

# Statistics and Learning

Lecturers (alphabetic order):

Julien Bect, Gilles Faÿ, Ziad Kobeissi, Laurent Le Brusquet,  
Vincent Lescarret, Arshak Minasyan, Arthur Tenenhaus<sup>†</sup> & Xujia Zhu

<sup>†</sup> Course coordinator

1/60

Lecture 2/9

## Point estimation

### Course objectives

- ▶ Learn how to quantify the performance of an estimator
- ▶ Learn how to compare estimators
- ▶ Introduce the asymptotic approach

2/60

## Lecture outline

- 1 – Point estimation: definition and notations
- 2 – Quadratic risk of an estimator
- 3 – A lower bound on the quadratic risk
- 4 – Asymptotic properties
- 5 – Standard exercises
- 6 – Appendices

3/60

## Lecture outline

- 1 – Point estimation: definition and notations
- 2 – Quadratic risk of an estimator
- 3 – A lower bound on the quadratic risk
- 4 – Asymptotic properties
- 5 – Standard exercises
- 6 – Appendices

## Recap: mathematical framework

### Data

- ▶ Formally, an element  $\underline{x}$  in a set  $\underline{\mathcal{X}}$ .
- ▶ ex:  $\underline{\mathcal{X}} = \mathbb{R}^n, \mathbb{R}^{n \times d}, \{\text{words}\}, \text{some functional space, etc.}$

### From data to random variables

- ▶ **A priori** point of view: before the data is actually collected.
- ▶ Modeling: RV  $\underline{X}$  taking values in  $(\underline{\mathcal{X}}, \underline{\mathcal{A}})$ ,
- ▶ but the **distribution of  $\underline{X}$  is unknown**.

### Statistical modeling

- ▶  $\underline{X}$  is assumed to be defined on  $(\Omega, \mathcal{F}, \mathbb{P})$ , with  $\mathbb{P} \in \mathcal{P}$ .
- ▶  $\mathcal{P}$  : a set of possible probability measures on  $(\Omega, \mathcal{F})$
- ▶ Formally,  $\mathcal{M} = (\underline{\mathcal{X}}, \underline{\mathcal{A}}, \mathcal{P}^{\underline{X}})$ , with  $\mathcal{P}^{\underline{X}} = \{\mathbb{P}^{\underline{X}}, \mathbb{P} \in \mathcal{P}\}$ .

Canonical construction:  $\Omega = \underline{\mathcal{X}}, \mathcal{F} = \underline{\mathcal{A}}, \underline{X} = \text{Id}_{\underline{\mathcal{X}}}$  et  $\mathcal{P} = \mathcal{P}^{\underline{X}}$ .

4/60

## Recap: mathematical framework (cont'd)

### Important

Since  $\mathbb{P} \in \mathcal{P}$  is unknown, we need to design statistical procedures that “work well” (in a sense to be specified) for **any** distribution  $\mathbb{P} \in \mathcal{P}$ .

### Parameterized family of probability distributions

- ▶ Usually, we write  $\mathcal{P} = \{\mathbb{P}_{\theta}, \theta \in \Theta\}$ .
- ▶  $\theta$ : **unknown parameter** (scalar, vector, function...)
- ▶ In the following, we assume a **parametric model**:  $\Theta \subset \mathbb{R}^p$ .

Important case:  $d$ -variate (iid)  $n$ -sample  $(\rightarrow n \times d$  data table)

- ▶  $\underline{\mathcal{X}} = \mathcal{X}^n$ , with  $\mathcal{X} \subset \mathbb{R}^d$ , endowed with their Borel  $\sigma$ -algebras,
- ▶  $\underline{X} = (X_1, \dots, X_n)$  with  $X_i \stackrel{\text{iid}}{\sim} \mathbb{P}_{\theta}$ , and thus  $\mathbb{P}_{\theta}^{\underline{X}} = \mathbb{P}_{\theta}^{\otimes n}$ .

5/60

## Point estimation

Parameter of interest

- ▶ We are interested in **parameter**  $\eta = g(\theta)$ , where  $g : \Theta \mapsto \mathbb{R}$  ou  $\mathbb{R}^q$ .
- ▶ Its value is **unknown**, since  $\theta$  is unknown.

### Informal definition: estimation

Guess (infer) the value of  $\eta$  based on a realization  $\underline{x}$  of  $\underline{X}$ .

### Definition: estimator

We call **estimator** any statistic  $\hat{\eta} = \varphi(\underline{X})$  taking value in the set  $N = g(\Theta)$  of possible values for  $\eta$ .

Remark: the word “estimator” can refer either to the RV  $\hat{\eta}$  or to the function  $\varphi$ . In practice, we identify the two and write (abusively)  $\hat{\eta} = \hat{\eta}(\underline{X})$ .

6/60

## Example 1 (reminder)

IID Gaussian  $n$ -sample:  $\underline{X} = (X_1, \dots, X_n)$  with

- ▶  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ ,
- ▶  $\theta = (\mu, \sigma^2)$ ,
- ▶  $\Theta = \mathbb{R} \times ]0; +\infty[$ .

In this example, we assume that we want to **estimate the mean**  $\mu$ ;

- ▶ here  $\eta = \mu$  and  $g : \theta = (\mu, \sigma^2) \mapsto \mu$ ,
- ▶  $\sigma^2$  is unknown too (nuisance parameter).

7/60

## Example 1 (cont'd)

Some possible estimators. . .

- ▶  $\hat{\mu}_1 = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  (method of moments / MLE),
- ▶  $\hat{\mu}_2 = \mu_0$  for a given  $\mu_0 \in \mathbb{R}$ ,
- ▶  $\hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n$ ,
- ▶  $\hat{\mu}_4 = \bar{X}_n + c$  for a given  $c \neq 0$ ,
- ▶  $\hat{\mu}_5 = \text{med}(X_1, \dots, X_n)$ ,
- ▶ . . .

Questions

- ▶ Is one of these estimators “better” than the others?
- ▶ Can we find an “optimal” estimator?
- ▶ In what sense?

8/60

## Other examples

Example 1'

- ▶ Same statistical model as in Example 1, but
- ▶  $g(\theta) = \sigma^2$ .
- ▶ In this case,  $\mu$  is seen as a nuisance parameter.

Example 1''

- ▶ Again the same statistical model, but
- ▶  $g(\theta) = \theta = (\mu, \sigma^2)$ .
- ▶ Here, the parameter to be estimated is a **vector**.

9/60

## Other examples (cont'd)

### Example 2

- ▶  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$ , i.e.,  $f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{x \geq 0}$ ,
- ▶  $\Theta = (0, +\infty)$ ,
- ▶  $g(\theta) = \mathbb{E}_\theta(X_1) = 1/\theta$ .

### Example 2'

- ▶ Same statistical model, but
- ▶  $g(\theta) = \mathbb{P}_\theta(X_1 > x_0) = e^{-\theta x_0}$  for a given  $x_0 > 0$ .

### Example 3 (optional)

- ▶ non-parametric statistics

 complement

10/60

## Lecture outline

- 1 – Point estimation: definition and notations
- 2 – Quadratic risk of an estimator
- 3 – A lower bound on the quadratic risk
- 4 – Asymptotic properties
- 5 – Standard exercises
- 6 – Appendices

## General concept of risk

### Goal

Quantify the performance of an estimator

Consider a **loss function**  $L : N \times N \rightarrow \mathbb{R}$ .

- ▶ Reminder:  $N = g(\Theta)$  is the set of all possible values for  $\eta$ .
- ▶ Interpretation: we lose  $L(\eta, \eta')$  if we choose  $\eta'$  as our estimate while  $\eta$  is the true value.

### Risk

For a given loss function  $L$ , we define the **risk**  $R_\theta(\hat{\eta})$  of the estimator  $\hat{\eta}$ , for the value  $\theta \in \Theta$  of the unknown parameter, by

$$R_\theta(\hat{\eta}) = \mathbb{E}_\theta(L(g(\theta), \hat{\eta})).$$

11/60

## Quadratic risk

### Quadratic risk

We call **quadratic risk** the risk associated with the loss function

$$L(\eta, \eta') = \|\eta - \eta'\|^2,$$

that is,

$$R_\theta(\hat{\eta}) = \mathbb{E}_\theta(\|g(\theta) - \hat{\eta}\|^2).$$

Remarks

- ▶ Also called “mean square error” (MSE).
- ▶ **Most commonly used** notion of risk (for the sake of simplicity, as we will see);
- ▶ in the rest of this lecture, **we will consider this risk exclusively**.

