell_\theta(Y_1)]$ by the SLLN (Strong Law of Large Numbers). Assuming that this limiting function achieves its maximum only at $\theta_\star$, it is natural to expect that $\hat{\theta}_n^{\text{ML}}$ will subsequently converge to the argmax with respect to θ of the limiting function, i.e., $\theta_\star$.

$$\begin{array}{ccc} \hat{\theta}_n^{\text{ML}} & \in \text{argmax}_{\theta\in\Theta} & n^{-1}\sum_{i=1}^n \log\ell_\theta(Y_i) \\ \vdots & & \Big\downarrow \text{SLLN} \\ \theta_\star & \in \text{argmax}_{\theta\in\Theta} & \mathbb{E}_{\theta_\star}[\log\ell_\theta(Y_1)] \end{array}$$

Actually, we have the property

$$\mathbb{E}_{\theta_\star}[\log \ell_{\theta_\star}(Y_1)] = \max_{\theta\in\Theta} \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] \,.$$

Indeed

$$\mathbb{E}_{\theta_\star}[\log \ell_{\theta_\star}(Y_1)] - \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] = KL(P_{\theta_\star}||P_\theta) \geqslant 0 \,,$$

where we have defined $\mathrm{d}P_\theta = \ell_\theta \mathrm{d}\mu$ for any $\theta \in \Theta$. This link with the Kullback-Leibler divergence shows that if we assume (A4)-(ii), then $\theta_\star$ turns out to be the only argmax of $\mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)]$ that is for any $\theta \neq \theta_\star$,

$$\mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] < \mathbb{E}_{\theta_\star}[\log \ell_{\theta_\star}(Y_1)] \,. \tag{2.10}$$

2.5.3 *M*-estimators

It turns out that MLEs are just specific instances of *M*-estimators. Let us provide a brief introduction to this broader class of estimators, as they are frequently employed in practical applications.

Definition 2.38 (***M*-estimators).** We say that $\hat{\theta}_n^M$ is an *M*-estimator if for some family of functions $(\varphi_\theta)_{\theta\in\Theta}$,

$$\hat{\theta}_n^M \in \mathrm{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \varphi_\theta(Y_i) \,.$$

►**Q-2.12.** Great! So, if I am not mistaken, the MLE is just an *M*-estimator with $\varphi_\theta(y) = \log \ell_\theta(y)$.

Exactly, but in general, we cannot choose any function φ_θ. We should, at least, have the property $\mathbb{E}_\theta[\varphi_\theta(Y_1)] = \max_\theta \mathbb{E}_\theta[\varphi_\theta(Y_1)]$, which is, of course, satisfied in the particular case: $\varphi_\theta(y) = \log \ell_\theta(y)$. Now, let's mention the strong consistency and the asymptotic normality properties for *M*-estimators. As before, I have decided to postpone the exact statements of the assumptions to the appendix, just to focus on the results. I have also omitted the proofs, as they are very similar to the MLE case. It would be a valuable exercise for you to show Proposition 2.39 and Proposition 2.40 on your own, by following the proofs of Proposition 2.35 and Proposition 2.36 in the general context of *M*-estimators.

Proposition 2.39 (**Strong Consistency for *M*-estimators).** Under (B1),

$$\lim_{n\to\infty} \hat{\theta}_n^M = \theta_\star \,, \quad \mathbb{P}_{\theta_\star} - a.s.$$

Proposition 2.40 (**Asymptotic normality for *M*-estimators).** Under (B2),

$$\sqrt{n}(\hat{\theta}_n^M - \theta_\star) \overset{\mathcal{L}_{\mathbb{P}_{\theta_\star}}}{\rightsquigarrow} \mathcal{N}(0, U_{\theta_\star} G_{\theta_\star} U_{\theta_\star}^T)$$

where

$$U_{\theta_\star} = \left(\mathbb{E}_{\theta_\star}\left[\frac{\partial^2 \varphi_\theta(Y_1)}{\partial\theta\partial\theta^T}\bigg|_{\theta=\theta_\star} \right] \right)^{-1} = U_{\theta_\star}^T \,,$$

$$G_{\theta_\star} = \mathbb{E}_{\theta_\star}\left[\frac{\partial \varphi_\theta(Y_1)}{\partial\theta} \left(\frac{\partial \varphi_\theta(Y_1)}{\partial\theta} \right)^T \bigg|_{\theta=\theta_\star} \right] \,.$$

►**Q-2.13.** If you have an estimator, let's say $\hat{\theta}_n$, and you are interested in its asymptotic properties, what step-by-step approach would you recommend?

I would suggest the following approach:

(i) First, check if you have an explicit expression for your estimator $\hat{\theta}_n$. Often, Method of Moments (MOM) estimators, Maximum Likelihood Estimators (MLEs), or M-estimators can be directly expressed from the observations as a function of $n^{-1}\sum_{i=1}^n h(Y_i)$, where h is a particular function.

(ii) Next, try to obtain the properties using either the "Law of Large Numbers (LGN)", "Continuous Mapping Theorem", "Central Limit Theorem (CLT)", "Slutsky's Theorem" or any other relevant theorem. Another lemma may be of interest for solving the exercises of this chapter (we will consistently come back to this lemma in the next chapter on confidence intervals):

Lemma 2.41 (The δ-method). Assume that there exist a sequence of random variables $\{\hat{\theta}_n : n \in \mathbb{N}\}$, a random variable U, a constant θ and a sequence of positive real numbers $\{r_n : n \in \mathbb{N}\}$ such that

$$\lim_n r_n = \infty, \quad \text{and} \quad r_n(\hat{\theta}_n - \theta) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} U .$$

Then, for any measurable function $g : \mathbb{R} \to \mathbb{R}$, differentiable at θ, we have

$$r_n(g(\hat{\theta}_n) - g(\theta)) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} g'(\theta)U .$$

(iii) If your estimator is a MLE, verify if you can apply Proposition 2.35 and Proposition 2.36. For M-estimators, check if you can apply Proposition 2.39 and Proposition 2.40.

(iv) Combine (iii) and (ii).

(v) If none of the above methods works, try to mimick/extend the proofs in the appendix.

Even if you have a MLE or an M-estimator, try (ii) before (iii); it may be quicker to obtain the results.

2.6 After studying this chapter...

a) I know the definition of a sufficient statistic and I know how to find a sufficient statistic by using the factorization theorem.

b) I can detect an exponential family and I can find the associated natural parameters.

c) I can calculate a likelihood, a score, a Fisher Information matrix, especially for iid models.

d) I know the definiton of an estimator and I can calculate its bias.

e) I can implement the MOM (Method of Moments), the MLE (Maximum Likelihood Estimator) and more generally, M-estimators and I know their properties (strong consistency and asymptotic normality). I can try to check its asymptotic properties using the approach given in Q-2.13.

f) I know the Rao-Blackwell theorem, the Cramér-Rao bound and the definition of MVUE (Minimum Variance Unbiased Estimators) and efficient estimators.

2.7 Highlights

Harald Cramér (source: Wikipedia)

Harald Cramér (25 September 1893 – 5 October 1985) was a Swedish mathematician, actuary, and statistician, specializing in mathematical statistics and probabilistic number theory. John Kingman described him as "one of the giants of statistical theory".

Harald Cramér was born in Stockholm, Sweden on 25 September 1893. Cramér remained close to Stockholm for most of his life. He entered the University of Stockholm as an undergraduate in 1912, where he studied mathematics and chemistry. During this period, he was a research assistant under the famous chemist, Hans von Euler-Chelpin, with whom he published his first five articles from 1913 to 1914.Following his lab experience, he began to focus solely on mathematics. He eventually began his work on his doctoral studies in mathematics which were supervised by Marcel Riesz at the University of Stockholm. Also influenced by G. H. Hardy, Cramér's research led to a PhD in 1917 for his thesis "On a class of Dirichlet series".

Following his PhD, he served as an Assistant Professor of Mathematics at Stockholm University from 1917 to 1929. Early on, Cramér was highly involved in analytic number theory. He also made some important statistical contributions to the distribution of primes and twin primes. His most famous paper on this subject is entitled "On the order of magnitude of the difference between consecutive prime numbers", which provided a rigorous account of the constructive role in which probability applied to number theory and included an estimate for prime gaps that became known as Cramér's conjecture.

In the late 1920s, Cramér became interested in the field of probability, which at the time was not an accepted branch of mathematics. Cramér knew that a radical change was needed in this field, and in a paper in 1926 said, "The probability concept should be introduced by a purely mathematical definition, from which its fundamental properties and the classical theorems are deduced by purely mathematical operations." Cramér took an interest in the rigorous mathematical formulation of probability in the work of French and Russian mathematicians such as Kolmogorov, Lévy, Bernstein, and Khinchin in the early 1930s. Cramér also made significant development to the revolution in probability theory. Cramér later wrote his careful study of the field in his Cambridge publication Random variables and probability distributions which appeared in 1937 (with a 2nd edition in 1962 and a 3rd edition in 1970). Shortly after World War II, Cramér went on to publish the influential Mathematical Methods of Statistics in 1946. This text was one that "showed the way in which statistical practice depended on a body of rigorous mathematical analysis as well as Fisherian intuition." His 1955 book Elements of Probability Theory and Some of its Applications introduces probability theory at a more elementary level than Mathematical Methods of Statistics.

In 1929, Cramér was appointed to a newly created chair in Stockholm University, becoming the first Swedish professor of Actuarial Mathematics and Mathematical Statistics. Cramér retained this position up until 1958. During his tenure at Stockholm University, Cramér was a PhD advisor for 10 students, most notably Herman Wold and Kai Lai Chung. In 1950 he was elected as a Fellow of the American Statistical Association. Starting in 1950, Cramér took on the additional responsibility of becoming the President of Stockholm University. In 1958, he was also appointed to be Chancellor of the entire Swedish university system. Cramér retired from the Swedish university system in 1961.

A large portion of Cramér's work concerned the field of actuarial science and insurance mathematics. During the period from 1920 to 1929, he was an actuary for the life insurance company Svenska livförsäkringsbolaget. His actuarial work during this time led him to study probability and statistics which became the main area of his research. In 1927 he published an elementary text in Swedish Probability theory and some of its applications. Following his work for Svenska livförsäkringsbolaget, he went on to work for Återförsäkringsaktiebolaget Sverige, a reinsurance company, up until 1948. He was also known for his pioneering efforts in insurance risk theory. After this period, he remained as a consultant actuary to Sverige from 1949 to 1961. Later in his life, he was elected to be the Honorary President of the Swedish Actuarial Society.

Cramér remained an active contributor to his professional career for an additional 20 years. Following his retirement in 1961, he became extremely active in research, which had been slowed due to his Chancellorship. During the years from 1961 to 1983, Cramér traveled throughout the United States and Europe to continue his research, making significant stops at Berkeley, Princeton, and at the Research Triangle Institute of North Carolina.

Cramér received an Honorary Doctorate from Heriot-Watt University in 1972.

His academic career spanned over seven decades, from 1913 to 1982.

Harald Cramér married Marta Hansson in 1918, and they remained together up until her death in 1973. He had often referred to her as his "Beloved Marta". Together they had one daughter, Marie-Louise, and two sons, Tomas and Kim.

2.8 Appendix

2.8.1 Proof of Proposition 2.35

(A4) (i) The set $\Theta \subset \mathbb{R}^d$ is compact.

(ii) for all $\theta, \theta' \in \Theta$ such that $\theta \neq \theta'$, we have $P_\theta \neq P_{\theta'}$ where we set $P_\theta(\mathrm{d}y_1) = \ell_\theta(y_1)\mu(\mathrm{d}y_1)$.

(iii) $\mathbb{E}_{\theta_\star}\left[\sup_{\theta\in\Theta} |\log \ell_\theta(Y_1)|\right] < \infty$.

