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)

J. Bect & L. Le Brusquet — 1A — Statistics and Learning

Lecture 1/9

Introduction and point estimation methods

In this lecture you will learn how to...

- Introduce statistical inference and illustrate its usefulness
- Define the mathematical framework
- Present some commonly used estimation methods

Lecture outline

Lecture outline

One word, several meanings...

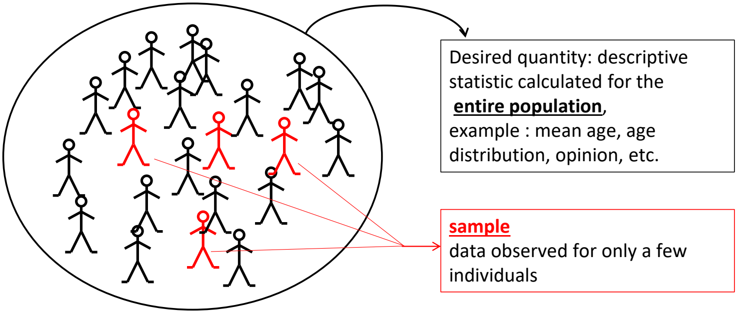
- One (or several) statistic(s) : numerical indicators, often simple, computed from data.

  Examples : average, standard deviation, median, etc....

- statistics : a mathematical discipline which has several branches, including
  - descriptive statistics,
  - statistical inference (part 1 of this course),
  - design of experiments,
  - statistical learning (part 2 of this course),
  - ...

Remark : a mathematical definition of the word "statistic" (first meaning) will be given later.

Historical example : the opinion survey case

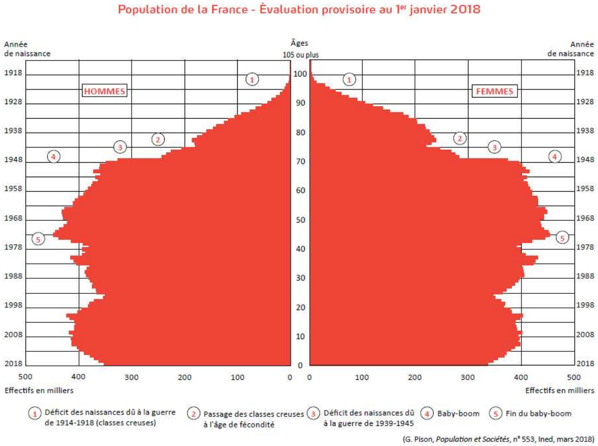


A descriptive statistic may be calculated on :

- the entire population $\rightarrow$ quantity of interest
- a sample $\rightarrow$ "approximate" value (sense to be defined)

To infer = to draw conclusions about a population from data collected for a sample

Demographic statistics (census)



Descriptive statistics are useful to “explore” data sets

Typical goals : obtain numerical summaries (of small dimension) and/or easily interpretable visualizations.

Other example : estimation of a proportion

Context. Consider a box with W white balls and R red balls, where W and R are unknown.

Goal. Estimate the proportion $\theta = \frac{W}{W+R}$ of white balls.

Data (observations). We perform n draws with replacement

➠ for the i-th draw, $x_i = 1$ if the ball is white, 0 otherwise.

Steps to estimate θ

1. **statistical modeling**
   x_i realization of a RV X_i, with $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$, $0 \leq \theta \leq 1$
2. **inference** (here, estimation)
   using the data $\underline{x} = (x_1, \ldots, x_n)$ and the statistical model.
   ➠ Consider $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$ (a possible descriptive statistic)
   ➠ Is it reasonable to use it a “substitute” for the unknown θ ?

Relation between statistical inference and probability theory

Probability theory provides the foundation for statistical inference :

- probability theory : a probability space is given ;
- statistical inference : several probabilistic models are assumed possible ; we want to extract (from data) information from data about the underlying probability measure.

Illustration on the "box" example :

	Probability (W and R known)	Inference (W and R unknown)
typical questions	• distribution of the number of white balls after n draws ; • distribution of the number of draws to get the first white ball	• estimate θ ; • give an interval containing θ ; • decide whether $\theta \leq 0.5$ or not.
type of conclusions	certain	for finite n, impossible to answer with certainty

Application fields & examples of statistical questions

Many fields of application :

- **Healthcare :** identify biomarkers responsible for a disease from data collected on cohorts.
- **Environment, safety :** estimate the probability of risk from measurement data.
- **Industry :** control the quality of a production line from data collected for only a few elements.
- **Opinion survey :** predict the winner of an election from a survey, quantify the uncertainty about the prediction.
- **Insurance :** evaluate the risk of ruin for an insurance company facing a disaster.