12/60

## Example 1 (reminder)

IID Gaussian  $n$ -sample:  $\underline{X} = (X_1, \dots, X_n)$  with

- ▶  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ ,
- ▶  $\theta = (\mu, \sigma^2)$ ,
- ▶  $\Theta = \mathbb{R} \times ]0; +\infty[$ .

In this example, we assume that we want to **estimate the mean  $\mu$** ;

- ▶ here  $\eta = \mu$  and  $g : \theta = (\mu, \sigma^2) \mapsto \mu$ ,
- ▶  $\sigma^2$  is unknown too (nuisance parameter).

13/60

## Example 1: risk of the estimator $\hat{\mu}_1$

Consider the estimator

$$\hat{\mu}_1 = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

For all  $\theta = (\mu, \sigma^2) \in \Theta$ , we have the following result:

### Quadratic risk of the sample mean

$$R_\theta(\hat{\mu}_1) = \mathbb{E}_\theta \left( (\hat{\mu}_1 - \mu)^2 \right) = \frac{\sigma^2}{n}.$$

Remark: the result holds as soon as the  $X_i$ 's have finite second-order moments  
(Gaussianity is not actually used)

14/60



## Example 1: risk of the estimator $\hat{\mu}_1$ (computation)

Notice that

$$\mathbb{E}_\theta(\hat{\mu}_1) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta(X_i) = \mu.$$

Therefore

$$\begin{aligned} R_\theta(\hat{\mu}_1) &= \text{var}_\theta(\hat{\mu}_1) = \frac{1}{n^2} \text{var}_\theta\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{var}_\theta(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

□

15/60

## Bias of an estimator

Let  $\hat{\eta}$  be an estimator of  $\eta = g(\theta)$  admitting a first-order moment, for all  $\theta \in \Theta$ .

**Definition: bias / unbiased estimator**

The **bias** of an estimator  $\hat{\eta}$  at  $\theta \in \Theta$  is defined as

$$b_\theta(\hat{\eta}) = \mathbb{E}_\theta(\hat{\eta}) - g(\theta).$$

We will say that  $\hat{\eta}_n$  is an **unbiased estimator** (UE) if

$$b_\theta(\hat{\eta}) = 0, \quad \forall \theta \in \Theta.$$

Example 1

- ▶ We have already seen that  $\hat{\mu}_1 = \bar{X}_n$  is an UE of  $\mu$ .
- ▶ More generally:  $\hat{\mu} = \alpha + \beta \bar{X}_n$  is an UE of  $\mu$  if, and only if,  $\alpha = 0$  et  $\beta = 1$ .

16/60

## Bias-variance decomposition

Reminder: we still consider the **quadratic risk**.

Let  $\hat{\eta}$  be an estimator  $\eta = g(\theta)$  admitting a second-order moment,  
 $\forall \theta \in \Theta$

### Proposition: Bias-variance decomposition (scalar case)

If the quantity of interest is scalar ( $\eta \in \mathbb{R}$ ), we have:

$$R_{\theta}(\hat{\eta}) = \mathbb{E}_{\theta}((\hat{\eta} - g(\theta))^2) = \text{var}_{\theta}(\hat{\eta}) + \text{b}_{\theta}(\hat{\eta})^2.$$

exercise 1

Remark: we can generalize to the vector case by summing over the components:

$$R_{\theta}(\hat{\eta}) = \mathbb{E}_{\theta}(\|\hat{\eta} - g(\theta)\|^2) = \text{tr}(\text{var}_{\theta}(\hat{\eta})) + \|\text{b}_{\theta}(\hat{\eta})\|^2,$$

where  $\text{var}_{\theta}(\hat{\eta})$  is the covariance matrix of  $\hat{\eta}$ .

17/60

## Example 1: risk of some estimators

$$\hat{\mu}_1 = \bar{X}_n \quad R_{\theta}(\hat{\mu}_1) = \frac{\sigma^2}{n} + 0^2$$

$$\hat{\mu}_2 = \mu_0 \quad R_{\theta}(\hat{\mu}_2) = 0^2 + (\mu - \mu_0)^2$$

$$\hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n \quad R_{\theta}(\hat{\mu}_3) = \frac{1}{4} \frac{\sigma^2}{n} + \frac{1}{4} (\mu - \mu_0)^2$$

$$\hat{\mu}_4 = \bar{X}_n + c \quad R_{\theta}(\hat{\mu}_4) = \frac{\sigma^2}{n} + c^2$$

$$\hat{\mu}_5 = \text{med}(X_1, \dots, X_n) \quad R_{\theta}(\hat{\mu}_5) \approx 1.57 \frac{\sigma^2}{n} + 0^2 \quad (n \rightarrow +\infty)$$

Exercise: Compute  $R_{\theta}(\hat{\mu}_j)$ ,  $2 \leq j \leq 4$

exercise 1

Remark: only the result for  $\hat{\mu}_5$  actually uses the Gaussianity assumption.

18/60

## Admissible estimators

### Definition: order relation on the set of estimators

We will say that  $\hat{\eta}'$  is (weakly) **preferable** to  $\hat{\eta}$  if

$$\blacktriangleright \forall \theta \in \Theta, R_{\theta}(\hat{\eta}') \leq R_{\theta}(\hat{\eta}),$$

We will say that it is **strictly preferable** to  $\hat{\eta}$  if, in addition,

$$\blacktriangleright \exists \theta \in \Theta, R_{\theta}(\hat{\eta}') < R_{\theta}(\hat{\eta}).$$

#### Remarks

- ▶ The relation “is preferable to” is a partial order on risk functions.
- ▶ In general there is no optimal estimator, i.e., no estimator that is preferable to all the others (unless we restrict the class of estimators that is considered).

### Admissibility

We will say that  $\hat{\eta}$  is **admissible** if there is no estimator  $\hat{\eta}'$  that is strictly preferable to it.

19/60

## Example 1 (cont'd)

$\hat{\mu}_1 = \bar{X}_n$	$R_{\theta}(\hat{\mu}_1) = \frac{\sigma^2}{n} + 0^2$
$\hat{\mu}_2 = \mu_0$	$R_{\theta}(\hat{\mu}_2) = 0^2 + (\mu - \mu_0)^2$
$\hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n$	$R_{\theta}(\hat{\mu}_3) = \frac{1}{4} \frac{\sigma^2}{n} + \frac{1}{4} (\mu - \mu_0)^2$
$\hat{\mu}_4 = \bar{X}_n + c$	$R_{\theta}(\hat{\mu}_4) = \frac{\sigma^2}{n} + c^2$

- ▶  $\hat{\mu}_1$  is strictly preferable to  $\hat{\mu}_4$ , therefore  $\hat{\mu}_4$  is not admissible.
- ▶ It can be proved
  - ▶ that  $\hat{\mu}_1$ ,  $\hat{\mu}_2$ , and  $\hat{\mu}_3$  are pairwise incomparable,
  - ▶ but that all three are admissible (proof out of scope)

exercise 1

20/60

## Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Standard exercises

6 – Appendices

## Context and motivation

We consider the class of **unbiased estimators** of  $g(\theta)$ ,

⇒ for an UE,  $R_\theta(\hat{\eta}) = \text{var}_\theta(\hat{\eta})$ .

Objective of this section: show that there exists a bound of the form

$$\text{var}_\theta(\hat{\eta}) \geq v_{\min}(\theta), \quad \forall \theta \in \Theta,$$

that holds for (nearly) **all** UE of  $g(\theta)$ .

Application of such a bound?

- ① Prove that a certain level of accuracy cannot be met by an unbiased estimator.
- ② Prove that a given UE is **optimal** (that is, it minimizes, within the class of UEs, the risk  $R_\theta(\hat{\eta}), \forall \theta \in \Theta$ ).

