CentraleSupélec

Statistics and Learning

Lecturers (alphabetic order):

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

† Course coordinator

Lecture 1/9

Introduction and point estimation methods

Course objectives

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

Lecture outline

Lecture outline

One word, several meanings...

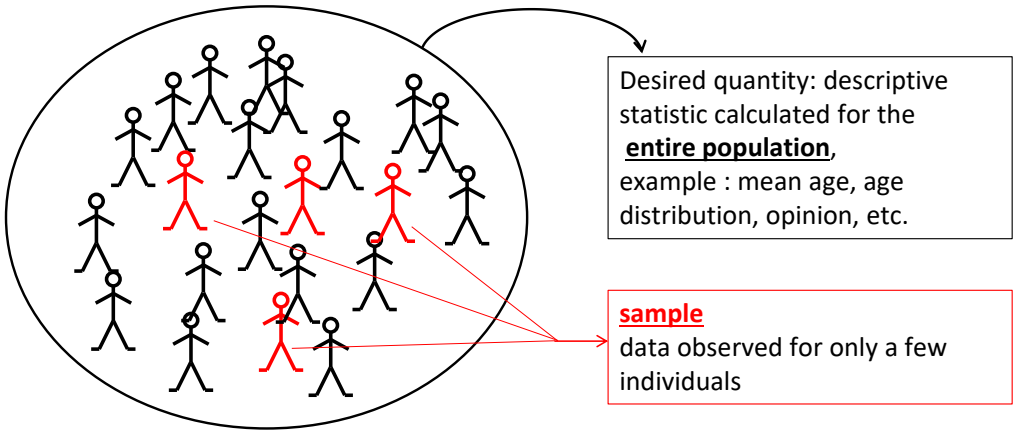
- One (or several) statistic(s): numerical indicators, often simple, computed from data.

  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

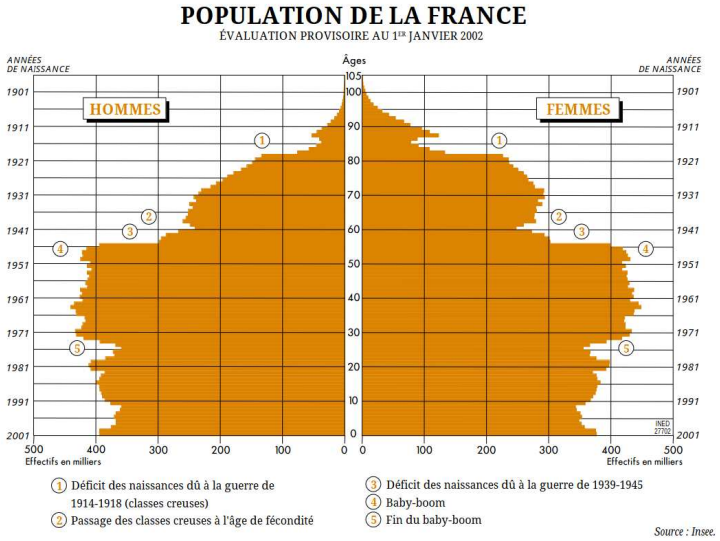


A descriptive statistic may be calculated on:

- the entire population → quantity of interest
- a sample → "approximate" value (to be defined)

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

Exhaustive census is not statistical inference



Descriptive statistics are useful to "explore" data sets

Goals: obtain numerical summaries (of small dimension)
and/or easily interpretable visualizations, etc.

Note: in France, for municipalities with more than 10,000 inhabitants, the systematic census has been replaced since 2004 by random (but not IID) sampling of addresses.

Another 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 $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, $0 \leq \theta \leq 1$
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 as 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 in general to answer with certainty

Example of questions addressed, in various fields

- **Healthcare:** identify biomarkers responsible for a disease using data collected from cohorts.
- **Insurance:** evaluate the risk of insolvency of an insurance company.
- **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.
- **Ecology:** estimate the size of a population of animals using partial observations (e.g., capture-mark-recapture).
- ...

Lecture outline

From data to random variables

Data (observations)

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

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 over $(\Omega, \mathscr{F})$ supposed to contain the "true" probability measure.

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 define a statistical model as 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 statistical procedures that are "applicable" to **any** probability $\mathbb{P} \in \mathscr{P}$.