(iv) $\mathbb{P}_{\theta_\star} - a.s.$, the function $\theta \mapsto \ell_\theta(Y_1)$ is upper-semicontinuous.

where we recall that a function $f : \Theta \subset \mathbb{R}^d \to \mathbb{R}$ is upper-semicontinuous if and only if for any $\theta_0 \in \Theta$, $\lim_{\theta\to\theta_0} f(\theta) \leqslant f(\theta_0)$. Note that in such a case, f attains its maximum in any compact set.

Proof. [of Proposition 2.35] For any $\theta \in \Theta$, assumption (A4)-(iii) allows to apply the SLLN and we have

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} \log \ell_\theta(Y_k) = \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)], \quad \mathbb{P}_{\theta_\star} - a.s. \tag{2.11}$$

For any $\rho > 0$ and $\theta_0 \in \Theta$, define $\mathsf{B}(\theta_0, \rho) = \{\theta \in \Theta : \|\theta - \theta_0\| < \rho\}$ where $\|\cdot\|$ is any norm on $\mathbb{R}^d$. Let K be a compact subset of Θ. For all $\theta_0 \in K$, $\mathbb{P}_{\theta_\star} - a.s.$,

$$\begin{aligned}\limsup_{\rho\searrow 0} \limsup_{n\to\infty} \sup_{\theta\in\mathsf{B}(\theta_0,\rho)} n^{-1}\sum_{k=0}^{n-1} \log \ell_\theta(Y_k) \\ \leqslant \limsup_{\rho\searrow 0} \limsup_{n\to\infty} n^{-1}\sum_{k=0}^{n-1}\left(\sup_{\theta\in\mathsf{B}(\theta_0,\rho)} \log \ell_\theta(Y_k)\right) = \limsup_{\rho\searrow 0} \mathbb{E}_{\theta_\star}\left[\sup_{\theta\in\mathsf{B}(\theta_0,\rho)} \log \ell_\theta(Y_1)\right].\end{aligned} \tag{2.12}$$

By the monotone convergence applied to the non-increasing function $\rho \mapsto \sup_{\theta\in\mathsf{B}(\theta_0,\rho)} \log \ell_\theta$, we have

$$\limsup_{\rho\searrow 0} \mathbb{E}_{\theta_\star}\left[\sup_{\theta\in\mathsf{B}(\theta_0,\rho)} \log \ell_\theta(Y_1)\right] = \mathbb{E}_{\theta_\star}\left[\limsup_{\rho\searrow 0} \sup_{\theta\in\mathsf{B}(\theta_0,\rho)} \log \ell_\theta(Y_1)\right] \leqslant \mathbb{E}_{\theta_\star}\left[\log \ell_{\theta_0}(Y_1)\right], \tag{2.13}$$

where the last inequality follows from (A4)-(iv). Combining (2.12) and (2.13), we obtain that for all $\eta > 0$ and all $\theta_0 \in K$, there exists $\rho^{\theta_0} > 0$ satisfying

$$\limsup_{n\to\infty} \sup_{\theta\in\mathsf{B}(\theta_0,\rho^{\theta_0})} n^{-1}\sum_{k=0}^{n-1} \log \ell_\theta(Y_k) \leqslant \mathbb{E}_{\theta_\star}\left[\log \ell_{\theta_0}(Y_1)\right] + \eta \leqslant \sup_{\theta\in K} \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] + \eta, \quad \mathbb{P}_{\theta_\star} - a.s.$$

Since K is a compact subset of Θ, we can extract a finite sub-cover of K from $\cup_{\theta_0\in K}\mathsf{B}(\theta_0, \rho^{\theta_0})$, so that

$$\limsup_{n\in\infty} \sup_{\theta\in K} n^{-1}\sum_{k=0}^{n-1} \log \ell_\theta(Y_k) \leqslant \sup_{\theta\in K} \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] + \eta, \quad \mathbb{P}_{\theta_\star} - a.s.$$

Since η is arbitrary, we obtain

$$\limsup_{n\in\infty} \sup_{\theta\in K} n^{-1}\sum_{k=0}^{n-1} \log \ell_\theta(Y_k) \leqslant \sup_{\theta\in K} \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)], \quad \mathbb{P}_{\theta_\star} - a.s. \tag{2.14}$$

Moreover, $\mathbb{P}_{\theta_\star} - a.s.$ by (2.13), we get

$$\limsup_{\rho\searrow 0} \sup_{\theta\in\mathsf{B}(\theta_0,\rho)} \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] \leqslant \limsup_{\rho\searrow 0} \mathbb{E}_{\theta_\star}\left[\sup_{\theta\in\mathsf{B}(\theta_0,\rho)} \log \ell_\theta(Y_1)\right] \leqslant \mathbb{E}_{\theta_\star}\left[\log \ell_{\theta_0}(Y_1)\right].$$

This shows that $\theta \mapsto \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)]$ is upper-semicontinuous. For all $\varepsilon > 0$, $K_\varepsilon := \{\theta \in \Theta : \|\theta - \theta_\star\| \geqslant \varepsilon\}$ is a compact subset of Θ. Using the upper-semicontinuity of $\theta \mapsto \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)]$, there exists $\theta_\varepsilon \in K_\varepsilon$ such that

$$\sup_{\theta \in K_\varepsilon} \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] = \mathbb{E}_{\theta_\star}\left[\log \ell_{\theta_\varepsilon}(Y_1)\right] < \mathbb{E}_{\theta_\star}[\log \ell_{\theta_\star}(Y_1)] \ ,$$

where the strict inequality follows from (2.10). Finally, combining this inequality with (2.14), we obtain $\mathbb{P}_{\theta_\star} - a.s.$,

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\theta \in K_\varepsilon} n^{-1} \sum_{k=0}^{n-1} \log \ell_\theta(Y_k) &\leqslant \sup_{\theta \in K_\varepsilon} \mathbb{E}_{\theta_\star}[\log \ell_\theta(Y_1)] \\ &< \mathbb{E}_{\theta_\star}[\log \ell_{\theta_\star}(Y_1)] \overset{(1)}{=} \lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \log \ell_{\theta_\star}(Y_k) \leqslant \liminf_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \log \ell_{\hat\theta_n^{\mathrm{ML}}}(Y_k) \ ,\end{aligned}$$

where $\overset{(1)}{=}$ follows from (A4)-(iii) and the SLLN. This property ensures that $\hat\theta_n^{\mathrm{ML}} \notin K_\varepsilon$, hence $\|\theta - \theta_\star\| \leqslant \varepsilon$ for all n larger to some $\mathbb{P}_{\theta_\star} - a.s.$ finite integer-valued random variable. The proof is completed since ε is arbitrary. ■

2.8.2 Proof of Proposition 2.36

(A5) (i) The set $\Theta \subset \mathbb{R}^d$ is compact and $\theta_\star \in \mathring{\Theta}$ where $\mathring{\Theta}$ denotes the interior of Θ.

(ii) $\mathbb{P}_{\theta_\star}$-a.s., $\lim_n \hat\theta_n^{\mathrm{ML}} = \theta_\star$.

(iii) For μ_1-almost all y_1, $\theta \mapsto \ell_\theta(y_1)$ is twice continuously differentiable on Θ and

$$\int \sup_{\theta\in\Theta} \left\| \frac{\partial \ell_\theta(y_1)}{\partial \theta} \right\| \mu_1(\mathrm{d}y_1) < \infty, \tag{2.15}$$

$$\int \sup_{\theta\in\Theta} \left\| \frac{\partial^2 \ell_\theta(y_1)}{\partial \theta \partial \theta^T} \right\| \mu_1(\mathrm{d}y_1) < \infty, \tag{2.16}$$

$$\mathbb{E}_{\theta_\star}\left[\sup_{\theta\in\Theta} \left\| \frac{\partial^2 \log \ell_\theta(Y_1)}{\partial \theta \partial \theta^T} \right\| \right] < \infty, \tag{2.17}$$

where by abuse of notation $\|\cdot\|$ denotes any norm on $\mathbb{R}^d$ or $\mathbb{R}^{d\times d}$.

(iv) $I_F^{\theta_\star}(Y_1)$ is invertible.

Note that under (A5)-(iii), the score $\xi_{\theta_\star}(Y_1)$ is centered and square-integrable under $\mathbb{P}_{\theta_\star}$ and

$$I_F^{\theta_\star}(Y_1) = \mathbb{V}\mathrm{ar}_{\theta_\star}(\xi_{\theta_\star}(Y_1)) = -\mathbb{E}_{\theta_\star}\left[\left. \frac{\partial^2 \log \ell_\theta(Y_1)}{\partial \theta \partial \theta^T} \right|_{\theta=\theta_\star} \right] .$$

Proof. [of Proposition 2.36] By (A5)-(i)-(ii), $\hat\theta_n^{\mathrm{ML}}$ is in $\mathring{\Theta}$ for n larger to some $\mathbb{P}_{\theta_\star}$-a.s. finite random integer. Then, for such n, writing a Taylor expansion with integral reminder yields

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[n^{-1/2} \sum_{k=1}^n \log \ell_\theta(Y_k) \right]\Big|_{\theta=\hat\theta_n^{\mathrm{ML}}} &= 0 \\ = \underbrace{n^{-1/2} \sum_{k=1}^n \left. \frac{\partial \log \ell_\theta(Y_k)}{\partial \theta} \right|_{\theta=\theta_\star}}_{I_n(\theta_\star)} &+ \underbrace{n^{-1} \sum_{k=1}^n \left(\int_0^1 \left. \frac{\partial^2 \log \ell_\theta(Y_k)}{\partial \theta \partial \theta^T} \right|_{\theta=(1-s)\theta_\star + s\hat\theta_n^{\mathrm{ML}}} \mathrm{d}s \right)}_{J_n(\hat\theta_n^{\mathrm{ML}})} \left[\sqrt{n}(\hat\theta_n^{\mathrm{ML}} - \theta_\star)\right] \ ,\end{aligned}$$

which implies

$$\sqrt{n}(\hat\theta_n^{\mathrm{ML}} - \theta_\star) = -J_n(\hat\theta_n^{\mathrm{ML}})^{-1} I_n(\theta_\star)$$

Then, provided that we can show that

$$I_n(\theta_\star) \overset{\mathcal{L}_{\mathbb{P}_{\theta_\star}}}{\rightsquigarrow} \mathcal{N}\left(0, I_F^{\theta_\star}(Y_1)\right) \ , \tag{2.18}$$

$$J_n(\hat\theta_n^{\mathrm{ML}}) \overset{\mathbb{P}_{\theta_\star}-prob}{\longrightarrow} -I_F^{\theta_\star}(Y_1) \ , \tag{2.19}$$

we finally obtain according to the multidimensional Slutsky lemma combined with (A5)-(iv)

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{ML}} - \theta_\star) \overset{\mathcal{L}_{\mathbb{P}_{\theta_\star}}}{\rightsquigarrow} \mathcal{N}(0, \underbrace{\left[I_F^{\theta_\star}(Y_1)\right]^{-1} I_F^{\theta_\star}(Y_1) \left[I_F^{\theta_\star}(Y_1)^T\right]^{-1}}_{\left[I_F^{\theta_\star}(Y_1)\right]^{-1}}) \,.$$

We now turn to the proof of (2.18) and (2.19). The Central Limit Theorem applied to the sequence of random vectors $\{\xi_{\theta_\star}(Y_k) \,:\, k \in \mathbb{N}\}$ which are centered under $\mathbb{P}_{\theta_\star}$ yields

$$I_n = n^{-1/2} \sum_{k=1}^{n} \xi_{\theta_\star}(Y_k) \overset{\mathcal{L}_{\mathbb{P}_{\theta_\star}}}{\rightsquigarrow} \mathcal{N}(0, I_F^{\theta_\star}(Y_1)) \,.$$