## Regularity condition $C_0$ and $C_1$

### Regularity condition $C_0$

Dominated model: there exists a ( $\sigma$ -finite) measure  $\nu$  on  $(\mathcal{X}, \mathcal{A})$ , and a family  $(f_\theta)$  of probability density functions wrt  $\nu$ , such that

$$\forall A \in \mathcal{A}, \quad \mathbb{P}_\theta(\underline{X} \in A) = \int_A f_\theta(\underline{x}) \nu(d\underline{x}).$$

### Regularity condition $C_1$

The densities  $f_\theta$  share a **common support**:  $\exists \mathcal{S} \in \mathcal{A}$ ,

$$\forall \theta \in \Theta, \quad \mathbb{1}_{f_\theta > 0} = \mathbb{1}_{\mathcal{S}} \quad \nu\text{-ae.}$$

► Consequently, it can be assumed that  $f_\theta(x) > 0 \Leftrightarrow x \in \mathcal{S}$ .

22/60

## Regularity condition $C_1$ : examples / counter-example

Consider an IID univariate  $n$ -sample:

$$\underline{X} \sim f_\theta(\underline{x}) = \prod_{i=1}^n f_\theta(x_i)$$

(with a usual abuse of notation for the pdf's).

Remark: if  $C_1$  holds for  $n = 1$  with  $\mathcal{S} = \mathcal{S}_1$ ,  
then it also holds for all  $n \geq 2$  with  $\mathcal{S} = \mathcal{S}_1^n$ .

A few examples. . .

- ①  $\mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2 > 0$ :  $C_1$  holds with  $\mathcal{S}_1 = \mathbb{R}$ ,
- ②  $\mathcal{E}(\theta)$ :  $C_1$  holds with  $\mathcal{S}_1 = [0, +\infty)$ .
- ③  $\mathcal{U}_{[0, \theta]}$ :  $C_1$  does not hold!

23/60

## Another regularity condition

We assume that  $C_0$  and  $C_1$  hold.

### Regularity condition $C_2$

- i  $\Theta$  is an open subset of  $\mathbb{R}^p$ ,
- ii  $\theta \mapsto f_\theta(\underline{x})$  is differentiable for  $\nu$ -almost all  $\underline{x}$ ,
- iii and, at any  $\theta \in \Theta$ , we have

$$\int_{\mathcal{S}} \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}) = \nabla_{\theta} \int_{\mathcal{S}} f_{\theta}(\underline{x}) \nu(d\underline{x}) = 0.$$

In other words:  $\forall \theta \in \Theta, \forall k \leq p$ ,

$$\int_{\mathcal{S}} \frac{\partial f_{\theta}(\underline{x})}{\partial \theta_k} \nu(d\underline{x}) = \frac{\partial}{\partial \theta_k} \int_{\mathcal{S}} f_{\theta}(\underline{x}) \nu(d\underline{x}) = 0.$$

24/60

## Score

### Definition / property: score

Assume that  $C_0$ ,  $C_1$ ,  $C_2$ -i and  $C_2$ -ii hold and define, for all  $\underline{x} \in \mathcal{S}$

$$S_{\theta}(\underline{x}) = \nabla_{\theta} (\ln f_{\theta}(\underline{x})) = \begin{pmatrix} \frac{\partial \ln f_{\theta}(\underline{x})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln f_{\theta}(\underline{x})}{\partial \theta_p} \end{pmatrix}.$$

Then

- i We call **score** the random vector  $S_{\theta} = S_{\theta}(X)$ .
- ii  $C_2$ -iii  $\Leftrightarrow \forall \theta \in \Theta$ , the score  $S_{\theta}$  is **centered** under  $\mathbb{P}_{\theta}$ .

Remarks:

- ▶ Well defined, since  $\underline{X} \in \mathcal{S}$   $\mathbb{P}_{\theta}$ -ps,  $\forall \theta \in \Theta$ .
- ▶ The score vanishes at the MLE (recall that  $\Theta \subset \mathbb{R}^p$  is assumed open).

25/60

## The score is centered (proof)

Notice that

$$\nabla_{\theta} (\ln f_{\theta}) = \frac{1}{f_{\theta}} \nabla_{\theta} f_{\theta},$$

and thus, for all  $\theta \in \Theta$ ,

$$\begin{aligned} \mathbb{E}_{\theta}(S_{\theta}) &= \int_{\mathcal{S}} S_{\theta}(\underline{x}) f_{\theta}(\underline{x}) \nu(d\underline{x}) \\ &= \int_{\mathcal{S}} \frac{1}{f_{\theta}(\underline{x})} \nabla_{\theta} f_{\theta}(\underline{x}) f_{\theta}(\underline{x}) \nu(d\underline{x}) \\ &= \int_{\mathcal{S}} \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}). \end{aligned}$$

Finally,

$$\mathbb{E}_{\theta}(S_{\theta}) = 0 \quad \Leftrightarrow \quad \int_{\mathcal{S}} \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}) = 0 \quad (\text{C}_2\text{-iii}). \quad \square$$

26/60

## Example 2

Recall that  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$  with  $\theta \in \Theta = ]0, +\infty[$ .

We compute the **likelihood**, for any  $x_1, \dots, x_n \geq 0$ :

$$\mathcal{L}(\theta; \underline{x}) = f_{\theta}(\underline{x}) = \prod_{i=1}^n f_{\theta}(x_i) = \theta^n e^{-\theta \sum x_i},$$

then the **log-likelihood**:

$$\ln \mathcal{L}(\theta; \underline{x}) = \ln f_{\theta}(\underline{x}) = n \ln \theta - \theta \sum x_i,$$

and, finally, the **score**:

$$S_{\theta}(\underline{X}) = \sum_{i=1}^n S_{\theta}(X_i) = n \left( \frac{1}{\theta} - \bar{X}_n \right).$$

27/60

## Remark on condition C<sub>2</sub>-iii

Recall C<sub>2</sub>-iii:  $\forall \theta \in \Theta$ ,

$$\int_{\mathcal{S}} \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}) = \nabla_{\theta} \int_{\mathcal{S}} f_{\theta}(\underline{x}) \nu(d\underline{x}) = 0,$$

or, equivalently:  $\mathbb{E}_{\theta}(S_{\theta}) = 0$ .

Two approaches are available to check this condition:

- ① Compute **explicitly**  $\mathbb{E}_{\theta}(S_{\theta}) = \int_{\mathcal{S}} \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x})$ .
- ② Use a **domination** condition: show that  $\forall \theta_0 \in \Theta$ ,  $\exists \mathcal{V} \subset \Theta$ , neighborhood of  $\theta_0$ , and a  $\nu$ -integrable function  $g : \mathcal{X} \rightarrow \mathbb{R}$  st

$$\forall \theta \in \mathcal{V}, \forall \underline{x} \in \mathcal{S}, \forall k \leq p, \quad \left| \frac{\partial f_{\theta}(\underline{x})}{\partial \theta_k} \right| \leq g(\underline{x}).$$

28/60

## Cramér-Rao inequality

1/2

Consider a statistical model where C<sub>0</sub>-C<sub>2</sub> hold, and let  $\hat{\eta}$  be an estimator of  $\eta = g(\theta) \in \mathbb{R}$ .

### Definition: regular estimator

We will say that  $\hat{\eta}$  is a **regular** estimator if

- ①  $\mathbb{E}_{\theta}(\hat{\eta}^2) < +\infty$ ,  $\forall \theta \in \Theta$ ,
- ②  $\theta \mapsto \mathbb{E}_{\theta}(\hat{\eta})$  is differentiable, with

$$\nabla_{\theta} \mathbb{E}_{\theta}(\hat{\eta}) = \int_{\mathcal{S}} \hat{\eta}(\underline{x}) \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}), \quad \forall \theta \in \Theta.$$

Remark: if  $\hat{\eta}$  is an unbiased regular estimator of  $g(\theta)$ , then

$$(\nabla g)(\theta) = \int_{\mathcal{S}} \hat{\eta}(\underline{x}) \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}), \quad \forall \theta \in \Theta.$$

29/60



## Theorem / definition: Cramér-Rao inequality

Consider a statistical model where  $C_0$ - $C_2$  hold, and assume that the score  $S_\theta$  admits second-order moments for all  $\theta \in \Theta$ .