Family of distributions

The set $\mathscr{P}$ is 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 Gaussian distributions on $\underline{\mathcal{X}} = \mathbb{R}$

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

(In this example we consider only one scalar observation.)

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_θ,

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 $(\mathcal{X}, \mathscr{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- $(\underline{\mathcal{X}}, \underline{\mathscr{A}}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

Example: component reliability

Modeling (cont'd): family of distributions

Typical* assumption for the lifetime of a component:

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

Hence the statistical model:

$$\left(\mathbb{R}^n,\ \mathcal{B}(\mathbb{R}^n),\ \{\mathcal{E}(\theta)^{\otimes n}, \theta > 0\}\right).$$

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

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

with respect to Lebesgue's measure

* 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 and 2
- provide **confidence intervals** for θ and η
  ➠ lecture 3
- **test the hypothesis** $\eta \leq 10$, in order to assess the value of an optional warranty extension
  ➠ lecture 4 on hypothesis testing
- **estimate** θ given prior information on its value (e.g., provided by the manufacturer of the production line)
  ➠ lecture 5 on Bayesian estimation

Example: component reliability (cont'd)

Data: a sample of size $n = 10$ [arbitrary unit]

0.5627	16.1121	5.4943	7.9374	1.2658
2.9885	8.6266	43.8877	2.1641	8.9138

Estimating η : a first estimator (see Lecture 2 for a definition)

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

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

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

Notations / vocabulary

Notations. We will often use notations such as

- $\mathbb{E}_\theta(.)$ (expectation),
- $\text{var}_\theta(.)$ (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*.

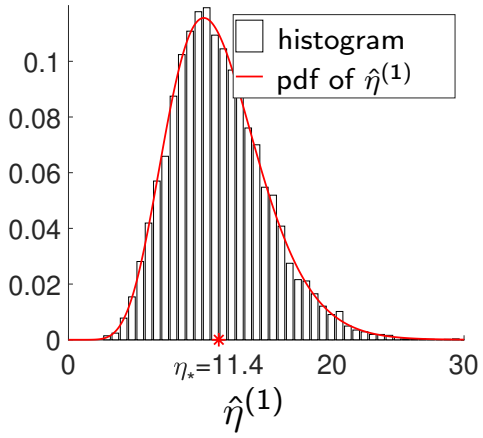
Example: the estimator $\hat{\eta}^{(1)} = \bar{X}$ is a statistic.

* 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 example we can compute the density of $\hat{\eta}^{(1)}$ analytically.

gamma distribution

$\hat{\eta}^{(2)}$: 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 ?
- Is there an estimator that is "better" than all 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,
  - 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

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

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:

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

Using $\hat{\eta}^{(1)} = \bar{X}$ as an estimator of $\eta = \frac{1}{\theta}$, 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:

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

Application: the empirical measure can be seen as an estimator of $\mathbb{P}^{X_1}$ ➠ allows us to construct 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:

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

Other examples:

- sample variance — exercise 3
- empirical cumultive distribution function — complement

Lecture outline

The method of moments

Assume that

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

Consider the function

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

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 : $\hat{\theta} = h^{-1}(\hat{m}_{1:p})$, 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, 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

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

Another example: Gaussian n-sample

✏ PC 1, Ex. 1.1

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

Considering the first two moments, we have:

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

Lecture outline

Likelihood function

Assume a dominated model: $\mathbb{P}_\theta^{\underline{X}}$ admits a pdf f_θ wrt a measure ν on $\underline{\mathcal{X}}$, for all $\theta \in \Theta$.

reminder: density

Definition: likelihood

We call likelihood the function:

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

We call log-likelihood the function $\ln \mathcal{L}$.

Remark. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{P}_\theta$, then,

$$\mathcal{L}(\theta; \underline{x}) = \prod_{i=1}^n f_\theta(x_i), \qquad \text{therefore} \quad \ln \mathcal{L}(\theta; \underline{x}) = \sum_{i=1}^n \ln f_\theta(x_i).$$

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

Maximum likelihood estimation

Definition: MLE

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

Remarks:

- Existence and uniqueness of the MLE: not guaranteed in general.
- Equivalently, $\hat{\theta}$ is a maximizer of $\theta \mapsto \ln \mathcal{L}(\theta; \underline{X})$.
- Assume $\Theta \subset \mathbb{R}^p$. If $\mathcal{L}$ is of class C^1 wrt θ on $\mathrm{int}(\Theta)$, a necessary condition for an interior point $\hat{\theta} \in \mathrm{int}(\Theta)$ to maximize the likelihood is:

$$\left(\nabla_\theta (\ln \mathcal{L})\right)(\hat{\theta}; \underline{X}) = 0.$$

This is called the likelihood equation.

