CentraleSupélec

Statistics and Learning

Julien Bect

(julien.bect@centralesupelec.fr)

Teaching : CentraleSupélec / dept. of Statistics and Signal Processing

Research : Laboratory of Signals and Systems (L2S)

Lecture 2/9

Point estimation

In this lecture you will learn how to…

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

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

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

Recap : mathematical framework

Data

- Formally, a point $\underline{x}$ in a set $\underline{\mathcal{X}}$.
- ex : $\underline{\mathcal{X}} = \mathbb{R}^n$, $\mathbb{R}^{n \times d}$, $\{\text{words}\}$, 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 $(\underline{\mathcal{X}}, \underline{\mathscr{A}})$,
- **but** the distribution of $\underline{X}$ is unknown.

Statistical modeling

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

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

Recap : mathematical framework (cont'd)

Important

Since $\mathbb{P} \in \mathscr{P}$ is unknown, we must design statistical procedure that "work well" (in a sense to be specified) for **any** distribution $\mathbb{P} \in \mathscr{P}$.

Parameterized family of probability distributions

- Usually, we write $\mathscr{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$.
- θ : 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 (→ $n \times d$ data table)

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

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 θ is unknown.

Informal definition : estimation

Guess (infer) the value of η 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 η.

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

Example 1 (reminder)

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

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

In this example, we assume that we want to estimate the mean μ ;

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

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 $c \neq 0$,
- $\hat{\mu}_5 = \mathrm{med}(X_1, \ldots, X_n)$,
- ...

Questions

- Is one these estimators "better" than the others ?
- Can we find an "optimal" estimator ?
- In what sense ?

Other examples

In the following examples, as in Example 1 :

- $\underline{X} = (X_1, \ldots, X_n)$ is an (IID) n-sample,
- the X_i's are scalar : univariate n-sample.

Example 1'

- Same statistical model as in Example 1, but
- $g(\theta) = \sigma^2$.
- In this case, μ 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.

Other examples (cont'd)

Example 2

- $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{E}(\theta)$, 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 $x_0 > 0$.

Other examples (cont. and end)

Example 3

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

Example 4

- $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{P}_\theta$,
- P_θ : probability density functions $\theta(x)$
- $\Theta = \left\{\text{pdf on } \mathbb{R}, \text{ of class } \mathscr{C}^2, \text{ with } \int \theta''(x)^2 \,\mathrm{d}x < +\infty\right\}$
- $g(\theta) = \sigma^2$.

Examples 3 et 4 : non-parametric statistics (not treated in this course).

Lecture outline

General concept of risk

Goal

Quantify the performance of an estimator

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

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

Risk

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

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

Quadratic risk

Quadratic risk

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

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

that is,

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

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 the lecture, we will consider this risk exclusively.

Example 1 (reminder)

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

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

In this example, we assume that we want to estimate the mean μ ;

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

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

Consider the estimator

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

Let $\theta = \left(\mu, \sigma^2\right) \in \Theta$. We have the following result :

Quadratic risk of the empirical mean

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

Remark : the result holds as soon as the X_i's have second order

(Gaussianity is not actually used)

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

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

□

Bias of an estimator

Let $\hat{\eta}$ be an estimator of $\eta = g(\theta)$ st $\mathbb{E}_\theta\left(\|\hat{\eta}\|\right) < +\infty$, $\forall\theta \in \Theta$.

Definition : bias / unbiased estimator

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

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

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

$$\mathrm{b}_\theta(\hat{\eta}) = 0, \qquad \forall\theta \in \Theta.$$

Example 1

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

Bias-variance decomposition

Reminder : we still consider the quadratic risk.

Let $\hat{\eta}$ be an estimator of $\eta = g(\theta)$ st $\mathbb{E}_\theta\left(\|\hat{\eta}\|^2\right) < +\infty$, $\forall \theta \in \Theta$.

Proposition : Bias-variance decomposition (scalar case)

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

$$R_\theta\left(\hat{\eta}\right) = \mathbb{E}_\theta\left((\hat{\eta} - g(\theta))^2\right) = \mathbb{V}_\theta\left(\hat{\eta}\right) + \mathsf{b}_\theta(\hat{\eta})^2.$$

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

$$R_\theta\left(\hat{\eta}\right) = \mathbb{E}_\theta\left(\|\hat{\eta} - g(\theta)\|^2\right) = \mathrm{tr}\left(\mathbb{V}_\theta\left(\hat{\eta}\right)\right) + \|\mathsf{b}_\theta(\hat{\eta})\|^2,$$

where $\mathbb{V}_\theta\left(\hat{\eta}\right)$ is the covariance matrix of $\hat{\eta}$.