Let  $\text{var}_\theta(S_\theta)$  denote the **covariance matrix** of the score, which is assumed invertible for all  $\theta \in \Theta$ .

Let  $\hat{\eta}$  be a **regular unbiased** estimator of  $g(\theta)$ . Then,  $\forall \theta \in \Theta$ ,

$$R_\theta(\hat{\eta}) = \text{var}_\theta(\hat{\eta}) \geq \nabla g(\theta)^\top \text{var}_\theta(S_\theta)^{-1} \nabla g(\theta).$$

An unbiased estimator is called **efficient** if this bound is met for all  $\theta$ .

 Proof

30/60

## Fisher information

We still assume that  $C_0$ - $C_2$  hold.

## Definition: Fisher information

We call **Fisher information** of  $\underline{X}$  the  **$p \times p$  matrix**

$$I(\theta) = \text{var}_\theta(S_\theta) = \mathbb{E}_\theta \left( S_\theta S_\theta^\top \right)$$

which appears in the Cramér-Rao lower bound.

## Proposition

Let  $I_n(\theta)$  denote the Fisher information in an IID  $n$ -sample. Then

$$I_n(\theta) = n I_1(\theta).$$

The CR inequality becomes:  $\text{var}_\theta(\hat{\eta}) \geq \frac{1}{n} \nabla g(\theta)^\top I_1(\theta)^{-1} \nabla g(\theta)$ .

31/60

## Proof

Notice that the score is additive in an IID sample:

$$\begin{aligned} S_\theta &= \nabla_\theta (\ln f_\theta(\underline{x})) \\ &= \nabla_\theta \left[ \ln \left( \prod_{i=1}^n f_\theta^{X_1}(X_i) \right) \right] = \sum_{i=1}^n \underbrace{\nabla_\theta \left( \ln f_\theta^{X_1}(X_i) \right)}_{Z_i}. \end{aligned}$$

Thus we have

$$\text{var}_\theta(S_\theta) = \sum_{i=1}^n \text{var}_\theta(Z_i) = n \text{var}_\theta(Z_1) = n l_1(\theta)$$

since  $Z_1, \dots, Z_n$  are IID, and distributed like the score in a sample of size 1.  $\square$

32/60

## Example 1: estimation of $\mu$

Reminder:  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$  and  $\theta = (\mu, \sigma^2)$

- ▶  $\hat{\mu}_n = \bar{X}_n$  is the MLE of  $\mu$ ,
- ▶  $\hat{\mu}_n$  is unbiased and  $R_\theta(\hat{\mu}_n) = \text{var}_\theta(\hat{\mu}_n) = \frac{\sigma^2}{n}$ .

The **Fisher information matrix** in this model is (see PC 2)

$$I_n(\theta) = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

Cramér-Rao inequality with  $g(\theta) = \mu$ :  $\forall \hat{\mu}'_n$  UE of  $\mu$ ,

$$R_\theta(\hat{\mu}'_n) = \text{var}_\theta(\hat{\mu}'_n) \geq \frac{\sigma^2}{n},$$

therefore  $\hat{\mu}_n = \bar{X}_n$  is efficient.

33/60

## Example 1': estimation of $\sigma^2$

Same statistical model, but we want to estimate  $g(\theta) = \sigma^2$ .

It is then possible to show (see PC 2) that

- ▶ the MLE  $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  is biased;
- ▶  $\hat{\sigma}_n^2 = (S'_n)^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  is an **UE of  $\sigma^2$** , with variance

$$\text{var}_{\theta}(\hat{\sigma}_n^2) = \frac{2\sigma^4}{n-1}.$$

Conclusion:  $\hat{\sigma}_n^2$  is **not an efficient estimator**, since

$$\text{var}_{\theta}(\hat{\sigma}_n^2) > \frac{2\sigma^4}{n}.$$

(Beware the misleading terminology: it can be proved, using Lehmann-Scheffé's theorem, that  $\hat{\sigma}_n^2$  is a *minimal variance* UE for this problem, and therefore is optimal for the quadratic risk among all UE's.)

34/60

## Lecture outline

- 1 – Point estimation: definition and notations
- 2 – Quadratic risk of an estimator
- 3 – A lower bound on the quadratic risk
- 4 – Asymptotic properties
- 5 – Standard exercises
- 6 – Appendices

## Motivation / notations

### Problem

It is sometimes (often !) difficult to obtain the exact properties of statistical procedures.

(point estimators, but also CIs, tests, etc. (cf. next lectures))

### Asymptotic approach(es) → approximate properties

- ▶  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P_\theta$ , defined on a common  $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$
- ▶ Sequences of estimators:  $\hat{\eta}_n = \hat{\eta}_n(X_1, \dots, X_n)$
- ▶ Properties of the estimators when  $n \rightarrow \infty$ ?

Remark: we have now not one but a **sequence  $(\mathcal{M}_n)_{n \geq 1}$  of statistical models**

$$\mathcal{M}_n = (\mathcal{X}^n, \mathcal{A}^{\otimes n}, \{P_\theta^{\otimes n}, \theta \in \Theta\}),$$

that we instantiate on a common underlying probability space  $(\Omega, \mathcal{F})$ .

35/60

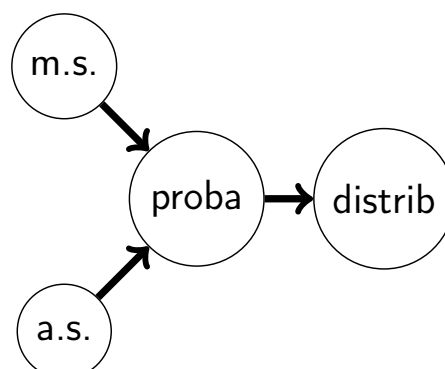
## Probability refresher: convergence modes

Main convergence modes that are useful in statistics:

- ▶ **almost sure** convergence ,
- ▶ convergence **in  $L^2$**  (in mean square),
- ▶ convergence **in probability**,
- ▶ convergence **in distribution**.

Implications between convergence modes:

Supplements



36/60

## Consistency

Let  $(\hat{\eta}_n)$  denote a sequence of estimators of  $\eta = g(\theta)$ .

### (weak) Consistency

We will say that  $\hat{\eta}_n$  is a **consistent** estimator of  $\eta = g(\theta)$  if,  $\forall \theta \in \Theta$ ,

$$\hat{\eta}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\theta} g(\theta). \quad (\text{with an obvious abuse of terminology})$$

### Strong and mean-square consistency

We will say that  $\hat{\eta}_n$  is **strongly consistent** (resp. **consistent in the mean-square sense**) if,  $\forall \theta \in \Theta$ ,

$$\hat{\eta}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\theta\text{-a.s.}} g(\theta) \quad \left( \text{resp., } \hat{\eta}_n \xrightarrow[n \rightarrow \infty]{L^2(\mathbb{P}_\theta)} g(\theta) \right).$$

Remark: the word “convergent” is sometimes used instead of “consistent”.

37/60

## Probability refresher: law of large numbers

Let  $(X_k)_{k \geq 1}$  be a sequence of real- or vector-valued RV.