This shows (2.18). It remains to prove (2.19). Let $V_\rho = \{\theta \in \Theta \,:\, \|\theta - \theta_\star\| \leqslant \rho\}$ and define for any $y \in \mathbb{Y}_1$,

$$H_\rho(y) = \sup_{\theta \in V_\rho} \left\| \frac{\partial^2 \log \ell_\theta(y)}{\partial\theta\partial\theta^T} - \frac{\partial^2 \log \ell_\theta(y)}{\partial\theta\partial\theta^T}\bigg|_{\theta=\theta_\star} \right\| .$$

Under (A5)-(iii), $\lim_{\rho\to 0} H_\rho(y) = 0$ for μ_1-almost all $y \in \mathbb{Y}_1$ and $|H_\rho(y)| \leqslant 2\sup_{\theta\in\Theta} \left\| \frac{\partial^2 \log \ell_\theta(y)}{\partial\theta\partial\theta^T} \right\|$. By (2.17), we can apply the Lebesgue dominated convergence theorem, so that

$$\lim_{\rho\to 0} \mathbb{E}[H_\rho(Y_1)] = 0 \,.$$

Hence, for any $\varepsilon > 0$, there exists $\rho_\varepsilon > 0$ sufficiently small such that $\mathbb{E}_{\theta_\star}\left[H_{\rho_\varepsilon}(Y_1)\right] < \varepsilon$. Then,

$$\begin{aligned}
\mathbb{P}_{\theta_\star}(\left\|J_n(\hat{\theta}_n^{\mathrm{ML}}) - J_n(\theta_\star)\right\| > \varepsilon) &\leqslant \mathbb{P}_{\theta_\star}(\left\|J_n(\hat{\theta}_n^{\mathrm{ML}}) - J_n(\theta_\star)\right\| > \varepsilon, \hat{\theta}_n^{\mathrm{ML}} \in V_{\rho_\varepsilon}) + \mathbb{P}_{\theta_\star}(\hat{\theta}_n^{\mathrm{ML}} \notin V_{\rho_\varepsilon}) \\
&\leqslant \mathbb{P}_{\theta_\star}(n^{-1} \sum_{k=1}^{n} H_{\rho_\varepsilon}(Y_k) > \varepsilon) + \mathbb{P}_{\theta_\star}(\left\|\hat{\theta}_n^{\mathrm{ML}} - \theta_\star\right\| > \varepsilon) \,. \qquad (2.20)
\end{aligned}$$

By the SLLN, $\lim_n n^{-1}\sum_{k=1}^n H_{\rho_\varepsilon}(Y_k) = \mathbb{E}_{\theta_\star}\left[H_{\rho_\varepsilon}(Y_1)\right] < \varepsilon$, $\mathbb{P}_{\theta_\star}$-a.s. This limiting result combined with $\hat{\theta}_n^{\mathrm{ML}} \overset{\mathbb{P}_{\theta_\star}-prob}{\longrightarrow} \theta_\star$ shows that the rhs of (2.20) tends to 0 as n tends to infinity. Hence, $J_n(\hat{\theta}_n^{\mathrm{ML}}) - J_n(\theta_\star) \overset{\mathbb{P}_{\theta_\star}-prob}{\longrightarrow} 0$ and since applying again the SLLN, $J_n(\theta_\star) \overset{\mathbb{P}_{\theta_\star}\text{-a.s.}}{\longrightarrow} -I_F^{\theta_\star}(Y_1)$, we can conclude that (2.19) holds. The proof is completed. ■

2.8.3 Assumptions for the asymptotic properties of M-estimators

(B1)
- (i) The set $\Theta \subset \mathbb{R}^d$ is compact.
- (ii) for all $\theta \neq \theta_\star$, $\mathbb{E}_{\theta_\star}[\varphi_\theta(Y_1)] < \mathbb{E}_{\theta_\star}[\varphi_{\theta_\star}(Y_1)]$.
- (iii) $\mathbb{E}_{\theta_\star}\left[\sup_{\theta\in\Theta} |\varphi_\theta(Y_1)|\right] < \infty$.
- (iv) $\mathbb{P}_{\theta_\star} - a.s.$, the function $\theta \mapsto \varphi_\theta(Y_1)$ is upper-semicontinuous.

(B2)
- (i) The set $\Theta \subset \mathbb{R}^d$ is compact and $\theta_\star \in \mathring{\Theta}$ where $\mathring{\Theta}$ denotes the interior of Θ.
- (ii) $\mathbb{P}_{\theta_\star}$-a.s., $\lim_n \hat{\theta}_n^M = \theta_\star$.
- (iii) For μ_1-almost all y_1, $\theta \mapsto \varphi_\theta(y_1)$ is twice continuously differentiable on Θ and

$$\mathbb{E}_{\theta_\star}\left[\left\| \frac{\partial\varphi_\theta(Y_1)}{\partial\theta}\bigg|_{\theta=\theta_\star} \right\|^2\right] < \infty, \qquad (2.21)$$

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left\| \frac{\partial^2\varphi_\theta(Y_1)}{\partial\theta\partial\theta^T} \right\|\right] < \infty. \qquad (2.22)$$

- (iv) $\mathbb{E}_{\theta_\star}\left[\frac{\partial^2\varphi_\theta(Y_1)}{\partial\theta\partial\theta^T}\Big|_{\theta=\theta_\star}\right]$ is invertible.

Chapter 3

Confidence regions

In the previous chapter, given a function of interest $\theta \mapsto g(\theta) \in \mathbb{R}^p$ and observations $\mathsf{Y} = (Y_1, \dots, Y_n) \in \mathbb{Y}$, we provided a function of the observation, written as a *vector* $T(\mathsf{Y})$ which is supposed to approximate $g(\theta)$. By doing so, we have defined the notion of an *estimator* of $g(\theta)$.

We now would like to be more careful and, instead of providing a single function which is supposed to approximate $g(\theta)$, we aim to offer an entire *region* $\mathcal{C}(\mathsf{Y}) \subset \mathbb{R}^p$ where our function of interest $g(\theta)$ should lie with a "high" probability. This leads to the notion of *confidence regions* (when the dimension is $p > 1$) and *confidence intervals* (when the dimension is $p = 1$).

3.1 Confidence regions for a finite sample

3.1.1 Level of a confidence region

In this section, the function of interest is given $\theta \mapsto g(\theta) \in \mathbb{R}^p$ and we denote by $\mathsf{Y} = (Y_1, \dots, Y_n) \in \mathbb{Y}$ the observations. We start with a formal definition of confidence regions.

Definition 3.1 (Confidence regions). Let $(\mathbb{Y}, \mathcal{F}, Q)$ be a parametric statistical model i.e. $Q = (\mathbb{P}_\theta)_{\theta \in \Theta}$. A *confidence region* $\mathcal{C}$ with level $1 - \alpha$ is defined by

$$\mathbb{P}_\theta(g(\theta) \in \mathcal{C}(\mathsf{Y})) \geqslant 1 - \alpha, \quad \forall \theta \in \Theta \,,$$

or equivalently,

$$\mathbb{P}_\theta(g(\theta) \notin \mathcal{C}(\mathsf{Y})) \leqslant \alpha, \quad \forall \theta \in \Theta \,.$$

The level of a confidence region is, therefore, between 0 and 1, and the higher the level, the better the confidence region. However, a higher level also leads to a larger region, which may become impractical. As an extreme case, when $\mathcal{C}(\mathsf{Y}) = g(\Theta)$, the level of this region is 1, but its size is the largest possible, providing no useful information about the location of $g(\theta)$. Thus, a good balance between the level of a confidence region and its size is necessary. It is worth noting that to ensure well-defined probabilities in Definition 3.1, we must assume that for any $\theta \in \Theta$,

$$\{y \in \mathbb{Y} \,:\, g(\theta) \in \mathcal{C}(y)\} \in \mathcal{F} \,.$$

As a final remark, the notation $\mathcal{C}(\mathsf{Y})$ implicitly means that the confidence region is a function of the observations Y only and does not depend on the parameter θ.

To illustrate the notion of confidence regions, let us first consider a poll example that will be treated extensively.

Example 3.2 (A poll example)**.** In this example, we have a group of n individuals, and we have collected their intentions to vote or not vote for a certain candidate, denoted as X. Specifically, let Y_i represent the intention of individual i to vote for candidate X. The variable Y_i takes only two values, 0 or 1, and we can model it as $Y_i \sim \mathcal{B}(\theta)$ under $\mathbb{P}_\theta$, where $\theta \in (0,1)$ represents the probability of voting for the candidate. Given the observed data $Y = (Y_1, \ldots, Y_n)$, our goal is to construct a confidence interval for θ with level $1-\alpha$.

As a numerical example, consider $n = 1500$, and out of this group, there are 789 individuals who intend to vote for candidate X. Now, we want to find a confidence interval with a 95% level for the probability of candidate X winning the election.

Considering Definition 3.1, we aim to construct an interval I_α such that for any $\theta \in [0,1]$,

$$\mathbb{P}_\theta(\theta \notin I_\alpha) \leqslant \alpha\,. \tag{3.1}$$

We will discuss two methods.

First method

Write $\hat{\theta}_n = \sum_{i=1}^n Y_i/n$. Since $\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[Y_1] = \theta$, the Bienaymé-Tchebychev inequality yields

$$\forall \varepsilon > 0, \quad \mathbb{P}_\theta(|\hat{\theta}_n - \theta| > \varepsilon) \leqslant \frac{\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n)}{\varepsilon^2} = \frac{\theta(1-\theta)}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}\,. \tag{3.2}$$

The approach consists of rewriting (3.2) as (3.1). To achieve this, we choose ε such that the right-hand sides in (3.2) and (3.1) coincide. That is, we choose ε such that $\frac{1}{4n\varepsilon^2} = \alpha$, which leads to $\varepsilon = \frac{1}{2\sqrt{n\alpha}}$. With this chosen value of ε, we now focus on the left-hand sides of (3.1) and (3.2).

Let I_α be such that $\{|\hat{\theta}_n - \theta| > \varepsilon\} = \{\theta \notin I_\alpha\}$. This condition implies that $I_\alpha = [\hat{\theta}_n \pm \varepsilon] = \left[\hat{\theta}_n \pm \frac{1}{2\sqrt{n\alpha}}\right]$. Finally, (3.1) is proven for the following choice of I_α:

$$I_\alpha = \left[\frac{\sum_{i=1}^n Y_i}{n} - \frac{1}{2\sqrt{n\alpha}}\,,\, \frac{\sum_{i=1}^n Y_i}{n} + \frac{1}{2\sqrt{n\alpha}}\right]\,. \tag{3.3}$$

In other words, this choice of I_α yields a confidence interval for θ with level $1-\alpha$.

In the numerical example, we have $\alpha = 5\%$, $\hat{\theta}_n = 789/1500$. Then $I_\alpha = [0.47, 0.58]$.