Example 1 : risk of some estimators

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

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

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

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

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

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

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

Admissible estimators

Definition : order relation on the set of estimators

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

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

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

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

Remarks

- The relation "is preferable to" is a partial order.
- 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.

Example 1 (cont'd)

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

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

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

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

- $\hat{\mu}_1$ is strictly preferable to $\hat{\mu}_4$, therefore $\hat{\mu}_4$ is not admissible.
- $\hat{\mu}_1$, $\hat{\mu}_2$, et $\hat{\mu}_3$ are pairwise incomparable.
- It can be proved that all three are admissible.
  Exercise : Prove that $\hat{\mu}_2$ is admissible.

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

Motivation

We will present in this section a lower bound of the form

$$\mathbb{V}_\theta(\hat{\eta}) \geq v_{\min}(\theta), \qquad \forall \theta \in \Theta,$$

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

Remark : for an UE, $R_\theta(\hat{\eta}) = \mathbb{V}_\theta(\hat{\eta})$.

Usefulness 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 (rare situation).
3. Prove that a given UE is nearly optimal.

Regularity condition C_1

Dominated model : there exists a (σ-finite) measure ν on $(\underline{\mathcal{X}}, \underline{\mathscr{A}})$ st

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

Regularity condition C_1

The densities f_θ share a common support : $\exists \mathcal{S} \in \underline{\mathscr{A}}$,

$$\forall \theta \in \Theta, \quad f_\theta(\underline{x}) > 0 \;\Leftrightarrow\; \underline{x} \in \mathcal{S}.$$

Remarks :

- $\mathcal{S}$ is only defined up to a ν-négligible set (as pdf's are).
- Strictly speaking, the « support » of the measure is the closure of $\mathcal{S}$.

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. $\mathscr{N}(\mu, \sigma^2)$: 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 !

Another regularity condition

We assume that C_1 holds.

Regularity condition C_2

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

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

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

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

Score

Definition / property : score

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

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

Then

i) We call score the random vector $S_\theta = S_\theta(\underline{X})$.

ii) C_2-iii $\Leftrightarrow \forall \theta \in \Theta$, the score S_θ 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.

The score is centered (proof)

Notice that

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

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

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

Finally,

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

Example 2

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

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

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

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 :

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

Remark on condition C_2-iii

Recall C_2-iii : $\forall \theta \in \Theta$,

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

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

Two approaches are available to check this condition :

1. Compute explicitely $\mathbb{E}_\theta(S_\theta) = \int_{\mathcal{S}} \nabla_\theta f_\theta(\underline{x})\, \nu(\mathrm{d}\underline{x})$.
2. Use a domination condition : show that $\forall \theta_0 \in \Theta$, $\exists \mathscr{V} \subset \Theta$, neighboorhood of θ_0, and a ν-integrable function $g : \underline{\mathcal{X}} \to \mathbb{R}$ st

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

Cramér-Rao inequality (scalar case)

Consider a statistical model where C_1 and C_2 hold.

Let $\hat{\eta}$ be an estimator of $\eta = g(\theta) \in \mathbb{R}$ st $\mathbb{E}_\theta\left(\hat{\eta}^2\right) < +\infty$, $\forall \theta \in \Theta$.

Definition : regular estimator

$\hat{\eta}$ is said to be regular if $\theta \mapsto \mathbb{E}_\theta\left(\hat{\eta}\right)$ is differentiable, with

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

Theorem / definition : Cramér-Rao inequality

If $\hat{\eta}$ is regular unbiased estimator, then $\forall \theta \in \Theta$

$$R_\theta\left(\hat{\eta}\right) = \mathbb{V}_\theta\left(\hat{\eta}\right) \;\geq\; \nabla g(\theta)^\top\, \mathbb{V}_\theta\left(S_\theta\right)^{-1} \nabla g(\theta).$$

Moreover, $\hat{\eta}$ is said to be efficient if the lower bound is met.