### Strong law of large numbers

If the  $X_k$ 's are **IID** and have finite **first-order moments**, then

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}(X_1).$$

### Law of large numbers in $L^2$

If the  $X_k$ 's are **IID** and have finite **second-order moments**, then

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{L^2} \mathbb{E}(X_1).$$

Proof (scalar case):  $\mathbb{E} \left( (\bar{X}_n - \mathbb{E}(X_1))^2 \right) = \text{var}_\theta(\bar{X}_n) = \frac{1}{n} \text{var}_\theta(X_1) \rightarrow 0.$   $\square$

38/60

## Consistency: examples

A) IID  $n$ -sample with finite first-order moment

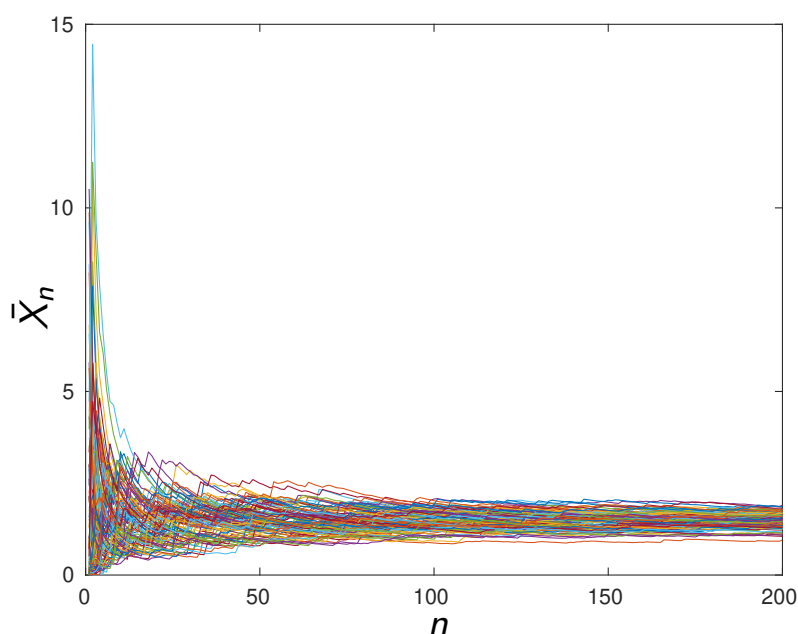
- ▶ i.e.,  $\mathbb{E}_\theta(\|X_1\|) < +\infty$ , for all  $\theta \in \Theta$ .
- ▶  $\bar{X}_n$  is a **strongly consistent** estimator of  $\eta = \mathbb{E}_\theta(X_1)$ .
- ▶ Nothing can be said about the quadratic risk without additional assumptions.

B) IID  $n$ -sample with finite second-order moment

- ▶ i.e.,  $\mathbb{E}_\theta(\|X_1\|^2) < +\infty$ , for all  $\theta \in \Theta$ .
- ▶  $\bar{X}_n$  is **strongly consistent** and **consistent in the mean-square sense** for  $\eta = \mathbb{E}_\theta(X_1)$ .

39/60

## Consistency: examples (cont'd)



Convergence of  $\bar{X}_n$  to the true mean  
(for a Gamma  $n$ -sample with true mean  $\mu = 1.5$ )

40/60

## Consistency: examples (cont'd)

### C) IID $n$ -sample (with any distribution)

- ▶ Let  $A \in \mathcal{A}$  and  $\eta = g(\theta) = \mathbb{P}_\theta(X_1 \in A)$ .
- ▶ Relative frequency:  $\hat{\eta}_n = \frac{1}{n} \text{card} \{i \leq n \mid X_i \in A\}$
- ▶  $\hat{\eta}_n$  is a **strongly** and **mean-square consistent** estimator of  $\eta$ .

Application: histograms

exercise 3

### D) MLE of an $n$ -sample distributed according to the uniform distribution (see PC 1)

- ▶  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}_{[0, \theta]}$
- ▶ We estimate  $\eta = \theta$  with  $\hat{\eta}_n = \max_{i \leq n} X_i$ .
- ▶  $\hat{\eta}_n$  is consistent, both **strongly** and in the **mean-square sense**.

### E) Maximum likelihood estimator

complement

41/60

## Summary and preview

### We have seen and will practice in PC 2:

- ▶ the quantitative assessment of an estimator's performance through risk computation,
- ▶ the comparison of estimators and a concept of optimality,
- ▶ the asymptotic analysis of estimators.

### We will cover in Lecture 3:

- ▶ the concept of convergence rate of an estimator,
- ▶ the definition and construction of confidence intervals/regions.

42/60

## Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

**5 – Standard exercises**

5.1 – Questions

5.2 – Answers

6 – Appendices

## Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

**5 – Standard exercises**

5.1 – Questions

5.2 – Answers

6 – Appendices



## Exercise 1 (quadratic risk)

[solution](#)

Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_*^+$ .

We want to estimate  $g(\theta) = \mu$ . We consider the estimators

$$\hat{\mu}_1 = \bar{X}_n, \quad \hat{\mu}_2 = \mu_0, \quad \hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n, \quad \hat{\mu}_4 = \bar{X}_n + c,$$

where  $\mu_0$  and  $c$  are given real numbers.

### Questions

- 1 Prove the bias-variance decomposition formula in the scalar case. [back to slide 17](#)
- 2 Compute the quadratic risk of each of these estimators
- 3 Prove that  $\hat{\mu}_2$  and  $\hat{\mu}_3$  are not comparable.
- 4 Prove that  $\hat{\mu}_4$  is not admissible. [back to slide ??](#)

43/60

## Exercise 2 (efficiency of an estimator)

[solution](#)

Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$  with  $\theta \in \Theta = ]0, 1[$ .

Recall that (see Exercises in Lecture 1):

- the log-likelihood of the  $n$ -sample is

$$\ln \mathcal{L}(\theta; \underline{x}) = \ln f_{\theta}(\underline{x}) = n \ln(1 - \theta) - \ln \left( \frac{\theta}{1 - \theta} \right) \sum_{i=1}^n x_i,$$

- the MLE is  $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$ .

### Questions

- 1 Check that the model satisfies the hypotheses for Cramér-Rao's inequality, and compute Cramér-Rao's bound.
- 2 Is the MLE  $\hat{\theta}_n$  efficient?

44/60

## Exercise 3 (consistency of an histogram)

 solution

Consider:

 back to slide 41

- ▶ an  $n$ -sample  $X_1, \dots, X_n$ , with  $X_i$  in  $]a, b] \subset \mathbb{R}$ ,
- ▶ a partition of  $]a, b]$  in  $K$  adjacent classes  $A_k = ]a_{k-1}, a_k]$ , for  $k \in \{1, \dots, K\}$ , with  $a_0 = a$ ,  $a_K = b$ ,
- ▶ the vector  $\eta \in \mathbb{R}^K$  with  $\eta^{(k)} = P(X_1 \in A_k)$ .

### Histogram

Graphical representation of the empirical distribution of a random variable using rectangles, where the bases are the intervals  $A_k$  and the areas are proportional to the relative frequencies  $\hat{\eta}_n^{(k)}$  of the classes:

$$\hat{\eta}_n^{(k)} = \frac{1}{n} \text{card} \{i \leq n \mid X_i \in A_k\}, \quad 1 \leq k \leq K.$$

**Question.** Prove that  $\hat{\eta}_n = (\hat{\eta}_n^{(1)}, \dots, \hat{\eta}_n^{(K)})$  is a **strongly consistent** and **mean-square consistent** estimator of  $\eta$ .

45/60

## Exercise 4 (mean-square consistency)

 solution

Let  $\hat{\eta}_n$  denote an estimator of a scalar parameter  $\eta = g(\theta) \in \mathbb{R}$ , indexed by the size  $n$  of the observed sample.

### Question

Prove that  $\hat{\eta}_n$  is consistent in the mean-square sense if, and only if, the following conditions are satisfied for all  $\theta \in \Theta$ :

- i  $b_\theta(\hat{\eta}_n) \rightarrow 0$ ,
- ii  $\text{var}_\theta(\hat{\eta}_n) \rightarrow 0$ .

46/60

## Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Standard exercises

5.1 – Questions

5.2 – Answers

6 – Appendices

## Exercise solution 1

[back to questions](#)

### 1 Bias-variance decomposition

$$R_{\theta}(\hat{\eta}) = \mathbb{E}_{\theta}((\hat{\eta} - g(\theta))^2) = \text{var}_{\theta}(\hat{\eta}) + \text{b}_{\theta}(\hat{\eta})^2.$$

#### Proof

$$\begin{aligned} R_{\theta}(\hat{\eta}) &= \mathbb{E}_{\theta}((\hat{\eta} - g(\theta))^2) \\ &= \mathbb{E}_{\theta}((\hat{\eta} - \mathbb{E}_{\theta}(\hat{\eta}) + \text{b}_{\theta}(\hat{\eta}))^2) \\ &= \underbrace{\mathbb{E}_{\theta}((\hat{\eta} - \mathbb{E}_{\theta}(\hat{\eta}))^2)}_{\text{var}_{\theta}(\hat{\eta})} + \text{b}_{\theta}(\hat{\eta})^2 + 2 \underbrace{\mathbb{E}_{\theta}(\hat{\eta} - \mathbb{E}_{\theta}(\hat{\eta}))}_{=0} \text{b}_{\theta}(\hat{\eta}) \\ &= \text{var}_{\theta}(\hat{\eta}) + \text{b}_{\theta}(\hat{\eta})^2. \end{aligned}$$

□

## Exercise solution 1

[back to questions](#)

② Compute the bias and variance of each estimator, and then conclude using the bias-variance decomposition.