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 $\sum_{i=1}^n x_i > 0$, unique solution in $\Theta = \mathbb{R}_+^*$ at $\theta = n \left(\sum_{i=1}^n x_i\right)^{-1}$.

✏ It is indeed a maximum of the likelihood function (cf. sign of the derivative).

➠ Since $\sum_{i=1}^n X_i > 0$ a.s., a unique MLE exists: $\hat{\theta} = \left(\bar{X}\right)^{-1}$.

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

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

Same approach as in the previous example.

1. First write the log-likelihood:

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

2. Solving the likelihood equation yields:

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

3. It can be proved that the maximum is indeed attained at this point.

✏ PC 1, Ex. 1.1

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

Summary and preview

We have seen and will practice in PC 1:

- the general framework of statistical inference,
- some classical methods for point estimtion.

We will cover in the next lecture:

- the quantitative assessment of an estimator's performance,
- the comparison of estimators,
- the asymptotic approach ($n \to \infty$).

Lecture outline

Lecture outline

Exercise 1 (Bernoulli model)

solution

Let $X_1, \ldots, X_n$ be an n-sample of binary observations, independent and identically distributed according to the Bernoulli $\mathrm{Ber}(p)$ distribution, with $p \in [0, 1]$.

Questions

1. Specify a formal statistical model $\mathscr{M} = \left(\underline{\mathscr{X}}, \underline{\mathscr{A}}, \mathscr{P}^{\underline{X}}\right)$ corresponding to this description.
2. Construct an estimator of p using the method of moments.
3. Construct an estimator of p using the maximum likelihood method.
4. Compute the expectation and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Exercise 2 (log-normal distribution)

solution

A bit of context

The association between childhood leukemia and exposure to very low-frequency magnetic fields (mainly due to electrical structures and appliances) is statistically significant for residential exposure averaged over 24 hours, with levels above $0.4\mu T$.

Source : anses

Modeling assumption. For dwellings located less than 50 meters from HV lines, residential exposure averaged over 24 hours follows a lognormal distribution.

log-normal distribution

Exercise 2 (log-normal distribution)

solution

Let

- $\underline{X} = (X_1, \ldots, X_n)$: n-sample with a log-normal $\mathscr{LN}(\mu, \sigma^2)$ distribution, where $\sigma^2 > 0$ is **known**.
- p_0 : probability that a RV following the $\mathscr{LN}(\mu, \sigma^2)$ distribution exceeds the threshold $s_0 = 0.4\mu$T.

Questions

1. Construct an estimator of μ using the maximum likelihood method.
2. Using the substitution method, derive an estimatoir of p_0.
3. Does the resulting estimator of p_0 converge almost surely? If so, to what limit?

Exercise 3 (sample variance)

solution

Let $X_1, \ldots, X_n$ be an n-sample of real-valued observations, independent and identically distributed, with a finite second order moment.

Let $\mathbb{M}$ denote the set of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with a finite second order moment.

Questions

1. Prove that $\text{var}(X_1) = \mathscr{G}\left(\mathbb{P}^{X_1}\right)$, where $\mathscr{G}$ is a function defined on $\mathbb{M}$, to be specified.
2. Using the substitution method, derive from $\mathscr{G}$ an estimator of the variance.
3. Study the convergence of the estimator when $n \to +\infty$.

back to slide ??

Lecture outline

Solution of exercise 1

back to questions

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

The "natural" (minimal) set to describe the values of a binary variable is $\mathcal{X} = \{0, 1\}$.

➠ $\underline{\mathcal{X}} = \{0, 1\}^n$ for an n-sample

On a finite or countable set, we use in general the discrete σ-algebra, i.e., the set of all subsets of $\underline{\mathcal{X}}$.