Proof

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

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

$$\mathbb{V}_\theta\left(\hat{\eta} - a^\top S_\theta\right) = \mathbb{V}_\theta\left(\hat{\eta}\right) - 2a^\top c + a^\top\,\mathbb{V}_\theta\left(S_\theta\right) a \;\geq\; 0.$$

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

$$\mathbb{V}_\theta\left(\hat{\eta}\right) - c^\top\,\mathbb{V}_\theta\left(S_\theta\right)^{-1} c \geq 0.$$

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

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

$\square$

Fisher information (scalar case)

We still assume that C_1 and C_2 hold.

Definition : Fisher information

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

$$I_{\underline{X}}(\theta) = \mathbb{V}_\theta(S_\theta(\underline{X})) = \mathbb{E}_\theta\left(S_\theta(\underline{X})\, S_\theta(\underline{X})^\top\right)$$

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 : $\mathbb{V}_\theta(\hat{\eta}) \geq \frac{1}{n} \nabla g(\theta)^\top I_1(\theta)^{-1} \nabla g(\theta)$.

Proof

Notice that the score is additive in an IID sample :

$$S_\theta(\underline{X}) = \sum_{i=1}^{n} S_\theta(X_i)$$

and thus

$$\mathbb{V}_\theta\left(S_\theta(\underline{X})\right) = \sum_{i=1}^{n} \mathbb{V}_\theta\left(S_\theta(X_i)\right) = n\ \mathbb{V}_\theta\left(S_\theta(X_1)\right)$$

since $S_\theta(X_1), \ldots, S_\theta(X_n)$ are IID. $\square$

Example 1 : estimation of μ

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

- $\hat{\mu}_n = \bar{X}_n$ is the MLE of μ ;
- $\hat{\mu}_n$ is unbiased and $R_\theta(\hat{\mu}_n) = \mathbb{V}_\theta(\hat{\mu}'_n) = \frac{\sigma^2}{n}$.

Exercise : the Fisher information matrix in this model is

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

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

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

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

Example 1' : estimation of σ^2

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

Exercise : show that

- the MLE $S_n^2 = \frac{1}{n} \sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ is biased ;
- $\sigma_n^2 = (S_n')^2 = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ is an UE of σ^2.

Lenghty computations (or Cochran's theorem) allow to get :

$$\mathbb{V}_\theta \left(\hat{\sigma}_n^2\right) = \frac{2\,\sigma^4}{n-1},$$

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

$$\mathbb{V}_\theta \left(\hat{\sigma}_n^2\right) > \frac{2\,\sigma^4}{n}.$$

(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.)

Exercise solution

Let us show that the sample variance S_n^2 is biased :

$$\mathbb{E}_\theta(S_n^2) = \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right) = \mathbb{E}_\theta\left(X_1^2\right) - \mathbb{E}_\theta\left(\bar{X}_n^2\right)$$
$$= \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

We conclude that the "corrected" sample variance is unbiased :

$$\mathbb{E}_\theta((S_n')^2) = \frac{n}{n-1}\,\mathbb{E}_\theta(S_n^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2. \quad \square$$

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

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) $\rightarrow$ approximate properties

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

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

$$\mathscr{M}_n = \left(\mathcal{X}^n, \mathscr{A}^{\otimes n}, \{\mathrm{P}_\theta^{\otimes n}, \theta \in \Theta\}\right),$$

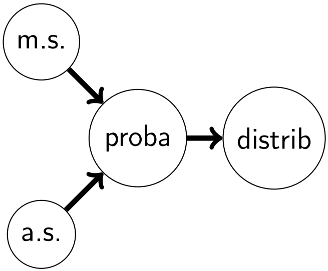
that we instantiate on a common underlying probability space $(\Omega, \mathscr{F})$.

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 :



Probability refresher : convergence modes

- almost sure convergence :

$$T_n \xrightarrow{\text{ps}} T \qquad \text{if} \qquad \mathbb{P}\left(T_n \to T\right) = 1$$

- convergence in L^2 (in mean square) :

$$T_n \xrightarrow{L^2} T \qquad \text{if} \qquad \mathbb{E}\left(\|T_n - T\|^2\right) \to 0$$

$$\text{iff} \qquad \forall j \leq p, \quad T_n^{(j)} \xrightarrow{L^2} T^{(j)}$$