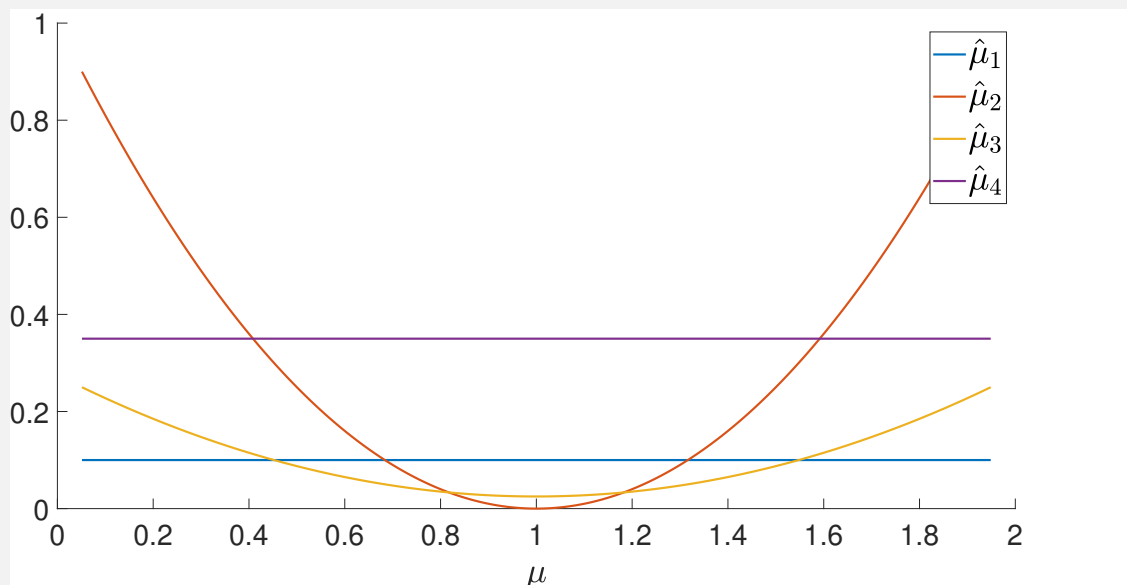
	expectation	bias	variance	quadratic risk
$\bar{X}_n$	$\mu$	0	$\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n}$
$\mu_0$	$\mu_0$	$\mu_0 - \mu$	0	$(\mu_0 - \mu)^2$
$\frac{1}{2} (\mu_0 + \bar{X}_n)$	$\frac{1}{2} (\mu_0 + \mu)$	$\frac{1}{2} (\mu_0 - \mu)$	$\frac{1}{4} \frac{\sigma^2}{n}$	$\frac{1}{4} \frac{\sigma^2}{n} + \frac{1}{4} (\mu_0 - \mu)^2$
$\bar{X}_n + c$	$\mu + c$	$c$	$\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n} + c^2$

Reminder:  $\text{var}_\theta(\alpha X + \beta) = \alpha^2 \text{var}_\theta(X)$ .

48/60

## Exercise solution 1

[back to questions](#)



Draw the four risks for  $\sigma^2 = 1$ ,  $n = 10$ ,  $\mu_0 = 1$  and  $c = 0.5$ .

49/60

③ Let us compute the risk two well-chosen points.

For  $\theta = (\mu_0, 1)$  we have

$$R_{\theta}(\hat{\mu}_2) = 0, \quad R_{\theta}(\hat{\mu}_3) = \frac{1}{4n}, \quad \text{therefore } R_{\theta}(\hat{\mu}_2) < R_{\theta}(\hat{\mu}_3).$$

For  $\theta = \left(\mu_0 + \frac{1}{\sqrt{n}}, 1\right)$  we have

$$R_{\theta}(\hat{\mu}_2) = \frac{1}{n}, \quad R_{\theta}(\hat{\mu}_3) = \frac{1}{2n}, \quad \text{therefore } R_{\theta}(\hat{\mu}_2) > R_{\theta}(\hat{\mu}_3).$$

Therefore the estimators  $\hat{\mu}_2$  and  $\hat{\mu}_3$  are not comparable.

④ We have:

$$\begin{cases} R_{\theta}(\hat{\mu}_4) = \frac{\sigma^2}{n} + c^2 \\ R_{\theta}(\hat{\mu}_1) = \frac{\sigma^2}{n} \end{cases}$$

Therefore,  $\forall \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_*^+$ ,  $R_{\theta}(\hat{\mu}_4) > R_{\theta}(\hat{\mu}_1)$

Thus  $\hat{\mu}_4$  is not admissible.

❶ Let us check that the model satisfies the regularity conditions  $C_1$  and  $C_2$ , and that Fisher's information does not vanish.

⇒  $C_1$ : since  $\Theta = ]0, 1[$ , the densities

$$f_{\theta}(\underline{x}) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

are all supported on  $\mathcal{S} = \{0, 1\}^n$ .

⇒  $C_2$ :  $\Theta = ]0, 1[$  is an open subset of  $\mathbb{R}$ ,  $\theta \mapsto f_{\theta}(\underline{x})$  is differentiable on  $\Theta$  for all  $\underline{x}$ , and the score

$$S_{\theta}(\underline{X}) = \frac{\partial(\ln f_{\theta})}{\partial \theta}(X_i) = \frac{n}{\theta(1 - \theta)} (\bar{X}_n - \theta)$$

is centered:  $\mathbb{E}_{\theta}(S_{\theta}(\underline{X})) = \frac{n}{\theta(1 - \theta)} (\mathbb{E}_{\theta}(\bar{X}_n) - \theta) = 0$ .

⇒ Finally, we check that the Fisher information does not vanish:

$$I(\theta) = \text{var}_{\theta}(S_{\theta}(\underline{X})) = \left( \frac{n}{\theta(1 - \theta)} \right)^2 \text{var}_{\theta}(\bar{X}_n) = \frac{n}{\theta(1 - \theta)} > 0.$$

⇒ The Cramér-Rao bound for  $\theta$  is

$$I(\theta)^{-1} = \frac{1}{n} \theta(1 - \theta).$$

❷ The estimator  $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$  is unbiased:

$$\mathbb{E}_{\theta}(\hat{\theta}_n) = \mathbb{E}_{\theta}(X_1) = \theta,$$

and its variance is

$$\text{var}(\hat{\theta}) = \frac{1}{n} \text{var}(X_1) = \frac{\theta(1-\theta)}{n} = I(\theta)^{-1}.$$

Therefore it is efficient. □

Remark: it is easy to check that  $\hat{\theta}_n$  is a regular estimator (see definition on slide 29), since

- a the density  $f_{\theta}$  is differentiable with respect to  $\theta$ ,
- b the integrals boil down to finite sums over  $\{0, 1\}^n$ .

### ❶ Strong consistency

Reminder:  $\hat{\eta}_n \xrightarrow{\text{as}} \eta$  iff  $\hat{\eta}_n^{(k)} \xrightarrow{\text{as}} \eta^{(k)}, \forall k$ .

For all  $k \in \{1, \dots, K\}$ , we have:

$$\hat{\eta}_n^{(k)} = \frac{1}{n} \text{card} \{i \leq n \mid X_i \in A_k\} = \frac{1}{n} \sum_{i=1}^n Z_i \quad \text{with } Z_i = 1_{A_k}(X_i).$$

The strong law of large numbers, applied to  $(Z_i)_{i \geq 1}$ , then yields:

$$\hat{\eta}_n^{(k)} = \frac{1}{n} \sum_{i=1}^n Z_i \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}(Z_1) = \eta^{(k)}.$$

since  $Z_1, Z_2, \dots \stackrel{\text{iid}}{\sim} \text{Ber}(\eta^{(k)})$ .