Lecture outline

From data to random variables

Data (observations)

Let $\underline{x} \in \underline{\mathcal{X}}$ denote the data that must be analyzed. For instance :

1. a scalar quantity, measured on n objects/individuals :
   ➠ $\underline{x} = (x_1, \ldots, x_n), \quad x_i \in \mathbb{R}, \quad \underline{\mathcal{X}} = \mathbb{R}^n$;
2. d scalar quantities, potentially of different natures, measured on n objects/individuals :
   ➠ $\underline{x} = (x_1, \ldots, x_n), \quad x_i \in \mathbb{R}^d, \quad \underline{\mathcal{X}} = \mathbb{R}^{n \times d}$;
3. any dataset of a more complex nature
   (times series, symbolic data, graphs, etc.).

The data is modeled, a priori, by a random variable (RV) $\underline{X}$
➠ $\underline{x}$ is considered as a realization of $\underline{X}$.

Statistical model

The observation space $(\underline{\mathcal{X}}, \underline{\mathscr{A}})$

It is the measurable space in which $\underline{X}$ takes its values.
Most of the time, we will use :

- $\underline{\mathcal{X}} = \mathbb{R}^n$ with $\underline{\mathscr{A}} = \mathscr{B}(\mathbb{R}^n)$
- or, more generally, $\underline{\mathcal{X}} = \mathbb{R}^{n \times d}$ with $\underline{\mathscr{A}} = \mathscr{B}\left(\mathbb{R}^{n \times d}\right)$.

Statistical modeling

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space carrying :

- the observed random variable $\underline{X}$,
- any other (unobserved) RV that we might need.

The probability $\mathbb{P}$ is not perfectly known : we consider a

- set $\mathscr{P}$ of probability distributions sur $(\Omega, \mathscr{F})$

Statistical model (cont'd)

Distribution of the observations

Let $\mathbb{P}^{\underline{X}}$ denote the distribution of $\underline{X}$ when $\mathbb{P} \in \mathscr{P}$ is the underlying probability measure.

➠ We have a set $\mathscr{P}^{\underline{X}} = \left\{\mathbb{P}^{\underline{X}}, \mathbb{P} \in \mathscr{P}\right\}$ of possible distributions.

Definition : Statistical model

Formally, we call statistical model the triplet

$$\mathscr{M} = \left(\underline{\mathcal{X}},\ \underline{\mathscr{A}},\ \mathscr{P}^{\underline{X}}\right).$$

Remarks :

- We can construct several models $(\Omega, \mathscr{F}, \mathscr{P}, \underline{X})$ for a given $\mathscr{M}$.
- In particular, when we only care about the observed RV $\underline{X}$, we can work on the *canonical* model : $\Omega = \underline{\mathcal{X}}$, $\mathscr{F} = \underline{\mathscr{A}}$, $\mathscr{P} = \mathscr{P}^{\underline{X}}$, $\underline{X} = \mathrm{Id}_{\underline{\mathcal{X}}}$.

Statistical inference

Reminder : the data $\underline{x} \in \underline{\mathcal{X}}$ is seen as a realization of $\underline{X} \sim \mathbb{P}^{\underline{X}}$, for a certain (unknown) probability $\mathbb{P} \in \mathscr{P}$.

The goal of statistical inference

Goal : to construct procedures allowing to extract information about $\mathbb{P}^{\underline{X}}$ from

- one realization of $\underline{X}$,
- the knowledge of the set $\mathscr{P}^{\underline{X}}$ of all possible distributions.

Important

Since the true probability $\mathbb{P}$ is unknown, we must design statsitical procedures that are "applicable" to **any** probability $\mathbb{P} \in \mathscr{P}$.

Family of distributions

The set $\mathscr{P}$ est represented by a parameterized family :

$$\mathscr{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}\,.$$

Parametric model

If Θ is finite-dimensional, the model is called parametric.

- the parameter vector θ is often of small size.
- we will denote by p the number of parameters ($\Theta \subset \mathbb{R}^p$).