Second method

As in the first method, let's set $\hat{\theta}_n = \sum_{i=1}^n Y_i/n$. Using Hoeffding's inequality for the independent and bounded random variables (Y_i), we obtain for any $\varepsilon > 0$,

$$\mathbb{P}_\theta(|\hat{\theta}_n - \theta| > \varepsilon) \leqslant 2\mathrm{e}^{-2n\varepsilon^2}\,. \tag{3.4}$$

As before, we rewrite (3.4) as (3.1). First, we work on the rhs of these two inequalities, and we choose ε such that $2\mathrm{e}^{-2n\varepsilon^2} = \alpha$, ie: $\varepsilon = \sqrt{-\frac{1}{2n}\log\left(\frac{\alpha}{2}\right)}$. With this selected ε, we choose a convenient I_α such that the lhs of (3.4) and (3.1) coincide. Writing $\left\{|\hat{\theta}_n - \theta| > \varepsilon\right\} = \{\theta \notin I_\alpha\}$ yields $I_\alpha = [\hat{\theta}_n \pm \varepsilon] = [\hat{\theta}_n \pm \sqrt{-\frac{1}{2n}\log\left(\frac{\alpha}{2}\right)}]$, which can be expressed as follows:

$$I_\alpha = \left[\frac{\sum_{i=1}^n Y_i}{n} - \sqrt{-\frac{1}{2n}\log\left(\frac{\alpha}{2}\right)}\,,\, \frac{\sum_{i=1}^n Y_i}{n} + \sqrt{-\frac{1}{2n}\log\left(\frac{\alpha}{2}\right)}\right]\,. \tag{3.5}$$

In the previous numerical example where $\alpha = 5\%$ and $\hat{\theta}_n = 789/1500$, we get $I_\alpha = [0.50, 0.55]$.

It is important to note that (3.3) and (3.5) provide confidence intervals for θ with a level of $1-\alpha$. However, if we denote by $\ell_1(\alpha)$ and $\ell_2(\alpha)$ the sizes of the confidence intervals in (3.3) and (3.5), respectively, we obtain:

$$\ell_1(\alpha) = \frac{1}{\sqrt{n\alpha}}\,, \quad \ell_2(\alpha) = \sqrt{-\frac{2}{n}\log\left(\frac{\alpha}{2}\right)}\,.$$

If we wish to have confidence intervals with a level larger than 0.8 (i.e., $\alpha < 0.2$), it can be readily checked that $\ell_2(\alpha) \leqslant \ell_1(\alpha)$. Hence, in this case, the second method is preferable to the first one.

►**Q-3.1.** Ok, thanks for this nice example on Bernoulli variables... Do you have another example with, for instance, real-valued random variables?

Sure, let's consider another example involving Gaussian random variables.

Example 3.3. Assuming that under $\mathbb{P}_\theta$, the random variables $(Y_i)_{i\in\mathbb{N}}$ are independent and identically distributed, with $Y_1 \sim \mathcal{N}(\theta,\sigma^2)$, where σ^2 is known, we can construct a confidence interval for θ based on the observations $Y = (Y_1,\dots,Y_n)$.

Setting $\hat{\theta}_n = \sum_{i=1}^n Y_i/n$, we have under $\mathbb{P}_\theta$,

$$Z_n := \frac{\hat{\theta}_n - \theta}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)\,.$$

Now, for any $\alpha > 0$, pick z_α s.t. $\mathbb{P}(Z \notin [-z_\alpha, z_\alpha]) = \alpha$ where $Z \sim \mathcal{N}(0,1)$. Then,

$$\mathbb{P}_\theta\left(Z_n \notin [-z_\alpha, z_\alpha]\right) = \alpha\,.$$

Then, $\{Z_n \notin [-z_\alpha, z_\alpha]\} = \{\theta \notin I_\alpha\}$ where

$$I_\alpha = \left[\hat{\theta}_n - z_\alpha\sqrt{\frac{\sigma^2}{n}}\,, \hat{\theta}_n + z_\alpha\sqrt{\frac{\sigma^2}{n}}\right]\,.$$

Hence, $\mathbb{P}_\theta(\theta \notin I_\alpha) = \alpha \leqslant \alpha$ and we can conclude that I_α is a confidence interval for θ with level $1-\alpha$.

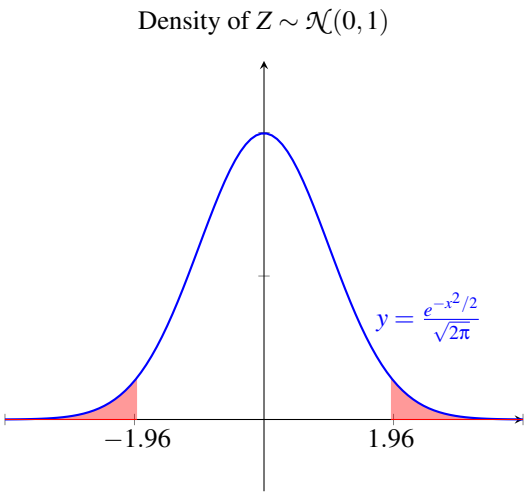


Figure 3.1: For $\alpha = 5\%$, we get $z_\alpha = 1.96$.

3.1.2 Pivot

In Example 3.3, the confidence interval has been constructed using a random variable Z_n such that for any $\theta \in \Theta$, under $\mathbb{P}_\theta$:

$$Z_n = \frac{\sum_{i=1}^n Y_i/n - \theta}{\sqrt{\sigma^2/n}} = G(\mathsf{Y},\theta) \sim \mathcal{N}(0,1)\,.$$

In other words, the random variable Z_n, which depends on both the observations Y and the parameter θ, has a distribution under $\mathbb{P}_\theta$ that *does not depend on* θ. This property shows that G is a pivot, as defined below.

Definition 3.4 (Pivot). Let $(\mathbb{Y}, \mathcal{F}, \mathcal{Q})$ be a parametric statistical model i.e. $\mathcal{Q} = (\mathbb{P}_\theta)_{\theta\in\Theta}$.
A measurable function $G : \mathbb{Y} \times \Theta \to \mathbb{R}^p$ is a *pivot* if and only if the law of $G(\mathsf{Y}, \theta)$ under $\mathbb{P}_\theta$ does not depend on θ.

Hence, $G(\mathsf{Y}, \theta)$ has a distribution under $\mathbb{P}_\theta$ that is free from the parameter θ. In other words, for any $\theta \in \Theta$, the distribution of $G(\mathsf{Y}, \theta)$ under $\mathbb{P}_\theta$ is the same for all $\theta \in \Theta$.

In the context of constructing confidence intervals, a pivot is a crucial concept. It allows us to create a probability statement that holds uniformly for all possible values of the parameter θ. In the case of Example 3.3, the pivot Z_n follows a standard normal distribution under $\mathbb{P}_\theta$ for all $\theta \in \Theta$, and this property ensures that the confidence interval based on Z_n will provide a valid confidence interval of level $1-\alpha$ for the parameter θ for any value of θ in the parameter space Θ.

More generally, if the function G is a pivot, then for any $\alpha > 0$, we first choose D_α such that for any $\theta \in \Theta$,

$$\mathbb{P}_\theta(G(\mathsf{Y}, \theta) \in D_\alpha) = 1 - \alpha \, .$$

Then,

$$\mathcal{C}(\mathsf{Y}) = \{\theta \in \Theta ; G(\mathsf{Y}, \theta) \in D_\alpha\} \tag{3.6}$$

is a confidence region for θ of level $1-\alpha$. Indeed, for any $\theta \in \Theta$,

$$\mathbb{P}_\theta(\theta \in \mathcal{C}(\mathsf{Y})) = \mathbb{P}_\theta(G(\mathsf{Y}, \theta) \in D_\alpha) = 1 - \alpha \, .$$

Note that (3.6) may be rewritten as:

$$G(\mathsf{Y}, \theta) \in D_\alpha \iff \theta \in \mathcal{C}(\mathsf{Y}) \, . \tag{3.7}$$

►**Q- 3.2.** May I interrupt you for a second? Does this equivalence provide the link between pivot functions and confidence intervals?

Exactly, the explicit expression of $\mathcal{C}(\mathsf{Y})$ crucially depends on how $G(\mathsf{Y}, \theta)$ relates to θ. If you have the choice between two pivot functions, I would advise you to choose the one where the dependence on θ is as simple as possible. This will allow you to obtain an explicit confidence region $\mathcal{C}(\mathsf{Y})$ more easily.

In any case, in Example 3.2, we used the Bienaymé-Tchebychev inequality or the Hoeffding inequality, which allows us to bound from above the probability that the absolute value of a centered random variable $|X - \mathbb{E}[X]|$ exceeds a certain threshold. These inequalities are quite useful tools for obtaining confidence intervals, and we will recall them in the next section.

3.1.3 Tools: some useful finite-sample inequalities.

Theorem 3.5 (The Bienaymé-Tchebychev inequality). Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[X^2] < \infty$. Then, for any $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leqslant \frac{\mathbb{V}\mathrm{ar}(X)}{t^2} \, .$$

PROOF. We have, $\mathbb{P}$-a.s.,

$$\mathbf{1}_{\{|X-\mathbb{E}[X]|>t\}} \leqslant \frac{|X - \mathbb{E}[X]|^2}{t^2} \, .$$

Taking the expectation on both sides of the previous inequality yields:

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{E}\left[\mathbf{1}_{\{|X-\mathbb{E}[X]|>t\}}\right] \leqslant \frac{\mathbb{V}\mathrm{ar}(X)}{t^2} \, .$$

■

Theorem 3.6 (The Hoeffding inequality). Let $(X_i)_{1\leqslant i\leqslant n}$ be independent random variables such that for any $i\in\{1,\dots,n\}$, there exist constants a_i, b_i such that

$$a_i \leqslant X_i \leqslant b_i\ , \quad \mathbb{P}-a.s.$$

Then, for any $t>0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \sum_{i=1}^n \mathbb{E}\left[X_i\right]\right| > t\right) \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)\ . \tag{3.8}$$

PROOF. Replacing if necessary X_i by $X_i - \mathbb{E}[X_i]$, we may assume without loss of generality that $\mathbb{E}[X_i]=0$ for all $1\leqslant i\leqslant n$. In that case, to obtain (3.8), it is sufficient to show for all $t>0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)\ . \tag{3.9}$$

Indeed, assume first that (3.9) holds for any sequence of independent centered bounded random variables. Then applying (3.9) with X_i replaced by $-X_i\in[-b_i,-a_i]$ we obtain the same upper-bound for $\mathbb{P}\left(\sum_{i=1}^n X_i > t\right)$ and $\mathbb{P}\left(-\sum_{i=1}^n X_i > t\right)$. The proof of (3.8) is then completed by using

$$\mathbb{P}\left(|\sum_{i=1}^n X_i| > t\right) = \mathbb{P}\left(\sum_{i=1}^n X_i > t\right) + \mathbb{P}\left(\sum_{i=1}^n X_i < -t\right)\ .$$

It now remains to prove (3.9). We have for all $s>0$,

$$\mathbf{1}_{\{\sum_{i=1}^n X_i > t\}} = \mathbf{1}_{\{\mathrm{e}^{-st}\mathrm{e}^{\sum_{i=1}^n sX_i}>1\}} \leqslant \mathrm{e}^{-st}\mathrm{e}^{\sum_{i=1}^n sX_i}\ .$$

Taking the expectation on both sides of the inequality and noting that the (X_i) are independent, we get for all $s,t>0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leqslant \mathrm{e}^{-st}\prod_{i=1}^n \mathbb{E}\left[\mathrm{e}^{sX_i}\right]\ . \tag{3.10}$$

In order to bound the rhs of (3.10), define for all $1\leqslant i\leqslant n$, $\phi_i(s)=\log\psi_i(s)$ and $\psi_i(s)=\mathbb{E}[\mathrm{e}^{sX_i}]$. Since $X_i\in[a_i,b_i]$, ψ_i is twice differentiable and $\psi_i^j(s)=\mathbb{E}[X_i^j\mathrm{e}^{sX_i}]$ for $j\in\{1,2\}$. Then, ϕ_i is twice differentiable and we have $\phi_i'=\frac{\psi_i'}{\psi_i}$ and $\phi_i''=\frac{\psi_i''}{\psi_i}-\left(\frac{\psi_i'}{\psi_i}\right)^2$. Therefore,

$$\phi_i''(s) = \frac{\mathbb{E}[X_i^2\mathrm{e}^{sX_i}]}{\mathbb{E}[\mathrm{e}^{sX_i}]} - \left(\frac{\mathbb{E}[X_i\mathrm{e}^{sX_i}]}{\mathbb{E}[\mathrm{e}^{sX_i}]}\right)^2 = \tilde{\mathbb{V}}(X_i) \quad \text{where } \tilde{\mathbb{V}}(Z)=\tilde{\mathbb{E}}[Z^2]-\tilde{\mathbb{E}}^2[Z] \text{ and } \tilde{\mathbb{E}}[Z]=\frac{\mathbb{E}[Z\mathrm{e}^{sX_i}]}{\mathbb{E}[\mathrm{e}^{sX_i}]}\ .$$