➠ $\underline{\mathscr{A}} = \mathcal{P}\left(\{0, 1\}^n\right) = \mathcal{P}\left(\{0, 1\}\right)^{\otimes n}$

The distribution of an n-tuple $(X_1, \ldots, X_n)$ of independent RVs is the product measure $\mathrm{P}^{X_1} \otimes \cdots \otimes \mathrm{P}^{X_n}$.

➠ $\mathscr{P}^{\underline{X}} = \{\mathrm{Ber}(p)^{\otimes n},\ p \in [0, 1]\}$

Remark: another possible choice would have been $\underline{\mathcal{X}} = \mathbb{R}^n$, $\underline{\mathscr{A}} = \mathcal{B}(\mathbb{R}^n)$.

Solution of exercise 1

back to questions

❷ Method of moments

If $X \sim \mathrm{Ber}(p)$, then $\mathbb{E}_p(X) = p$.

➠ The method of moments, applied to the first-order moment, directly yields the estimator $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$.

❸ Maximum likelihood

First write the likelihood:

$$\begin{aligned}\mathcal{L}(p; \underline{X}) &= \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} \\ &= p^N(1-p)^{n-N},\end{aligned}$$

where $N = \sum_{i=1}^n X_i$ and $0^0 = 1$,

Solution of exercise 1

back to questions

then the log-likelihood for $p \in (0,1)$:

$$\begin{aligned}\ell(p;\underline{X}) &= \ln(\mathcal{L}(p;\underline{X})) \\ &= N\ln(p) + (n-N)\ln(1-p).\end{aligned}$$

The log-likelihood is differentiable on $(0,1)$, with derivative

$$\begin{aligned}\frac{\partial \ell}{\partial p}(p;\underline{X}) &= \frac{N}{p} - \frac{n-N}{1-p} \\ &= \frac{n}{p(1-p)} \cdot \left(\bar{X}_n - p\right).\end{aligned}$$

We have $\frac{\partial \ell}{\partial p}(p;\underline{X}) > 0$ iff $p < N/n = \bar{X}_n$,
$\frac{\partial \ell}{\partial p}(p;\underline{X}) < 0$ iff $p > N/n = \bar{X}_n$.

Solution of exercise 1

back to questions

If $\bar{X}_n = 0$, the log-likelihood is strictly decreasing

➠ the likelihood is maximal at $p = 0$.

If $\bar{X}_n = 1$, the log-likelihood is strictly increasing

➠ the likelihood is maximal at $p = 1$.

If $0 < \bar{X}_n < 1$, the log-likelihood is maximal at $p = \bar{X}_n$.

Summary: $\hat{p}_n = \bar{X}_n$ is the unique MLE.

Remark: the log-likelihood takes infinite values at $p = 0$ and/or $p = 1$, but the likelihood itself is well defined and continuous on $[0,1]$.

Solution of exercise 1

back to questions

❹ Expectation and variance of $\bar{X}$

Reminders

- $\mathbb{E}_p(X_1) = p$ and $\mathrm{var}_p(X_1) = p(1-p)$.
- independence $\Rightarrow$ decorrelation $\Rightarrow \mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i)$.

Using that the X_i's are identically distributed:

$$\mathbb{E}_p\left(\bar{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_p[X_1] = p.$$

Using that the X_i's are IID:

$$\mathrm{var}_p(\bar{X}_n) = \frac{1}{n^2}\,\mathrm{var}_p\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}_p\left(X_i\right) = \frac{p(1-p)}{n}.$$

Solution of exercise 2

back to questions

❶ Maximum likelihood

First write the log-likelihood:

$$\ell(\mu; \underline{X}) = -\frac{n}{2}\ln(2\pi\sigma_0^2) - \sum_{i=1}^{n}\ln(X_i) + \frac{(\ln(X_i)-\mu)^2}{2\sigma^2} 1_{(\mathbb{R}_*^+)^n}(\underline{x}).$$

The log-likelihood is differentiable, with derivative (for $X_1, \ldots, X_n > 0$):

$$\frac{\partial \ell}{\partial \mu}(\mu; \underline{X}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln(X_i) - \mu)$$

Finally, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ is indeed the MLE since:

$$\frac{\partial \ell}{\partial \mu}(\mu; \underline{X}) > 0 \quad \text{for } \mu < \hat{\mu}, \qquad , \frac{\partial \ell}{\partial \mu}(\mu; \underline{X}) < 0 \quad \text{for } \mu > \hat{\mu}.$$