Example. Family of (scalar) Gaussian distributions

$$\mathscr{P}^{\underline{X}} = \left\{\mathscr{N}(\mu,\sigma^2),\quad \mu \in \mathbb{R},\quad \sigma^2 \in \mathbb{R}_*^+\right\}$$

Assumptions on the family of distributions

Dominated model

The model

$$\mathscr{M} = \left(\underline{\mathcal{X}},\ \underline{\mathscr{A}},\ \left\{\mathbb{P}_\theta^{X},\ \theta \in \Theta\right\}\right)$$

is said to be dominated if there exists a (σ-finite) measure ν on $(\underline{\mathcal{X}}, \underline{\mathscr{A}})$ such that

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

➟ f_θ is the density of $\mathbb{P}_\theta^{X}$ with respect to ν.

In this course, we will consider the following cases :

- "continuous" RV : reference measure ν = Lebesgue's measure,
- discrete RV : reference measures ν = counting measure.

Assumptions on the family of distributions (cont'd)

Identifiable model

The model

$$\mathscr{M} = \left(\underline{\mathcal{X}},\ \underline{\mathscr{A}},\ \left\{\mathbb{P}_\theta^{\mathsf{X}},\ \theta \in \Theta\right\}\right)$$

is identifiable if the mapping $\theta \mapsto \mathbb{P}_\theta^{\mathsf{X}}$ is injective.

In the rest of this course, all the models will be

- dominated by a reference measure ν,
- identifiable.

Sampling models

n-sample

If $\underline{X} = (X_1, \ldots, X_n)$ is such that :

- the X_i's are (mutually) independent,
- all the X_i's have the same distribution P,

then the X_i's are called independent et identically distributed (iid) and we say that $\underline{X}$ is an (iid) *n*-sample.

Distribution of an *n*-sample.
Consider the model that describes each of the X_i's individually :

- $(\mathcal{X}, \mathscr{A}, \{\mathrm{P}_\theta, \theta \in \Theta\})$

Then we have :

- $(\underline{\mathcal{X}}, \underline{\mathscr{A}}) = (\mathcal{X}^n, \mathscr{A}^{\otimes n})$ (product space),
- $\forall \theta \in \Theta,\ \mathbb{P}_\theta^{\underline{X}} = \mathrm{P}_\theta^{\otimes n}$ (product distribution).

Example : component reliability

This application will be used as an illustration in several lectures.

Context

- We are interested in the reliability of components from a production line.
- Reliability : measured by the lifetime of the components.
- Data (observations) : a sample of $n = 10$ components, for which the lifetime has been recorded : $\underline{x} = (x_1, \ldots, x_n)$.

Modeling

- Each x_i is modeled by a scalar RV X_i.
- The X_i's are assumed iid, with values in $(\mathcal{X}, \mathscr{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Example : component reliability

Modeling (cont'd) : family of distributions

Typical* assumption for the lifetime of a component :

$$X_1 \sim \mathcal{E}(\theta), \quad \theta > 0.$$

Hence the statistical model for one observation :

$$(\mathbb{R},\ \mathcal{B}(\mathbb{R}),\ \{\mathcal{E}(\theta), \theta > 0\}).$$

Note : this assumption on X_1 holds for all the X_i's, $i \geq 1$.

Density. The exponential distribution $\mathcal{E}(\theta)$ has the density :

$$f_\theta(x) = \theta \exp(-\theta x)\, \mathbb{1}_{[0,\infty[}(x).$$

* in the case of unpredictable failures, not related to the age of the component

Example : component reliability

A few problems of (statistical) interest

- **estimate** θ, or
- **estimate** $\eta = \frac{1}{\theta} = \mathbb{E}(X_1)$ (average lifetime)

  ➠ lectures #1 et #2
- provide **confidence intervals** for θ and η

  ➠ lecture #3
- **estimate** θ given prior information on its value (e.g., provided by the manufacturer of the production line)

  ➠ lecture #4 on Bayesian estimation
- **test the hypothesis** $\eta \leq 10$, in order to assess the value of an optional warranty extension

  ➠ lecture #5 on hypothesis testing

Data.

0.5627	16.1121	5.4943	7.9374	1.2658
2.9885	8.6266	43.8877	2.1641	8.9138

Table – Measured values (arbitrary units) for a sample of size $n = 10$

Estimating η **:** a first estimtor

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}_{\theta}(X_1) = \eta \quad \text{(SLLN)}.$$

➠ $\hat{\eta}^{(1)} = \bar{X}$ seems to be a reasonable "estimator" of η.

Numerical application $\quad \hat{\eta}^{(1)} = 10.1960$

Notations / vocabulary

Notations. We will often use notations such as

- $\mathbb{E}_{\theta}(.)$ (expectation),
- $\mathbb{V}_{\theta}(.)$ (variance ou covariance matrix),
- $f_{\theta}(.)$ (density), ...

to indicate that theses operators or functions depend on a probability $\mathbb{P}_{\theta}$ for a particular value of θ.