- convergence in probability :

$$T_n \xrightarrow{\mathrm{P}} T \qquad \text{if} \qquad \forall \varepsilon > 0, \quad \mathbb{P}\left(\|T_n - T\| \geq \varepsilon\right) \to 0$$

- convergence in distribution :

$$T_n \xrightarrow{\text{loi}} T \qquad \text{if} \qquad \forall \varphi, \quad \mathbb{E}\left(\varphi(T_n)\right) \to \mathbb{E}\left(\varphi(T)\right),$$

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

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, $\forall \theta \in \Theta$,

$$\hat{\eta}_n \xrightarrow[n\to\infty]{\mathbb{P}_\theta} g(\theta).$$

(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, $\forall \theta \in \Theta$,

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

Remark : the work « convergent » is sometimes used instead of « consistent ».

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 $\mathbb{E}\left(\|X_1\|\right) < +\infty$, then

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

Law of large numbers in L^2

If the X_k's are IID and $\mathbb{E}\left(\|X_1\|^2\right) < +\infty$, then

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

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

Consistency : examples

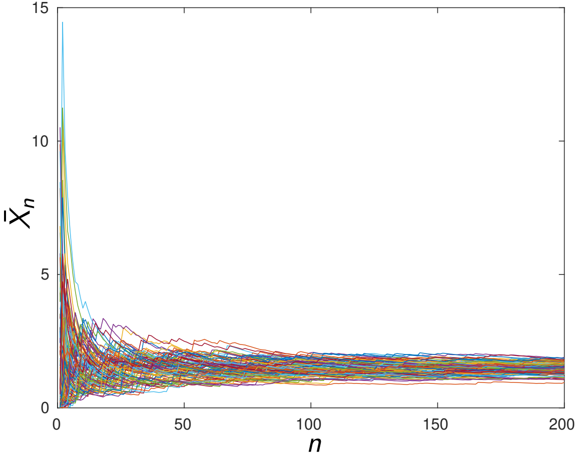
A) IID n-sample with finite first order moment

- 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 $\eta = \mathbb{E}_\theta(X_1)$.

Consistency : examples (cont'd)



Convergence of $\bar{X}_n$ to the true mean

(for a Gamma n-sample with true mean $\mu = 1.5$)

Consistency : examples (cont'd)

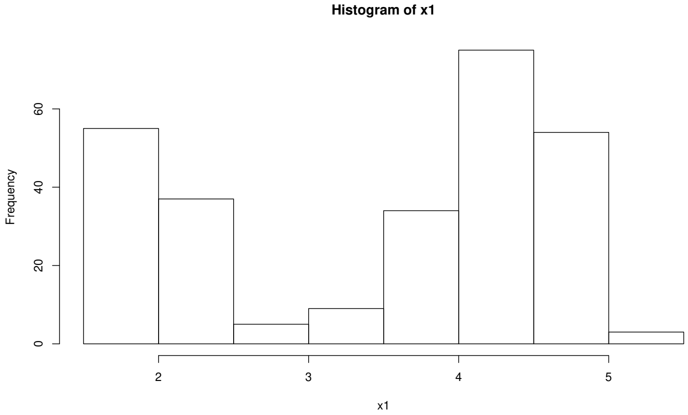
C) IID n-sample (with any distribution)

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

Application : histograms

- Let $\mathcal{X} = \cup_{k=1}^{K} A_k$ denote a partition of $\mathcal{X}$
- vector-valued $\hat{\eta}_n$: $\hat{\eta}_n^{(k)} = \frac{1}{n}\,\mathrm{card}\left\{i \leq n \mid X_i \in A_k\right\}$
- $\hat{\eta}_n$ is a strongly and mean-square consistent estimator of $\eta = \left(\mathbb{P}_\theta\left(X_1 \in A_k\right)\right)_{1 \leq k \leq K}$.

Consistency : examples (cont'd)



Example of a (un-normalized) histogram

Consistency : examples (cont'd)

D) Maximum of a uniform IID n-sample

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

  (Hint : start by proving almost sure consistency, using the Borel-Cantelli criterion : if $\sum_k \mathbb{P}\left(|Z_n - Z| > \varepsilon\right) < +\infty$ for all $\varepsilon > 0$,
  then $Z_n \xrightarrow{\text{a.s.}} Z$; then deduce form this that m.s. consistency holds as well.)