Setting U_i such that $X_i=(b_i-a_i)U_i+a_i$, we have $U_i\in[0,1]$, and therefore:

$$\begin{aligned}\tilde{\mathbb{V}}(X_i) &= (b_i-a_i)^2\tilde{\mathbb{V}}(U_i) = (b_i-a_i)^2(\tilde{\mathbb{E}}[U_i^2]-\tilde{\mathbb{E}}^2[U_i])\\ &\leqslant (b_i-a_i)^2(\tilde{\mathbb{E}}[U_i]-\tilde{\mathbb{E}}^2[U_i]) = (b_i-a_i)^2\underbrace{\tilde{\mathbb{E}}[U_i]}_{\in[0,1]}(1-\tilde{\mathbb{E}}[U_i]) \leqslant (b_i-a_i)^2/4\ .\end{aligned}$$

where we have used $p(1-p)\leqslant 1/4$ for all $p\in[0,1]$. This shows that for all $s>0$,

$$\phi_i''(s) \leqslant \left(\frac{b_i-a_i}{2}\right)^2\ .$$

Combining with $\phi_i(0)=\phi_i'(0)=0$, we obtain from a Taylor expansion that for all $s>0$, $\phi_i(s)\leqslant s^2(b_i-a_i)^2/8$. Recalling that $\mathbb{E}\left[\mathrm{e}^{sX_i}\right]=\mathrm{e}^{\phi_i(s)}$ and plugging this inequality into (3.10) yields for all $s,t>0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leqslant \mathrm{e}^{-st}\mathrm{e}^{s^2\sum_{i=1}^n \frac{(b_i-a_i)^2}{8}}\ .$$

The proof of (3.9) then follows by taking:

$$s=\frac{t}{\sum_{i=1}^n (b_i-a_i)^2/4}\ ,$$

and straightforward algebra. ∎

Corollary 3.7. Let $(X_i)_{1\leqslant i\leqslant n}$ be iid random variables, such that

$$a \leqslant X_1 \leqslant b \quad a.s.$$

Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right| > \varepsilon\right) \leqslant 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

PROOF. Define $D = \left\{\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_1]\right| > \varepsilon\right\} = \left\{\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| > n\varepsilon\right\}$. Now, applying the Hoeffding inequality (3.8) with $t = n\varepsilon$, $a_i = a$ and $b_i = b$, we then obtain

$$\mathbb{P}(D) \leqslant 2\exp\left(\frac{-2(n\varepsilon)^2}{n(b-a)^2}\right) \leqslant 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right),$$

and the proof is completed. ■

3.2 Asymptotic confidence regions

So far, the sample size n being fixed, we defined the level of confidence regions constructed from the observations $Y_1, \ldots, Y_n$. Now, we will relax the definition of the level by introducing the notion of "asymptotic level". In this context, we still consider confidence regions as functions of $Y_1, \ldots, Y_n$, but the constraint on the level will not be met for a fixed n, but rather asymptotically as n goes to infinity.

To be specific, let's consider a function of interest $g(\theta) \in \mathbb{R}^p$. To avoid confusion, we will slightly redefine notation. Assume that we are given a sequence of observations $Y_1, Y_2, \ldots$, where each observation Y_i belongs to a measurable set $(\mathbb{Y}_1, \mathcal{F}_1)$. Then, $Y_{1:n} = (Y_1, \ldots, Y_n) \in \mathbb{Y}_1^n$ represents the first n data points.

In order to approximate $g(\theta)$, we wish to provide a sequence of *regions* $\mathcal{C}_n(Y_{1:n}) \subset \mathbb{R}^p$ whose relevance is evaluated asymptotically. This allows us to define *asymptotic confidence regions* (when the dimension is $p > 1$) or *asymptotic confidence intervals* (when the dimension is $p = 1$).

3.2.1 Asymptotic level

Definition 3.8 (Asymptotic confidence regions). Let $(\mathbb{Y}_1^{\mathbb{N}}, \mathcal{F}_1^{\otimes\mathbb{N}}, \mathcal{Q})$ be a parametric statistical model i.e. $\mathcal{Q} = (\mathbb{P}_\theta)_{\theta\in\Theta}$.

A sequence of *confidence regions* $\{\mathcal{C}_n : n \in \mathbb{N}\}$ for $g(\theta)$ is with asymptotic level $1-\alpha$ if and only if

$$\liminf_{n\to\infty} \mathbb{P}_\theta(g(\theta) \in \mathcal{C}_n(Y_{1:n})) \geqslant 1-\alpha, \quad \forall\theta \in \Theta,$$

or equivalently

$$\limsup_{n\to\infty} \mathbb{P}_\theta(g(\theta) \notin \mathcal{C}_n(Y_{1:n})) \leqslant \alpha, \quad \forall\theta \in \Theta.$$

As in the previous section, we implicitly assume that for any $n \in \mathbb{N}$ and any $\theta \in \Theta$,

$$\{y_{1:n} \in \mathbb{Y}_1^n : g(\theta) \in \mathcal{C}_n(y_{1:n})\} \in \mathcal{F}_1^{\otimes n},$$

so that every probability in Definition 3.8 is well-defined.

Similarly to the non-asymptotic case, constructing confidence regions with a given asymptotic level will crucially depends on the existence of asymptotically pivotal functions as defined below.

Definition 3.9 (Asymptotically pivotal functions). Let $(\mathbb{Y}_1^{\mathbb{N}}, \mathcal{F}_1^{\otimes\mathbb{N}}, Q)$ be a parametric statistical model i.e. $Q = (\mathbb{P}_\theta)_{\theta\in\Theta}$.

A sequence of measurable functions $G_n : \mathbb{Y}_1^n \times \theta \to \mathbb{R}^p$ is *asymptotically pivotal* if and only if for any measurable set A,

$$\lim_{n\to\infty} \mathbb{P}_\theta(G_n(Y_{1:n},\theta) \in A) \text{ exists and does not depend on } \theta.$$

We now reconsider the poll example and provide several confidence intervals with a given asymptotic level, these confidence intervals are being constructed from asymptotically pivotal functions.

Example 3.10 (A poll example, revisited)**.** In this example, (Y_i) are iid random variables where Y_i stands for the intention of voting for a candidate X for the individual i. In that case, $Y_i \sim \mathcal{B}(\theta)$ under $\mathbb{P}_\theta$ and we observe $Y_{1:n} = (Y_1, \dots, Y_n)$ where $n \in \mathbb{N}$. Here, we are searching for a sequence of confidence intervals $I_{n,\alpha}$ for θ of asymptotic level $1-\alpha$ i.e. such that

$$\limsup_{n\to\infty} \mathbb{P}_\theta(\theta \notin I_{n,\alpha}) \leqslant \alpha\,.$$

First method (Wilson's interval)

Write $\hat{\theta}_n = \sum_{i=1}^n Y_i/n$ and note that $\mathbb{E}_\theta[Y_1] = \theta$ and $\mathbb{V}\mathrm{ar}_\theta(Y_1) = \theta(1-\theta)$. Then, the central limit theorem applies and we have

$$Z_n := \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\theta(1-\theta)}} \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} Z \quad \text{where} \quad Z \sim \mathcal{N}(0,1)\,.$$

In other words, Z_n, under $\mathbb{P}_\theta$, converges in law to Z which has the distribution $\mathcal{N}(0,1)$. Since this distribution does not depend on θ, we deduce that Z_n is *asymptotically pivotal*.

Now, for any $\alpha > 0$, pick z_α s.t. $\mathbb{P}(Z \notin [-z_\alpha, z_\alpha]) = \alpha$. Then,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(Z_n \notin [-z_\alpha, z_\alpha]\right) = \mathbb{P}_\theta\left(Z \notin [-z_\alpha, z_\alpha]\right) = \alpha\,.$$

But, by straightforward algebra, we obtain $\{Z_n \notin [-z_\alpha, z_\alpha]\} = \{\theta \notin I_{n,\alpha}\}$ if and only if $I_{n,\alpha}$ writes

$$I_{n,\alpha} = \left[\frac{\hat{\theta}_n + \frac{z_\alpha^2}{2n} \pm z_\alpha\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n} + \frac{z_\alpha^2}{4n^2}}}{1 + z_\alpha^2/n}\right]\,. \tag{3.11}$$

Finally,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\theta \notin I_{n,\alpha}\right) = \lim_{n\to\infty} \mathbb{P}_\theta\left(Z_n \notin [-z_\alpha, z_\alpha]\right) = \alpha\,,$$

and we conclude that $\{I_{n,\alpha} \,:\, n \in \mathbb{N}\}$ is a sequence of confidence intervals for θ with asymptotic level $1-\alpha$.

Second method (Wald's interval)

Recall that $\hat{\theta}_n = \sum_{i=1}^n Y_i/n$. Slutsky's theorem combined with the central limit theorem yields

$$\tilde{Z}_n := \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} Z \quad \text{where} \quad Z \sim \mathcal{N}(0,1)\,. \tag{3.12}$$

The detailed arguments for proving this convergence in law will be explained in Section 3.2.2. For now, let's assume that we have (3.12) and we will see how to deduce a confidence interval with an asymptotic level of $1-\alpha$. Since the distribution of Z does not depend on θ, we deduce that $\tilde{Z}_n$ is *asymptotically pivotal*.

Now, for any $\alpha > 0$, pick z_α s.t. $\mathbb{P}(Z \notin [-z_\alpha, z_\alpha]) = \alpha$. Then,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\tilde{Z}_n \notin [-z_\alpha, z_\alpha]\right) = \mathbb{P}(Z \notin [-z_\alpha, z_\alpha]) = \alpha\,.$$

Moreover, $\{\tilde{Z}_n \notin [-z_\alpha, z_\alpha]\} = \{\theta \notin \tilde{I}_{n,\alpha}\}$ if and only if $\tilde{I}_{n,\alpha}$ writes

$$\tilde{I}_{n,\alpha} = \left[\hat{\theta}_n \pm z_\alpha \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}\right] . \tag{3.13}$$

Finally,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\theta \notin \tilde{I}_{n,\alpha}\right) = \lim_{n\to\infty} \mathbb{P}_\theta\left(\tilde{Z}_n \notin [-z_\alpha, z_\alpha]\right) = \alpha\,.$$

Hence, $\{\tilde{I}_{n,\alpha} : n \in \mathbb{N}\}$ is a sequence of confidence intervals for θ with asymptotic level $1-\alpha$. When comparing $\tilde{I}_{n,\alpha}$ with $I_{n,\alpha}$, we note that

$$\text{length}(I_{n,\alpha}) \approx \text{length}(\tilde{I}_{n,\alpha}) = 2z_\alpha \sqrt{\frac{\theta(1-\theta)}{n}} + o(1/\sqrt{n})\,.$$

To summarize, $I_{n,\alpha}$ and $\tilde{I}_{n,\alpha}$ have the same asymptotic level of $1-\alpha$, and their lengths are equivalent up to the first order in $1/\sqrt{n}$. Therefore, their properties are quite similar, but since the bounds in $\tilde{I}_{n,\alpha}$ are simpler, it is more commonly used in practice. Note that the expression of $\tilde{I}_{n,\alpha}$ is straightforward because the dependence of the asymptotic pivotal function $\tilde{Z}_n$ on θ is very simple. This simplicity is due to Slutsky's theorem, which allows us to replace $\theta(1-\theta)$ in the denominator of Z_n with its consistent estimator $\hat{\theta}_n(1-\hat{\theta}_n)$. Because of its significance in the design of confidence regions, we will now provide a section with some useful tools for obtaining *asymptotically pivotal functions*.