Definition : Statistic

A statistic is a random variable (often scalar- or vector-valued) that can be computed from $\underline{X}$ alone*.

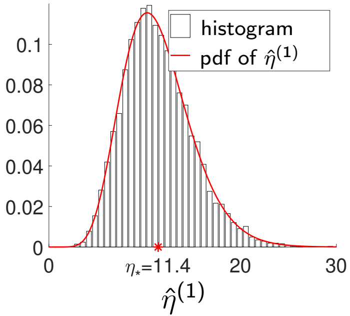
Example : the estimator $\hat{\eta}^{(1)} = \bar{X}$ is a statistic.

* Technically : can be written as a measurable function of $\underline{X}$.
In particular, depends neither on other (unobserved) RVs nor on θ.

Numerical assessment of the performance of $\hat{\eta}^{(1)}$

With numerical simulations, (almost) everything is possible !

- we choose a particular value of η (here, $\eta_* = 11,4$), then
- we simulate on a computer a large number m of n-samples (here, $m = 10000$).



Remarks

- Our estimates are, in this case, not very accurate.
- Providing confidence intervals would be very relevant here.
- In this simple we can compute the density of $\hat{\eta}^{(1)}$ analytically.

A few words on the Gamma distribution $\Gamma(p, \lambda)$

Let $X \sim \Gamma(p, \lambda)$, $p > 0$, $\lambda > 0$). Its pdf is

$$f(x) = \frac{\lambda}{\Gamma(p)} x^{p-1} \exp(-\lambda x) \mathbb{1}_{\mathbb{R}^+}(x).$$

Moments

- mean : $\mathbb{E}_\theta(X) = \frac{p}{\lambda}$
- variance : $\mathbb{V}_\theta(X) = \frac{p}{\lambda^2}$

Particular cases

- $\mathcal{E}(\lambda) = \Gamma(p = 1, \lambda)$
- $\Gamma(p = n, \lambda = \frac{n}{2}) = \chi^2(n)$

Properties

- Let $a > 0$. If $X \sim \Gamma(p, \lambda)$, then $aX \sim \Gamma\left(p, \frac{\lambda}{a}\right)$.
- If $X \sim \Gamma(p, \lambda)$, $Y \sim \Gamma(q, \lambda)$, and X and Y are independent, then $X + Y \sim \Gamma(p + q, \lambda)$.

Exercise. Show that $\hat{\eta}^{(1)} \sim \Gamma\left(n, \frac{n}{\eta}\right)$.

$\hat{\eta}^{(2)}$: another estimator.

With a convergence argument similar to the one used earlier :

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}_\theta\left(X_1^2\right) = \frac{2}{\theta^2} = 2\eta^2,$$

therefore using $\hat{\eta}^{(2)} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} X_i^2}$ seems “reasonable” as well.

Numerical application $\hat{\eta}^{(2)} = 11.2228$

Questions

- How can we compare two estimators ?
- If there an estimator that is “better” than the others ?
- How to construct “good” estimators ?

Lecture outline

Mathematical framework

In this section :

- we consider a statistical model

$$\mathscr{M} = \left(\underline{\mathcal{X}}, \underline{\mathscr{A}}, \left\{\mathbb{P}_\theta^{\underline{X}}, \theta \in \Theta\right\}\right),$$

  most of the time assumed to be parametric ($\Theta \subset \mathbb{R}^p$) ;

- when $\underline{X}$ is an IID n-sample, we write
  - $\underline{X} = (X_1, \ldots, X_n)$
  - $\underline{\mathcal{X}} = \mathcal{X}^n$, with $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathbb{R}^d$,
  - $\mathbb{P}_\theta^{\underline{X}} = \mathrm{P}_\theta^{\otimes n}$;

- we want to estimate a "quantity of interest" :
  - either θ itself ($\Rightarrow$ parametric model),
  - or, more generally, $\eta = g(\theta)$.

Lecture outline

The substitution method

Assume that

- we already have an estimator $\hat{\eta}$ of $\eta = g(\theta)$
- and we want to estimate another quantity of interest η' that can be written as $\eta' = h(\eta)$, with h a continuous function.