Solution of exercise 2

back to questions

❷ First express the probability of exceeding s_0 as a function of μ:

$$\begin{aligned} p_0 &= \mathbb{P}(X > s_0) \quad \text{with } X \sim \mathscr{LN}(\mu, \sigma^2) \\ &= 1 - F_{\mu,\sigma}(s_0) \\ &= 1 - \Phi_0\left(\frac{\ln(s_0) - \mu}{\sigma}\right). \end{aligned}$$

Then construct an estimator of p_0 by substitution, using $\hat{\mu}$:

$$\hat{p}_0 = 1 - \Phi_0\left(\frac{\ln(s_0) - \hat{\mu}}{\sigma}\right).$$

Solution of exercise 2

back to questions

❸ Let $Z_i = \ln(X_i)$, $i \geq 1$. The random variables Z_i are IID, and admit a first order moment equal to μ, since $Z_i \sim \mathcal{N}(\mu, \sigma^2)$.

Thus, by the strong law of large numbers:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln(X_i) \xrightarrow[n\to\infty]{\text{as}} \mathbb{E}(Z_1) = \mu.$$

Hence, using the continuity of $h : \mu \mapsto 1 - \Phi_0\left(\frac{\ln(s_0) - \mu}{\sigma}\right)$,

$$\hat{p}_0 = h(\hat{\mu}) \xrightarrow[n\to\infty]{\text{as}} h(\mu) = p_0.$$

Remark. Almost-sure convergence towards the parameter of interest is called *strong consistency* (see next lecture).

Solution of exercise 3

back to questions

❶ Using the Huygens-König and transfer theorems, we have:

$$\mathrm{var}\,(X_1) \;=\; \mathbb{E}\,(X_1^2) - \mathbb{E}\,(X_1)^2 \;=\; \mathscr{G}\left(\mathbb{P}^{X_1}\right)$$

where, for all $\mu \in \mathbb{M}$,

$$\mathscr{G}(\mu) \;=\; \int_{\mathcal{X}} x^2 \mu(\mathrm{d}x) - \left(\int_{\mathcal{X}} x \mu(dx)\right)^2.$$

❷ We use the substitution principle, with the empirical distribution as an estimator of P^{X_1}:

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

Solution of exercise 3

back to questions

We find the estimator

$$\begin{aligned} S_n^2 \;&=\; \int_{\mathcal{X}} x^2\, \hat{\mathbb{P}}_n^{X_1}(\mathrm{d}x) - \left(\int_{\mathcal{X}} x\, \hat{\mathbb{P}}_n^{X_1}(dx)\right)^2 \\ &=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2, \end{aligned}$$

which is called the *sample variance*.

❸ Applying the strong law of large numbers to the sequences (X_i) and $\left(X_i^2\right)$, which are IID RVs with a first order moment, we find

$$\bar{X} \;\xrightarrow{\text{as}}\; \mathbb{E}(X_1), \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \;\xrightarrow{\text{as}}\; \mathbb{E}(X_1^2),$$

Solution of exercise 3

back to questions

and therefore

$$S_n^2 \xrightarrow{\text{as}} \text{var}(X_1).$$

Remarks: on the other hand, we don't have convergence in L^2 in general, since the X_i^2's do not necessarily have a second order moment (for this, the X_i's would need to have a moment of order four).

Lecture outline

Lecture outline

The gamma family of distributions

A random variable X follows the $\Gamma(p, \lambda)$ distribution, with parameters $p > 0$ and $\lambda > 0$, if it has the pdf

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

Moments

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

Particular cases

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

Properties

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

back to slide ??