3.2.2 Tools: the Slutsky theorem and the δ-method

We provide more details for proving (3.12). We have already seen the Slutsky theorem in Theorem 1.9 but we recall it here for ease of reading:

Theorem 3.11 (The Slutsky theorem). Assume that $X_n \overset{\mathbb{P}-prob}{\longrightarrow} c$ where c is a constant and if $Z_n \overset{\mathcal{L}_\mathbb{P}}{\rightsquigarrow} Z$, then $(X_n, Z_n) \overset{\mathcal{L}_\mathbb{P}}{\rightsquigarrow} (c, Z)$, that is, for any real-valued continuous function f, we have $f(X_n, Z_n) \overset{\mathcal{L}_\mathbb{P}}{\rightsquigarrow} f(c, Z)$.

In the poll example, write

$$\tilde{Z}_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} = \underbrace{\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}}_{X_n} \times \underbrace{\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\theta(1-\theta)}}}_{Z_n} .$$

By the central limit theorem, $Z_n \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} Z$ where $Z \sim \mathcal{N}(0,1)$. Recall that strong law of large numbers yields

$$\hat{\theta}_n = \sum_{i=1}^n Y_i/n \to \mathbb{E}_\theta[Y_1] = \theta\,, \quad \mathbb{P}_\theta - a.s.$$

Hence, by Lemma 1.7, $\hat{\theta}_n \overset{\mathbb{P}_\theta - prob}{\longrightarrow} \theta$ and consequently, applying Theorem 1.8, $X_n \overset{\mathbb{P}_\theta - prob}{\longrightarrow} 1$. Since in addition, the function $(x,z) \mapsto f(x,z) = xz$ is continuous, we can apply Slutsky's theorem and deduce that

$$\tilde{Z}_n = f(X_n, Z_n) \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} f(1, Z) = Z\,,$$

which shows (3.12). Another tool (which is actually a byproduct of the Slutsky theorem) for finding asymptotically pivotal functions is the δ-method, and we will now state and prove it.

Lemma 3.12 (The δ-method). Assume that there exist a sequence of random variables $\{\hat{\theta}_n : n \in \mathbb{N}\}$, a random variable U, a constant θ and a sequence of positive real numbers $\{r_n : n \in \mathbb{N}\}$ such that

$$\lim_n r_n = \infty, \quad \text{and} \quad r_n(\hat{\theta}_n - \theta) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} U .$$

Then, for any measurable function $g : \mathbb{R} \to \mathbb{R}$, differentiable at θ, we have

$$r_n(g(\hat{\theta}_n) - g(\theta)) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} g'(\theta)U .$$

PROOF. Define

$$G(x) = \begin{cases} \frac{g(x)-g(\theta)}{x-\theta} & \text{if } x \neq \theta \\ g'(\theta) & \text{otherwise} \end{cases}$$

Write

$$r_n(g(\hat{\theta}_n) - g(\theta)) = G(\hat{\theta}_n) \times \underbrace{r_n\left(\hat{\theta}_n - \theta\right)}_{U_n} .$$

By assumption, $U_n \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} U$. Moreover, $\hat{\theta}_n \overset{\mathbb{P}-prob}{\longrightarrow} \theta$ (see Example 1.10). Moreover, g being differentiable at θ, we deduce that G is continuous and hence, by Theorem 1.8, $G(\hat{\theta}_n) \overset{\mathbb{P}-prob}{\longrightarrow} G(\theta) = g'(\theta)$. Then, applying the Slutsky lemma to the continuous function $f(x,u) = xu$, we get

$$r_n(g(\hat{\theta}_n) - g(\theta)) = G(\hat{\theta}_n)U_n = f(G(\hat{\theta}_n), U_n) \overset{\mathcal{L}_{\mathbb{P}}}{\rightsquigarrow} f(g'(\theta), U) = g'(\theta)U ,$$

which concludes the proof. ■

►**Q-3.3.** You said that the δ-method can be used for getting asymptotically pivotal functions. Can you tell me more? Can you explain the δ-method approach for the poll example?

Ok, let us do that. Recalling that $\hat{\theta}_n = \sum_{i=1}^n Y_i/n$, the central limit theorem yields

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} U , \quad \text{where} \quad U \sim \mathcal{N}(0, \theta(1-\theta)) .$$

Applying the δ-method, we obtain, for any measurable function g that is differentiable at θ,

$$Z_n = \sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} g'(\theta)U \sim \mathcal{N}\left(0, \sigma^2(\theta)\right) , \quad \text{where} \quad \sigma^2(\theta) = \theta(1-\theta)g'^2(\theta) . \tag{3.14}$$

Now, choose g such that $\sigma^2(\theta) = 1$. Then, for this choice of g,

$$Z_n \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} \mathcal{N}(0,1) .$$

Hence, $\{Z_n : n \in \mathbb{N}\}$ is *asymptotically pivotal*, from which we can easily deduce a *confidence interval* for θ. Since the choice of g allows to obtain a variance $\sigma^2 = 1$, we call it *a variance-stabilizing method*.

►**Q-3.4.** Easily ? Can you be more explicit?

Your wishes are my commands. By definition of σ^2, the condition $\sigma^2(\theta) = 1$ means

$$g'(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} .$$

You can check that the function $g : (0,1) \to \mathbb{R}$ defined by $g(\theta) = 2\arcsin(\sqrt{\theta})$ satisfies the above identity. With this choice of g, (3.14) becomes

$$Z_n = \sqrt{n}\left(2\arcsin\left((\hat{\theta}_n^{\text{ML}})^{1/2}\right) - 2\arcsin\left(\theta^{1/2}\right)\right) \overset{\mathcal{L}_{\mathbb{P}_\theta}}{\rightsquigarrow} \mathcal{N}(0,1) ,$$

And since Z_n is asymptotically pivotal, you can obtain a confidence interval for θ using similar techniques as in Example 3.10. I'll leave the rest to you to ensure that you have mastered the course.

►**Q-3.5.** Oh... thanks so much. I will do that immediately. Statistics is so exciting !!!

3.3 After studying this chapter...

a) I know the definition of the level of a confidence region and I can distinguish between level and asymptotic level.

b) I understand the definition of pivotal functions or asymptotically pivotal functions and how they are related to confidence regions.

c) I can construct confidence regions with a given level, using the Bienaymé-Tchebychev inequality or the Hoeffding inequality.

d) I can use Slutsky's theorem or the δ-method to obtain asymptotically pivotal functions.

3.4 Highlights

Abraham Wald (source: Wikipedia)

Abraham Wald (October 1902 – 13 December 1950) was a Jewish Hungarian mathematician who contributed to decision theory, geometry and econometrics, and founded the field of sequential analysis. One of his well-known statistical works was written during World War II on how to minimize the damage to bomber aircraft and took into account the survivorship bias in his calculations. He spent his research career at Columbia University.

Wald was born on 31 October 1902 in Kolozsvár, Transylvania, in the Kingdom of Hungary. A religious Jew, he did not attend school on Saturdays, as was then required by the Hungarian school system, and so he was home-schooled by his parents until college. His parents were quite knowledgeable and competent as teachers.

In 1928, he graduated in mathematics from the King Ferdinand I University.In 1927, he had entered graduate school at the University of Vienna, from which he graduated in 1931 with a Ph.D. in mathematics. His advisor there was Karl Menger.

Despite Wald's brilliance, he could not obtain a university position because of Austrian discrimination against Jews. However, Oskar Morgenstern created a position for Wald in economics. When Nazi Germany annexed Austria in 1938, the discrimination against Jews intensified. In particular, Wald and his family were persecuted as Jews. Wald immigrated to the United States at the invitation of the Cowles Commission for Research in Economics, to work on econometrics research.

The damaged portions of returning planes show locations where they can sustain damage and still return home; those hit in other places presumably do not survive. During World War II, Wald was a member of the Statistical Research Group (SRG) at Columbia University, where he applied his statistical skills to various wartime problems. They included methods of sequential analysis and sampling inspection. One of the problems that the SRG worked on was to examine the distribution of damage to aircraft returning after flying missions to provide advice on how to minimize bomber losses to enemy fire. Wald derived a useful means of estimating the damage distribution for all aircraft that flew from the data on the damage distribution of all aircraft that returned. His work is considered seminal in the discipline of operational research, which was then fledgling.

Wald and his wife died in 1950 when the Air India plane (VT-CFK, a DC-3 aircraft) in which they were traveling crashed near the Rangaswamy Pillar in the northern part of the Nilgiri Mountains, in southern India, on an extensive lecture tour at the invitation of the Indian government. He had visited the Indian Statistical Institute at Calcutta and was to attend the Indian Science Congress at Bangalore in January. Their two children were back at home in the United States.

After his death, Wald was criticized by Sir Ronald A. Fisher FRS. Fisher attacked Wald for being a mathematician without scientific experience who had written an incompetent book on statistics. Fisher particularly criticized Wald's work on the design of experiments and alleged ignorance of the basic ideas

of the subject, as set out by Fisher and Frank Yates. Wald's work was defended by Jerzy Neyman the next year. Neyman explained Wald's work, particularly with respect to the design of experiments. Lucien Le Cam credits him in his own book, Asymptotic Methods in Statistical Decision Theory: "The ideas and techniques used reflect first and foremost the influence of Abraham Wald's writings."

He was the father of the noted American physicist Robert Wald.

Chapter 4

Statistical Tests

4.1 Terminology and principles of statistical tests

In what follows,

- $\mathsf{Y} = (Y_1, \dots, Y_n)$ are the observations.
- $\mathsf{Y} \sim \mathbb{P}_\star$ where $\mathbb{P}_\star$ belongs to a family of possible distributions $\mathcal{Q}$.

Definition 4.1 (**Statistical hypothesis test).**

- We split $\mathcal{Q}$ into two disjoint subsets $\mathcal{Q}_0$ and $\mathcal{Q}_1$.
- Based on Y, we decide between two hypothesis:

$$\boxed{H_0 : \mathbb{P}_\star \in \mathcal{Q}_0 \quad \text{versus} \quad H_1 : \mathbb{P}_\star \in \mathcal{Q}_1}$$

where H_0 is called the *null hypothesis* and H_1 the *alternative hypothesis*.

Definition 4.2 (**Simple hypothesis).** An hypothesis H_0 (resp. H_1) is called simple if and only if

$$\mathcal{Q}_0 \text{ (resp. } \mathcal{Q}_1\text{) contains exactly one element.}$$

When the data $\mathsf{Y} = (Y_1, \dots, Y_n)$ is observed, the statistician chooses between the two hypothesis according to $T(\mathsf{Y}) \in \{0, 1\}$ where

- 0 is the decision of accepting H_0
- 1 is the decision of rejecting H_0

Note in practice, when $T(\mathsf{Y}) = 1$, we often say that we reject H_0 instead of saying that we accept H_1. This is because we often choose for H_0 the hypothesis which seems the most commonly accepted by now or the most simple hypothesis.

Remark 4.3. A statistical test is therefore defined by the function T which is called the *statistic* of the test. Note that $T(\mathsf{Y}) = \mathbf{1}_W(\mathsf{Y})$ for some region W, which is called the *rejection region* or *critical region*. Thus,

$$Y \in W \quad \Longleftrightarrow \quad T(\mathsf{Y}) = 1 \, .$$

Definition 4.4 (Type-I and Type-II errors)**.** Two types of errors are associated to a statistical test $T(\mathsf{Y})$:

(i) rejecting H_0 when H_0 is true (Type-I error).