**② Mean-square consistency**

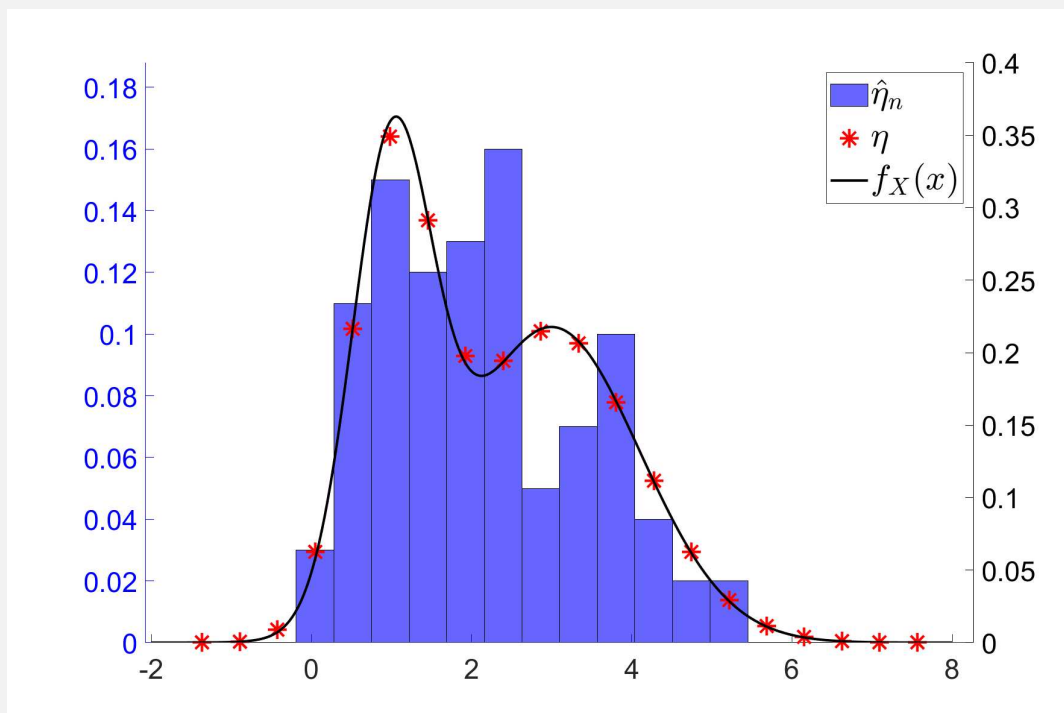
$$\mathbb{E} \left( \|\hat{\eta}_n - \eta\|^2 \right) = \sum_{k=1}^K \mathbb{E} \left( \left( \hat{\eta}_n^{(k)} - \eta^{(k)} \right)^2 \right)$$

with  $k$  fixed:  $\hat{\eta}_n^{(k)} = \bar{Z}_n$  with  $Z_i \sim \text{Ber}(\eta^{(k)})$  of finite variance.

The law of large number in  $L^2$  gives:

$$\bar{Z}_n = \hat{\eta}_n^{(k)} \xrightarrow[n \rightarrow \infty]{L^2} \eta^{(k)}, \quad \text{i.e.} \quad \mathbb{E} \left( \left( \hat{\eta}_n^{(k)} - \eta^{(k)} \right)^2 \right) \xrightarrow[n \rightarrow \infty]{} 0$$

$$\text{Hence } \mathbb{E} \left( \|\hat{\eta}_n - \eta\|^2 \right) \xrightarrow[n \rightarrow \infty]{} 0 \text{ et } \hat{\eta}_n \xrightarrow[n \rightarrow \infty]{L^2} \eta$$

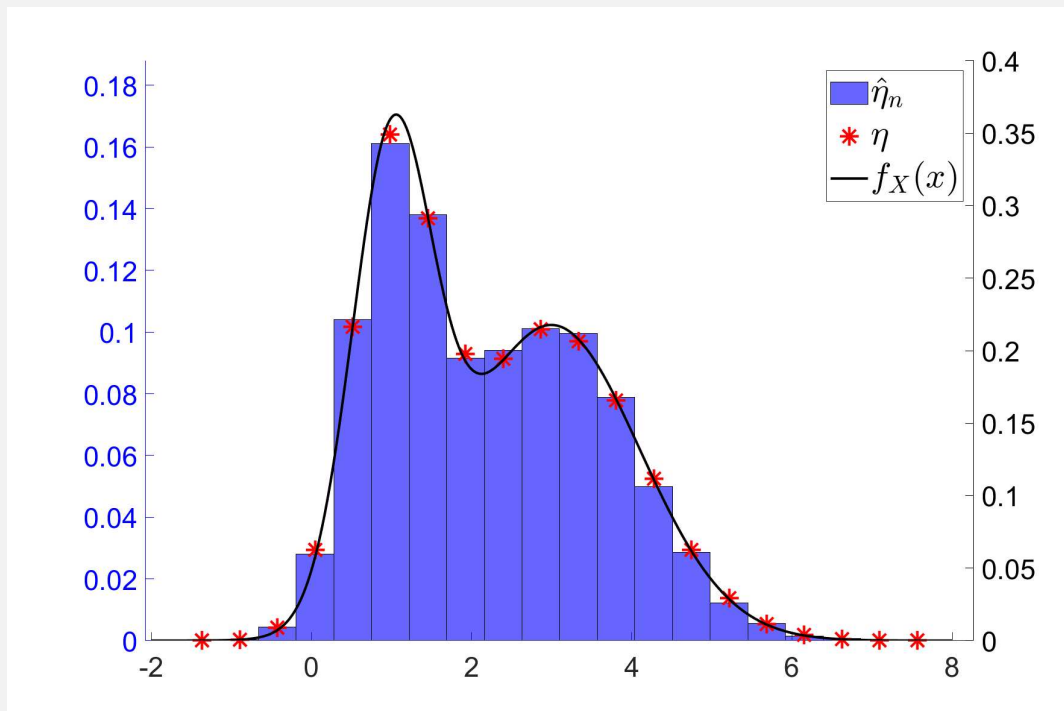


*Histogram calculated for  $N = 100$  and  $K = 20$ .*

Note. The law used in the example has a density  $f_X(x)$ .



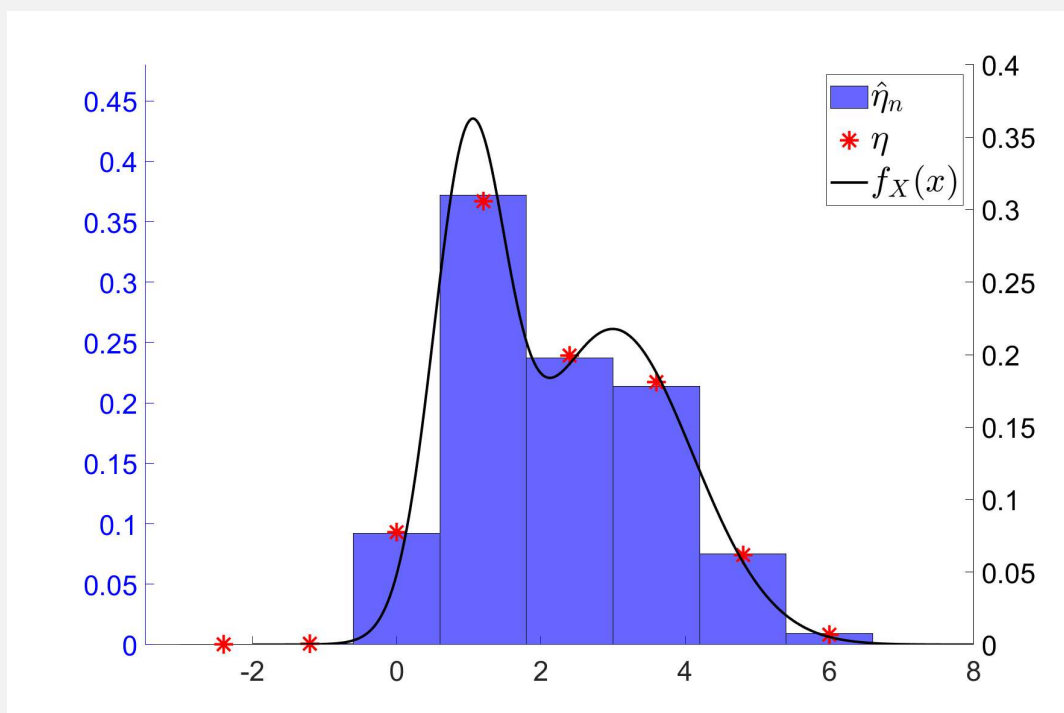
## Exercise solution 3

[back to questions](#)