The substitution method

The substitution method consists in using

$$\hat{\eta}' = h\left(\hat{\eta}\right) \text{ as an estimator of } \eta.$$

Example : component reliability

Reminder : $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{E}(\theta), \quad \theta > 0.$

We are interested in the probability that a failure occurs before t_0 :

$$\Rightarrow \quad \eta' = \mathbb{P}_\theta\left(X_1 \leq t_0\right) = \int_0^{t_0} \theta \exp(-\theta x)\mathrm{d}x$$
$$= 1 - \exp(-\theta t_0) = 1 - \exp\left(-\frac{t_0}{\eta}\right).$$

Using $\hat{\eta}^{(1)} = \bar{X}$ as an estimator of η, we get

$$\hat{\eta}' = 1 - \exp\left(-\frac{t_0}{\bar{X}}\right).$$

Empirical measure

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathbb{P}^{X_1}$.

Recall the Dirac measure at $x \in \mathcal{X}$:

$$\forall A \in \mathscr{A}, \quad \delta_x(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Definition : empirical measure

The empirical measure is the (random) measure defined by :

$$\hat{\mathbb{P}}^{X_1} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Usefulness : the empirical measure can be seen as an estimator of $\mathbb{P}^{X_1}$ ➠ allows us to construct of other estimators using the substitution method.

Example : estimator of the k-th order moment

Assume $X_1 \in L^k$. Then

$$m_k = \mathbb{E}\left(X_1^k\right) = \mathscr{G}\left(\mathbb{P}^{X_1}\right)$$

is well defined, with $\mathscr{G}(\mu) = \int_{\mathcal{X}} x^k \mu(\mathrm{d}x)$. By substitution :

$$\hat{m}_k = \mathscr{G}\left(\hat{\mathbb{P}}^{X_1}\right) = \int_{\mathcal{X}} x^k \, \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Similar example : the sample variance. If $X_1 \in L^2$ and $\eta' = \mathbb{V}(X_1) = \mathscr{G}\left(\mathbb{P}^{X_1}\right)$, where $\mathscr{G}(\mu) = \int_{\mathcal{X}} x^2 \mu(\mathrm{d}x) - \left(\int_{\mathcal{X}} x\mu(dx)\right)^2$, we get by substitution :

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \qquad \text{(sample variance).}$$

One last example : the empirical cdf

Let $x \in \mathbb{R}$. The cumulative distribution function (cdf) of X_1 at x is

$$F(x) = \mathbb{P}^{X_1}(X_1 \leq x) = \mathscr{G}_x\left(\mathbb{P}^{X_1}\right) \quad \text{with} \quad \mathscr{G}_x(\mu) = \int_{-\infty}^{x} \mu(\mathrm{d}x).$$

Hence the empirical cdf :

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

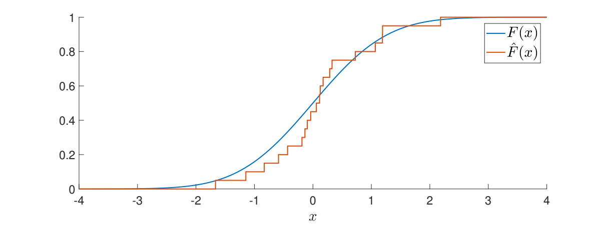


Figure – Empirical cdf for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $n = 20$.

Lecture outline

The method of moments

Assume that

- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{P}_\theta$, with $\theta \in \Theta$;
- most of the time assumed to be parametric : $\Theta \subset \mathbb{R}^p$,
- we want to estimate θ itself

Consider the function

$$\begin{array}{rccl} h: & \Theta \subset \mathbb{R}^p & \to & h(\Theta) \subset \mathbb{R}^p, \\ & \theta & \mapsto & h(\theta) = \begin{pmatrix} \mathbb{E}_\theta\left(X_1\right) \\ \vdots \\ \mathbb{E}_\theta\left(X_1^p\right) \end{pmatrix}. \end{array}$$

Remark : sometimes other moments can be used (not necessarily the first p).

The method of moments (cont'd)

Assume $h : \Theta \to h(\Theta)$ injective, and thus bijective.

The method of moments

The method of moments consists in

- estimating the first p moments $\hat{m}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$, $k \leq p$,
- then applying h^{-1} to construct an estimator of θ.

Hence moment-of-moments estimator : $\hat{\theta} = h^{-1}\left(\hat{m}_{1:p}\right)$, where

$$\hat{m}_{1:p} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^p \end{pmatrix}.$$

Remark : well defined only if $\hat{m}_{1:p} \in h(\Theta)$ $\mathbb{P}_\theta$-ps, pour tout θ.
Otherwise $\to$ minimization of some distance (generalized method of moments).