(ii) accepting H_0 when H_0 is false (Type-II error).

Remark 4.5. These two errors are respectively associated to the probabilities:

$$\begin{aligned} \text{Type-I error:} \quad & P(T(\mathsf{Y})=1) \quad \text{with} \quad P \in Q_0\,, \\ \text{Type-II error:} \quad & P(T(\mathsf{Y})=0) \quad \text{with} \quad P \in Q_1\,. \end{aligned}$$

Definition 4.6.

(i) The power function of $T(\mathsf{Y})$ is defined by

$$P \mapsto \beta_T(P) = P(T(\mathsf{Y})=1), \quad P \in Q\,.$$

(ii) The size of T is by definition $\sup_{P\in Q_0} \beta_T(P)$.

(iii) A test T is of (significance) level α if its size is less than α.

►**Q-4.1.** So, if I understand correctly, the size and level are linked with Type I-errors. You never consider Type II-errors?

Thanks for this excellent question, as it smoothly transitions to the notion of uniformly most powerful tests where Type-I and Type-II errors come into play.

Definition 4.7 (UMP). A test T_* of size α is a **uniformly most powerful** (UMP) test if and only if

$$\text{If Type-I error } (T) \leqslant \text{ Type-I error } (T_*) \text{ Then Type-II error } (T) \geqslant \text{ Type-II error } (T_*),$$

or equivalently, for any test T of level α,

$$\beta_{T_*}(P) \geqslant \beta_T(P)\,, \quad \forall P \in Q_1\,.$$

►**Q-4.2.** There always exist UMP tests?

I don't know, but if you find any, I advise you to use it because you can be assured that there is no other test that can achieve both better Type I and Type II errors.

4.2 The Neyman-Pearson lemma

We now consider a dominated parametric statistical model

- $Q = \{P_{\theta_0}, P_{\theta_1}\}$ where $P_{\theta_i}(\mathrm{d}y) = \ell_{\theta_i}(y)\mu(\mathrm{d}y)$ and $\ell_{\theta_i}(\cdot) > 0$.

Consider the following hypothesis

$$\boxed{H_0 :\ \theta = \theta_0 \quad \text{versus} \quad H_1 :\ \theta = \theta_1}$$

Lemma 4.8 (The Neyman-Pearson lemma). Assume that for any $\lambda > 0$, $\mathbb{P}_{\theta_0}\left(\frac{\ell_{\theta_1}(Y)}{\ell_{\theta_0}(Y)} = \lambda\right) = 0$. Then, for any $\alpha \in [0,1]$, there exists c_α such that the test

$$T_*(Y) = \begin{cases} 1 & \text{if } \frac{\ell_{\theta_1}(Y)}{\ell_{\theta_0}(Y)} > c_\alpha \\ 0 & \text{otherwise} \end{cases}$$

is a UMP test of size α.

PROOF. Under the assumptions of the Theorem, the function $\lambda \mapsto \mathbb{P}_{\theta_0}[\ell_{\theta_1}(Y) > \lambda \ell_{\theta_0}(Y)]$ is continuous, is equal to 1 if $\lambda = 0$ and converges to 0 as λ goes to infinity. This implies that there exists a constant c such that

$$\mathbb{P}_{\theta_0}[\ell_{\theta_1}(Y) > c\ell_{\theta_0}(Y)] = \alpha\,.$$

Consider now $T_*(Y) = \mathbf{1}_{\{\ell_{\theta_1}(Y) > c\ell_{\theta_0}(Y)\}}$. The previous equation shows that it is a test of size α. Consider now a test T of level α, it remains to show that $\beta_{T_*}(\theta_1) \geqslant \beta_T(\theta_1)$. It can be easily checked that

$$\forall y, \quad [T_*(y) - T(y)](\ell_{\theta_1}(y) - c\ell_{\theta_0}(y)) \geqslant 0\,,$$

which gives, by integrating with respect to μ,

$$\mathbb{P}_{\theta_1}(T_*(Y) = 1) - \mathbb{P}_{\theta_1}(T(Y) = 1) - c\left[\mathbb{P}_{\theta_0}(T_*(Y) = 1) - \mathbb{P}_{\theta_0}(T(Y) = 1)\right] \geqslant 0\,.$$

This is equivalent to:

$$\beta_{T_*}(\theta_1) - \beta_T(\theta_1) \geqslant c(\alpha - \beta_T(\theta_0)) \geqslant 0\,.$$

And the proof is concluded. ■

►**Q-4.3.** Great! So it's an example where you can be sure to find a UMP test.

You are right, but you must pay attention to the assumptions. The null and alternative hypotheses must be simple. Anyway, this test is based on a critical region which can be expressed in terms of the likelihood ratio $\ell_{\theta_1}(Y)/\ell_{\theta_0}(Y)$. That's why it is sometimes called a likelihood ratio test, which is a type of test often used in practice.

Sufficient statistics

Lemma 4.9. Assume that there is a sufficient statistic S for θ, i.e., the likelihood writes

$$\ell_\theta(y) = \psi_\theta(S(y))\phi(y)\,.$$

Then the critical region only depends on S.

PROOF. According to the Neyman-Pearson lemma,

$$T_*(Y) = \begin{cases} 1 & \text{if } \frac{\ell_{\theta_1}(Y)}{\ell_{\theta_0}(Y)} = \frac{\psi_{\theta_1}(S(Y))\phi(Y)}{\psi_{\theta_0}(S(Y))\phi(Y)} = \frac{\psi_{\theta_1}(S(Y))}{\psi_{\theta_0}(S(Y))} > c_\alpha \\ 0 & \text{otherwise} \end{cases}$$

And the critical region only depends on S which concludes the proof.

■

4.3 Some classical parametric tests

We now describe some usual parametric tests of **size** α. Some of these statistical tests are UMP but most of the time, the only requirement concerns the Type-I error: the test is of size α.

(i) For a single set of normal observations:

- testing the mean.
- testing the variance.

(ii) For two sets of normal observations:

- testing if their variance coincide.
- if so, testing if their mean coincide.

4.3.1 Testing the mean of the distribution $\mathcal{N}(m,\sigma^2)$

Mean of the distribution $\mathcal{N}(m,\sigma^2)$ when σ^2 is known

Consider $(Y_i)_{1\leqslant i\leqslant n}$ iid and $Y_i \sim \mathcal{N}(m,\sigma^2)$. Write $\bar{Y}_n = \frac{\sum_{i=1}^n Y_i}{n}$.

$$\boxed{H_0 : m = m_0 \quad \text{versus} \quad H_1 : m = m_1} \quad (\text{with } m_1 > m_0)\,.$$

Let

- $S = \frac{\bar{Y}_n - m_0}{\sqrt{\frac{\sigma^2}{n}}}$,
- c_α be such that $\mathbb{P}(Z > c_\alpha) = \alpha$ with $Z \sim \mathcal{N}(0,1)$.

Lemma 4.10. The test defined by the critical region:

$$\begin{cases} S \leqslant c_\alpha & \to H_0 \text{ is accepted} \\ S > c_\alpha & \to H_0 \text{ is rejected} \end{cases}$$

is an UMP test of size α. Moreover, under H_0, $S \sim \mathcal{N}(0,1)$.

This is called a *one-sided test* (or *one-tailed test*), since the decision depends on whether the test statistic is larger or less than a given threshold. The proof follows from the Neyman-Pearson lemma. Note that under H_0, S is distributed according to $\mathcal{N}(0,1)$. Hence, $\mathbb{P}_{H_0}(S > c_\alpha) = \mathbb{P}(Z > c_\alpha) = \alpha$. Conversely, under H_1, S is distributed according to $\mathcal{N}(m_1 - m_0, 1)$ with $m_1 > m_0$ by assumption. In Fig. 4.1, we summarize how the statistical test works with a graphic interpretation of Type-I and Type-II errors.

►**Q-4.4.** If I understand correctly, the result of the test is very linked to the size α that you choose a priori. Is there any indicator from the observations that allows you to say that the hypothesis H_0 is more likely to be accepted (or rejected) without choosing beforehand the size of the test?

The p-value, denoted by p_{val} in this course, may answer to your question. It allows to quantify the statistical significance of the observed statistic under the null hypothesis. Mathematically speaking, for this one-sided test, it is defined by: $p_{\text{val}} = F(S)$ where $F(t) = \mathbb{P}_{H_0}(S > t)$. It is a random variable between 0 and 1 which satisfies: $\{\alpha \leqslant p_{\text{val}}\} \Leftrightarrow \{H_0 \text{ accepted}\} \Leftrightarrow \{S \leqslant c_\alpha\}$ which can be seen in Fig. 4.2 and Fig. 4.3. Usually, if the p-value falls below 0.05, the null hypothesis is rejected.

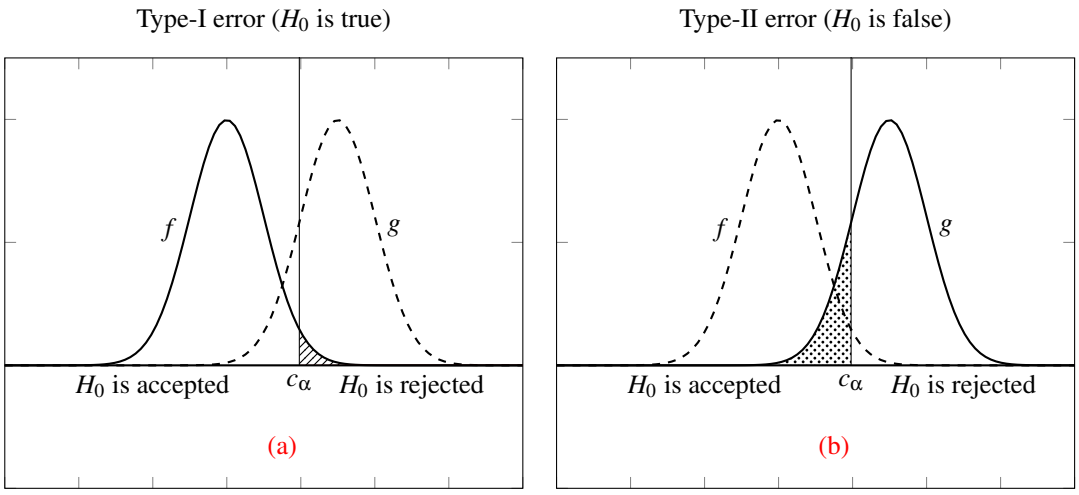


Figure 4.1: The test statistic S takes values on the x-axis. If $S \leqslant c_\alpha$, H_0 is accepted, and if $S > c_\alpha$, H_0 is rejected. Let f be the density of $\mathcal{N}(0,1)$ and g be the density of $\mathcal{N}(m_2 - m_1, 1)$. Then, under H_0 and H_1 respectively, the test statistic S has densities f and g. In the left panel, the Type-I error corresponds to the area of the hashed region and is equal to α. In the right panel, the Type-II error corresponds to the area of the dotted region.

Graphical interpretation of the p-value for one-sided tests

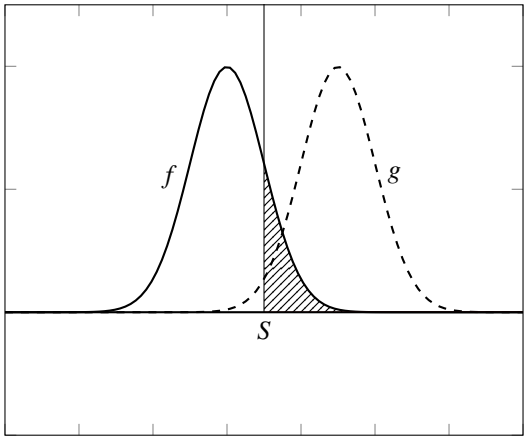


Figure 4.2: The p-value is defined as the area of the hashed region. Even though it corresponds to the situation of Fig. 4.1-(a), where c_α is replaced by S, you must pay attention that S is a random variable. In this example, we can see that the hashed area here is larger than the hashed area in Fig. 4.1-(a), showing that the $p_{\text{val}} \geqslant \alpha$, whereas $S \leqslant c_\alpha$.