E) Maximum likelihood estimator

- see below

Exercise solution

Let us first prove almost sure consistency. We have, $\forall \varepsilon \leq \eta$,

$$\begin{aligned}\mathbb{P}\left(|\hat{\eta}_n - \eta| > \varepsilon\right) &= \mathbb{P}\left(\hat{\eta}_n < \eta - \varepsilon\right) \\ &= \mathbb{P}\left(\forall i \leq n,\ X_i < \eta - \varepsilon\right) = \left(\frac{\varepsilon}{\eta}\right)^n,\end{aligned}$$

therefore $\sum_n \mathbb{P}\left(|\hat{\eta}_n - \eta| > \varepsilon\right) < +\infty$, which implies that $\hat{\eta}_n \xrightarrow{\text{a.s.}} \eta$ using the Borel-Cantelli criterion.

Mean-square consistency (and also consistency in L^q for any $q \geq 1$) follows by application of the monotone convergence theorem, since the $\hat{\eta}_n$'s form an increasing sequence of positive functions. $\square$

(For the second part, we could also have applied the dominated convergence theorem.)

Asymptotically unbiased estimator

Recall that $\mathrm{b}_\theta(\hat{\eta}) = \mathbb{E}_\theta(\hat{\eta}) - g(\theta)$.

Definition : asymptotically unbiased

We will say that an estimator $\hat{\eta}_n$ is asymptotically unbiased if

$$b_\theta(\hat{\eta}) \xrightarrow{n\to+\infty} 0, \quad \forall\theta \in \Theta.$$

Proposition

$\hat{\eta}_n$ is consistent in the mean-square sense if, and only if, the two following conditions met :

i) $\hat{\eta}_n$ is asymptotically unbiased,

ii) $\mathbb{V}_\theta(\hat{\eta}_n) \to 0$, for all $\theta \in \Theta$. $\quad$ ($\mathrm{tr}\,(\mathbb{V}_\theta(\hat{\eta})) \to 0$ in the vector case)

Proof : Use the bias-variance decomposition ! $\square$

Asymptotically unbiased estimator : example

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}_{[0,\theta]}$, and we want to estimate θ.

Let us prove that $\hat{\theta}_n = \max_{i \leq n} X_i$ is asymptotically unbiased.

Method 1 : direct computation

- Compute the expectation : $\mathbb{E}_\theta(\hat{\theta}_n) = \frac{n}{n+1}\,\theta$ (cf. TD),
- hence the bias : $b_\theta(\hat{\theta}) = -\frac{\theta}{n+1} \to 0$.

Method 2 : dominated convergence theorem

- We already know that $\hat{\theta}_n$ is strongly consistent ;
- besides $\left|\hat{\theta}_n\right| \leq \theta, \;\; \mathbb{P}_\theta -$ a.s. ;
- therefore $\mathbb{E}_\theta\left(\hat{\theta}_n\right) \to \theta$ by the dominated convergence theorem.

Consistency of the MLE

The MLE minizes 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 $c = \mathrm{cov}_\theta\left(S_\theta,\, \hat{\eta}\right) \in \mathbb{R}^p$. Then, $\forall \theta \in \Theta$,

$$\gamma_n(\theta) - \gamma_n(\theta_\star) = \frac{1}{n} \sum_{k=1}^{n} \ln \frac{f_{\theta_\star}(X_i)}{f_\theta(X_i)} \xrightarrow[\text{ps}]{n \to +\infty} \int_{\mathcal{S}_1} \ln \frac{f_{\theta_\star}(x)}{f_\theta(x)}\, f_{\theta_\star}(x)\, \nu_1(\mathrm{d}x).$$

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

Definition / property : Kullback-Leibler divergence

$$D_{\mathsf{KL}}\left(f_{\theta_\star} || f_\theta\right) = \int_{\mathcal{S}_1} \ln \frac{f_{\theta_\star}(x)}{f_\theta(x)}\, f_{\theta_\star}(x)\, \nu_1(\mathrm{d}x) \geq 0$$

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 $\Delta(\theta_\star, \theta) = D_{\mathrm{KL}}\left(f_{\theta_\star} || f_\theta\right)$.

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

Theorem : Consistency of the MLE

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

i) $\sup_{\theta \in \Theta} |\Delta_n(\theta_\star, \theta) - \Delta(\theta_\star, \theta)| \xrightarrow[n \to +\infty]{\mathbb{P}_{\theta_\star}} 0$

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.