Method of moments : examples

Example : component reliability

We have $\mathbb{E}_\theta(X_1) = \theta^{-1}$ (exponential distribution), therefore

$$\theta = \left(\mathbb{E}_\theta(X_1)\right)^{-1} \quad \text{and} \quad \hat{\theta} = \left(\bar{X}\right)^{-1}.$$

Example : $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$

We have $h(\theta) = \begin{pmatrix} \mathbb{E}_\theta(X_1) \\ \mathbb{E}_\theta(X_1^2) \end{pmatrix} = \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix}$,

therefore $\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}_\theta(X_1) \\ \mathbb{E}_\theta(X_1^2) - \left(\mathbb{E}_\theta(X_1)\right)^2 \end{pmatrix}$,

and finally $\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \end{pmatrix}$

Exercise. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}_{[a,b]}$. Method-of-moments estimator of (a, b) ?

Lecture outline

Maximum likelihood estimation

Reminder : dominated model $\to \mathbb{P}_\theta^X$ admits a pdf f_θ.

Definition : likelihood

We call likelihood the function :

$$\begin{array}{rccc} \mathcal{L}: & \Theta \times \underline{\mathcal{X}} & \to & \mathbb{R}_+ \\ & (\theta; \underline{x}) & \mapsto & f_\theta\left(\underline{x}\right) \end{array}$$

Remark. Si $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{P}_\theta$, then $\mathcal{L}(\theta; \underline{x}) = \prod_{i=1}^n f_\theta\left(x_i\right)$.

(usual abuse of notation : here $f_\theta = f_\theta^{X_1}$)

Definition : MLE

If $\hat{\theta}$ is a maximizer of $\theta \mapsto \mathcal{L}\left(\theta; \underline{X}\right)$, then
$\hat{\theta}$ is a maximum likelihood estimator (MLE) of θ.

MLE : practical details

- Existence and uniqueness of the MLE are not guaranteed in general.
- For an IID n-sample, we often use the log-likelihood :

$$\ln \mathcal{L}(\theta; \underline{x}) = \sum_{i=1}^{n} \ln f_\theta(x_i).$$

- If $\mathcal{L}$ is twice differentiable, a necessary condition for $\hat{\theta}$ to be an MLE is :

$$\begin{cases} \left(\nabla_\theta \left(\ln \mathcal{L}\right)\right)\left(\hat{\theta}; \underline{X}\right) \; = \; 0, \\ \left(\nabla_\theta \nabla_\theta^\top \left(\ln \mathcal{L}\right)\right)\left(\hat{\theta}; \underline{X}\right) \text{ has negative eigenvalues.} \end{cases}$$

(locally concave function ; $\nabla_\theta \nabla_\theta^\top$ is the Hessian operator)

MLE example : component reliability

For $x_1, \ldots, x_n \geq 0$, we have $\mathcal{L}(\theta; \underline{x}) = \prod_{i=1}^{n} \theta \exp(-\theta x_i)$, and thus

$$\ln \mathcal{L}(\theta; \underline{x}) = n \ln(\theta) - \theta \sum_{i=1}^{n} x_i .$$

Stationarity condition ("likelihood equation")

$$\frac{\partial(\ln \mathcal{L})}{\partial \theta}(\theta; \underline{x}) = 0 \iff \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0.$$

➠ If $\sum_{i=1}^{n} x_i \neq 0$, the MLE exists and is equal to $\hat{\theta} = \left(\bar{X}\right)^{-1}$.

(we check that, at this point, $\frac{\partial^2(\ln \mathcal{L})}{\partial \theta^2}(\hat{\theta}; \underline{x}) = -\frac{n}{\hat{\theta}^2} < 0$)

Remark : the same estimator was obtained by the method of moments.

MLE example : Gaussian IID n-sample, $\theta = \left(\mu, \sigma^2\right)$

Same approach as in the previous example :

$$\ln \mathcal{L}\left(\theta; \underline{x}\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2},$$

$$\left(\nabla_\theta \ln \mathcal{L}\right)\left(\theta; \underline{x}\right) = \frac{n}{\sigma^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i - \mu \\ -\frac{1}{2} + \frac{1}{2\sigma^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}.$$

Solving the liklihood equation yields :

$$\hat{\theta} = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \end{pmatrix}$$

et we can check that $\left(\nabla_\theta \nabla_\theta^\top \ln \mathcal{L}\right)\left(\hat{\theta}; \underline{x}\right)$ is negative definite.

Remark : the same estimator was obtained by the method of moments.