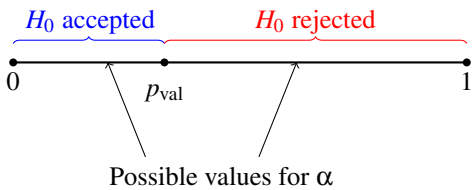


Figure 4.3: The size α of the test may take values between 0 and 1 on the x-axis.

Mean of the distribution $\mathcal{N}(m, \sigma^2)$ when σ^2 is unknown

Consider (Y_i) iid and $Y_i \sim \mathcal{N}(m, \sigma^2)$. Write $\bar{Y}_n = \frac{\sum_{i=1}^n Y_i}{n}$.

$$\boxed{H_0 : m = m_0 \quad \text{versus} \quad H_1 : m \neq m_0}$$

Let

- $T = \frac{\bar{Y}_n - m_0}{\sqrt{\frac{\tilde{\sigma}^2}{n}}}$ where $\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$,
- c_α be such that $\mathbb{P}(|Z| > c_\alpha) = \alpha$ where $Z \sim \mathcal{T}(n-1)$ i.e. Z follows the t-distribution with $n-1$ degrees of freedom.

Lemma 4.11. The test defined by the critical region:

$$\begin{cases} |T| \leqslant c_\alpha & \to H_0 \text{ is accepted} \\ |T| > c_\alpha & \to H_0 \text{ is rejected} \end{cases}$$

is of level α. Moreover, under H_0, $T \sim \mathcal{T}(n-1)$.

We call it a *two-sided test* (or *two-tailed test*) because the rejection region corresponds to $T > c_\alpha$ or $T < -c_\alpha$. The p-value is then defined in Fig. 4.4.

Graphical interpretation of the p-value for two-sided tests

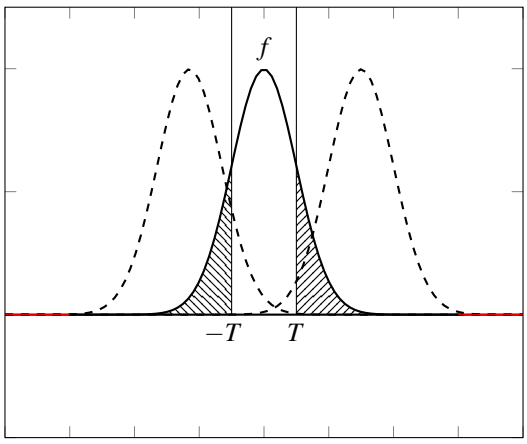


Figure 4.4: Let f be the density of $\mathcal{T}(n-1)$. Under H_0, the test statistic T has density f. The p-value of this two-sided test is then defined by the area of the hashed region. It corresponds to the area associated to the critical region where the threshold c_α is replaced by the observed statistic T. The dotted lines represent the possible densities for T under H_1.

4.3.2 Testing the variance of the distribution $\mathcal{N}(m, \sigma^2)$

Variance of the distribution $\mathcal{N}(m, \sigma^2)$ when m is known

Consider (Y_i) iid and $Y_i \sim \mathcal{N}(m, \sigma^2)$.

$$\boxed{H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1 : \sigma^2 = \sigma_1^2} \quad (\text{with } \sigma_1^2 > \sigma_0^2)\,.$$

Let

- $D = \sum_{i=1}^n \frac{(Y_i - m)^2}{\sigma_0^2}$.

- c_α be such that $\mathbb{P}(Z > c_\alpha) = \alpha$ where $Z \sim \chi_2(n)$ i.e. Z follows the χ_2 distribution with n degrees of freedom.

Lemma 4.12. The test defined by the critical region

$$\begin{cases} D \leqslant c_\alpha & \to H_0 \text{ is accepted} , \\ D > c_\alpha & \to H_0 \text{ is rejected} . \end{cases}$$

is of level α. Moreover, under H_0, $D \sim \chi_2(n)$.

Variance of the distribution $\mathcal{N}(m, \sigma^2)$ when m is unknown

Consider (Y_i) iid and $Y_i \sim \mathcal{N}(m, \sigma^2)$.

$$\boxed{H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1 : \sigma^2 = \sigma_1^2} \quad (\text{with } \sigma_1^2 > \sigma_0^2).$$

Let

- $D = \sum_{i=1}^n \frac{(Y_i - \bar{Y}_n)^2}{\sigma_0^2}$.
- c_α be such that $\mathbb{P}(Z > c_\alpha) = \alpha$ with $Z \sim \chi_2(n-1)$.

Lemma 4.13. The test defined by the critical region

$$\begin{cases} D \leqslant c_\alpha & \to H_0 \text{ is accepted} , \\ D > c_\alpha & \to H_0 \text{ is rejected} . \end{cases}$$

is of level α. Moreover, under H_0, $D \sim \chi_2(n-1)$.

4.3.3 Comparing two normal distributions $\mathcal{N}(m_0, \sigma_0^2)$ and $\mathcal{N}(m_1, \sigma_1^2)$

Assume that the observations are divided into two independent subsets $(Y_{0,1}, \ldots, Y_{0,n_0})$ and $(Y_{1,1}, \ldots, Y_{1,n_1})$ and assume that

$$\begin{cases} (Y_{0,i})_{1 \leqslant i \leqslant n_0} & \text{are iid and} \quad Y_{0,i} \sim \mathcal{N}(m_0, \sigma_0^2) , \\ (Y_{1,i})_{1 \leqslant i \leqslant n_1} & \text{are iid and} \quad Y_{1,i} \sim \mathcal{N}(m_1, \sigma_1^2) . \end{cases}$$

Comparing two normal distributions. To decide whether these two normal distributions are the same, we usually use a two-level test:

- We start by using an hypothesis test for deciding whether σ_0^2 and σ_1^2 are equal or not.
- Assume that the first test concludes that $\sigma_0^2 = \sigma_1^2$. Then we use a second hypothesis test to decide whether $m_0 = m_1$ or not.

Testing the equality of the variance

$$\boxed{H_0 : \sigma_0^2 = \sigma_1^2 \quad \text{versus} \quad H_1 : \sigma_0^2 \neq \sigma_1^2}$$

Write

$$\bar{Y}_{0,n_0} = \sum_{i=1}^{n_0} Y_{0,i}/n_0 \,, \quad \bar{Y}_{1,n_1} = \sum_{i=1}^{n_1} Y_{1,i}/n_1 \,.$$

- Define $R = \frac{\sum_{i=1}^{n_0}(Y_{0,i}-\bar{Y}_{0,n_0})^2/(n_0-1)}{\sum_{i=1}^{n_1}(Y_{1,i}-\bar{Y}_{1,n_1})^2/(n_1-1)}$.
- Letting $Z \sim F(n_0-1, n_1-1)$, define c_α and d_α with

$$\mathbb{P}(Z \leqslant c_\alpha) = \alpha/2, \qquad \mathbb{P}(Z \geqslant d_\alpha) = \alpha/2 \,.$$

Lemma 4.14. The test defined by the critical region

$$\begin{cases} R \in [c_\alpha, d_\alpha] & \to H_0 \text{ is accepted} \,, \\ R \notin [c_\alpha, d_\alpha] & \to H_0 \text{ is rejected} \,. \end{cases}$$

is of level α. Moreover, under H_0, $R \sim F(n_0-1, n_1-1)$.

Testing the equality of the mean when the variances are equal

$$\boxed{H_0 : m_0 = m_1 \quad \text{versus} \quad H_1 : m_0 \neq m_1}$$

Write

- $S = \dfrac{\bar{Y}_{0,n_0} - \bar{Y}_{1,n_1}}{\sqrt{\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\frac{1}{n_0+n_1-2}\left(\sum_{i=1}^{n_0}(Y_{0,i}-\bar{Y}_{0,n_0})^2 + \sum_{i=1}^{n_1}(Y_{1,i}-\bar{Y}_{1,n_1})^2\right)}}$
- c_α is defined by $\mathbb{P}(|Z| > c_\alpha) = \alpha$ where $Z \sim \mathcal{T}(n_0+n_1-2)$.

Lemma 4.15. The test defined by the critical region

$$\begin{cases} |S| \leqslant c_\alpha & \to H_0 \text{ is accepted} \,, \\ |S| > c_\alpha & \to H_0 \text{ is rejected} \,. \end{cases}$$

is of level α. Moreover, under H_0, $S \sim \mathcal{T}(n_0+n_1-2)$.

4.4 After studying this chapter...

a) I understand Statistical Hypothesis Testing,

b) I can distinguish Type-I and Type-II errors,

c) I know the definitions of the size, level, power of the test, UMP tests,

d) From the Neyman-Pearson Theorem, I can understand the likelihood ratio test.

e) For normal samples, I can choose between all the different classical hypothesis tests.

4.5 Highlights

4.5.1 Jersy Neyman (source: Wikipedia)

Jerzy Neyman (April 16, 1894 – August 5, 1981; born Jerzy Spława-Neyman) was a Polish mathematician and statistician who spent the first part of his professional career at various institutions in Warsaw, Poland and then at University College London, and the second part at the University of California, Berkeley. Neyman first introduced the modern concept of a confidence interval into statistical hypothesis testing and co-revised Ronald Fisher's null hypothesis testing (in collaboration with Egon Pearson).

He was born into a Polish family in Bendery, in the Bessarabia Governorate of the Russian Empire, the fourth of four children of Czesław Spława-Neyman and Kazimiera Lutosławska. His family was Roman Catholic, and Neyman served as an altar boy during his early childhood. Later, Neyman would become an agnostic. Neyman's family descended from a long line of Polish nobles and military heroes. He graduated from the Kamieniec Podolski gubernial gymnasium for boys in 1909 under the name Yuri Cheslavovich Neyman. He began studies at Kharkiv University in 1912, where he was taught by Ukrainian probabilist Sergei Natanovich Bernstein. After he read 'Lessons on the integration and the research of the primitive functions' by Henri Lebesgue, he was fascinated with measure and integration.

In 1921, he returned to Poland in a program of repatriation of POWs after the Polish-Soviet War. He earned his Doctor of Philosophy degree at University of Warsaw in 1924 for a dissertation titled "On the Applications of the Theory of Probability to Agricultural Experiments". He was examined by Wacław Sierpiński and Stefan Mazurkiewicz, among others. He spent a couple of years in London and Paris on a fellowship to study statistics with Karl Pearson and Émile Borel. After his return to Poland, he established the Biometric Laboratory at the Nencki Institute of Experimental Biology in Warsaw.

He published many books dealing with experiments and statistics, and devised the way which the FDA tests medicines today.

Neyman proposed and studied randomized experiments in 1923. Furthermore, his paper "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection", given at the Royal Statistical Society on 19 June 1934, was the groundbreaking event leading to modern scientific sampling. He introduced the confidence interval in his paper in 1937. Another noted contribution is the Neyman–Pearson lemma, the basis of hypothesis testing.

He was an Invited Speaker of the ICM in 1928 in Bologna and a Plenary Speaker of the ICM in 1954 in Amsterdam.

In 1938, he moved to Berkeley, where he worked for the rest of his life. Thirty-nine students received their Ph.Ds under his advisorship. In 1966, he was awarded the Guy Medal of the Royal Statistical Society and three years later the U.S. National Medal of Science. He died in Oakland, California in 1981.

Index