*Histogram calculated for  $N = 10000$  and  $K = 20$ .*

58/60

## Exercise solution 3

[back to questions](#)

*Histogram calculated for  $N = 10000$  and  $K = 8$ .*

59/60

Consider the bias-variance decomposition of the quadratic risk:

$$\mathbb{E}_{\theta} ((\hat{\eta} - g(\theta))^2) = \text{var}_{\theta}(\hat{\eta}) + \text{b}_{\theta}(\hat{\eta})^2.$$

The two terms in the sum are positive, therefore

$$\mathbb{E}_{\theta} ((\hat{\eta} - g(\theta))^2) \rightarrow 0 \quad \Leftrightarrow \quad \begin{cases} \text{var}_{\theta}(\hat{\eta}) \rightarrow 0, \\ \text{b}_{\theta}(\hat{\eta}) \rightarrow 0. \end{cases}$$

This proves the claimed equivalence. □

## Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Standard exercises

6 – Appendices

6.1 – Reminders & supplements

## Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Standard exercises

6 – Appendices

6.1 – Reminders & supplements

## Other examples, not treated in this course (nonparametric statistics)

### Example 3

- ▶  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} P$ ,
- ▶  $\theta = P$ , unknown distribution,
- ▶  $\Theta = \{\text{distributions on } (\mathbb{R}, \mathcal{B}(\mathbb{R}))\}$ ,
- ▶  $g(\theta) = F$ : cumulative distribution functions of the  $X_i$ 's.

### Example 4

- ▶  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} P_\theta$ ,
- ▶  $P_\theta$ : probability density functions  $\theta(x)$
- ▶  $\Theta = \{\text{pdf on } \mathbb{R}, \text{ of class } \mathcal{C}^2, \text{ with } \int \theta''(x)^2 dx < +\infty\}$
- ▶  $g(\theta) = \theta$ .

## Proof of the Cramér-Rao inequality

Preliminary remark: since  $\hat{\eta}$  is a regular UE of  $g(\theta)$ ,  $g$  is differentiable.

Let  $\theta \in \Theta$ , and set  $c = \text{cov}_{\theta}(S_{\theta}, \hat{\eta}) \in \mathbb{R}^p$ . Then,  $\forall a \in \mathbb{R}^p$ ,

$$\text{var}_{\theta}(\hat{\eta} - a^{\top} S_{\theta}) = \text{var}_{\theta}(\hat{\eta}) - 2a^{\top} c + a^{\top} \text{var}_{\theta}(S_{\theta}) a \geq 0.$$

In particular, for  $a = \text{var}_{\theta}(S_{\theta})^{-1} c \in \mathbb{R}^p$ , we get:

$$\text{var}_{\theta}(\hat{\eta}) - c^{\top} \text{var}_{\theta}(S_{\theta})^{-1} c \geq 0.$$

Finally, since  $S_{\theta}$  is centered and  $\hat{\eta}$  is a regular UE,

$$\begin{aligned} c &= \mathbb{E}_{\theta}(\hat{\eta} S_{\theta}) = \int_{\mathcal{S}} \hat{\eta}(\underline{x}) \cdot \frac{1}{f_{\theta}(\underline{x})} \nabla_{\theta} f_{\theta}(\underline{x}) \cdot f_{\theta}(\underline{x}) \nu(d\underline{x}) \\ &= \int_{\mathcal{S}} \hat{\eta}(\underline{x}) \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}) = \nabla_{\theta} \mathbb{E}_{\theta}(\hat{\eta}) = \nabla g(\theta). \end{aligned} \quad \square$$

[back to slide ??](#)

## Probability refresher: convergence modes

 **almost sure** convergence :

$$T_n \xrightarrow{\text{as}} T \quad \text{if} \quad \mathbb{P}(T_n \rightarrow T) = 1$$

 convergence **in  $L^2$**  (in mean square):

$$\begin{aligned} T_n &\xrightarrow{L^2} T \quad \text{if} \quad \mathbb{E}(\|T_n - T\|^2) \rightarrow 0 \\ &\text{iff} \quad \forall j \leq p, \quad T_n^{(j)} \xrightarrow{L^2} T^{(j)} \end{aligned}$$

 convergence **in probability**:

$$T_n \xrightarrow{P} T \quad \text{if} \quad \forall \varepsilon > 0, \quad \mathbb{P}(\|T_n - T\| \geq \varepsilon) \rightarrow 0$$

 convergence **in distribution**:

$$T_n \xrightarrow{d} T \quad \text{if} \quad \forall \varphi, \quad \mathbb{E}(\varphi(T_n)) \rightarrow \mathbb{E}(\varphi(T)),$$

with  $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$  continuous and bounded.

[back to slide 36](#)

## Consistency of the MLE

The MLE minimizes the following criterion:

$$\gamma_n(\theta) = -\frac{1}{n} \ln f_\theta(\underline{X}) = -\frac{1}{n} \sum_{k=1}^n \ln f_\theta(X_i).$$

Let  $\theta \in \Theta$ , and set  $c = \text{cov}_\theta(S_\theta, \hat{\eta}) \in \mathbb{R}^p$ . Then,  $\forall \theta \in \Theta$ ,

$$\gamma_n(\theta) - \gamma_n(\theta_\star) = \frac{1}{n} \sum_{k=1}^n \ln \frac{f_{\theta_\star}(X_i)}{f_\theta(X_i)} \xrightarrow[n \rightarrow +\infty]{\text{as}} \int_{S_1} \ln \frac{f_{\theta_\star}(x)}{f_\theta(x)} f_{\theta_\star}(x) \nu_1(dx).$$

(assuming that  $Z_i = \frac{f_{\theta_\star}(X_i)}{f_\theta(X_i)}$  has a first-order moment).

**Definition / property: Kullback-Leibler divergence**

$$D_{\text{KL}}(f_{\theta_\star} || f_\theta) = \int_{S_1} \ln \frac{f_{\theta_\star}(x)}{f_\theta(x)} f_{\theta_\star}(x) \nu_1(dx) \geq 0$$

## Consistency of the MLE (cont'd)

Set  $\Delta_n(\theta_\star, \theta) = \frac{1}{n} \sum_{k=1}^n \ln \frac{f_{\theta_\star}(X_i)}{f_\theta(X_i)}$  and  $\Delta(\theta_\star, \theta) = D_{\text{KL}}(f_{\theta_\star} || f_\theta)$ .

We have  $\Delta_n(\theta_\star, \theta) \xrightarrow[n \rightarrow +\infty]{\mathbb{P}_{\theta_\star}\text{-ps}} \Delta(\theta_\star, \theta)$  for all  $\theta$ , and  $\Delta(\theta_\star, \theta_\star) = 0$ .

**Theorem: Consistency of the MLE**

Assume that, for all  $\theta_\star \in \Theta$ ,

- i  $\sup_{\theta \in \Theta} |\Delta_n(\theta_\star, \theta) - \Delta(\theta_\star, \theta)| \xrightarrow[n \rightarrow +\infty]{\mathbb{P}_{\theta_\star}} 0$
- ii and, for all  $\epsilon > 0$ ,

$$\inf_{\theta \in \Theta, \|\theta - \theta_\star\| \geq \epsilon} \Delta(\theta_\star, \theta) > 0.$$

Then the MLE is (weakly) consistent.