The log-normal distribution $\mathscr{LN}(\mu, \sigma^2)$

Definition

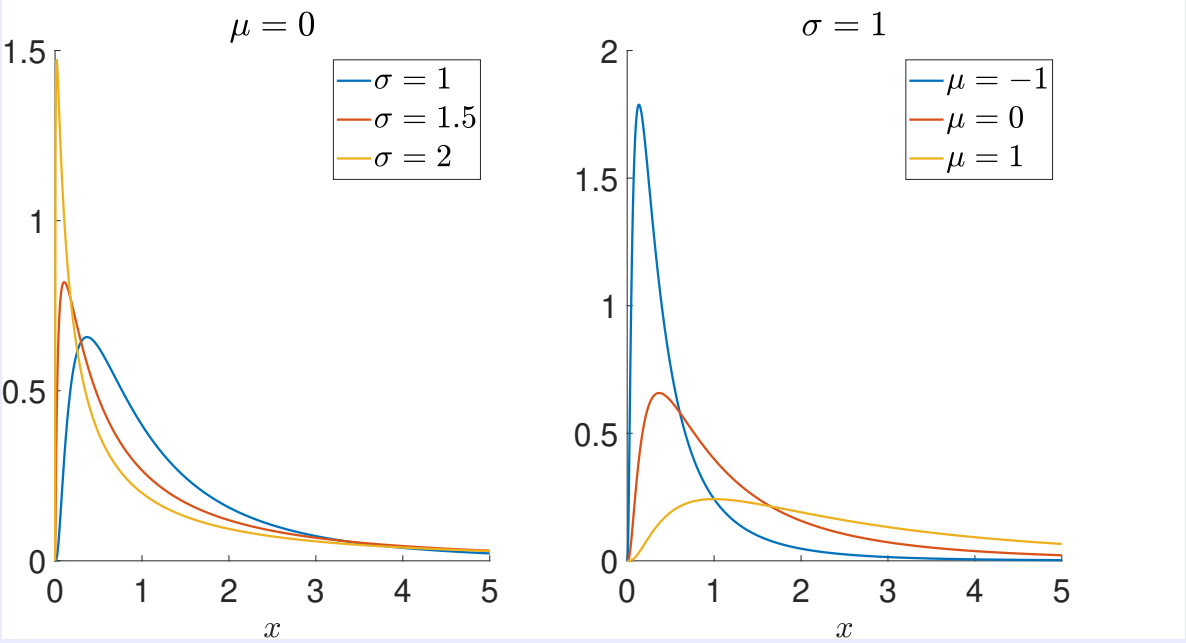
$X \sim \mathscr{LN}(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, if it admits the density

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) 1_{\mathbb{R}_*^+}(x).$$

Properties

- mean : $\mathbb{E}_{\mu,\sigma}(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$
- variance : $\mathrm{var}_{\mu,\sigma}(X) = \left(\exp\left(\sigma^2\right) - 1\right) \exp\left(2\mu + \sigma^2\right)$
- cumulative distribution function: $F_{\mu,\sigma} = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$, where Φ is the cumulative distribution function of the $\mathscr{N}(0, 1)$ distribution.
- $X \sim \mathscr{LN}(\mu, \sigma^2)$ iff $\ln(X) \sim \mathscr{N}(\mu, \sigma^2)$.

Density of the $\mathscr{LN}(\mu, \sigma^2)$ distribution



back to exercise 2

Lecture outline

Reminder: Probability density function wrt a measure

Let ν denote a positive masure on $(\underline{\mathcal{X}}, \underline{\mathscr{A}})$.

Definition: probability density function

The distribution $\mathbb{P}^{\underline{X}}$ of a RV $\underline{X}$ taking values in $(\underline{\mathcal{X}}, \underline{\mathscr{A}})$ admits a density with respect to ν if there exists $f : \underline{X} \to \mathbb{R}_+$, $\underline{\mathscr{A}}$-measurable and positive, st

$$\forall A \in \underline{\mathscr{A}}, \qquad \mathbb{P}\left(\underline{X} \in A\right) \;=\; \mathbb{P}^{\underline{X}}(A) \;=\; \int_A f(\underline{x})\, \nu(\mathrm{d}\underline{x}).$$

- f is the probability density function of $\mathbb{P}^{\underline{X}}$ with respect to ν.
- It satisfies $\int f \,\mathrm{d}\nu = 1$.

In this course, we will consider the following cases:

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

back to slide 31

Complement: the empirical cumulative distribution function

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, by substitution, the empirical cdf (ECDF):

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

For an IID n-sample $X_1, \ldots, X_n$ IID, with cumulative distribution function F, it can be proved (Glivenko-Cantelli theorem) that $\hat{F}_n \to F$ uniformly on $\mathbb{R}$, almost surely.

Complement: the empirical cumulative distribution function

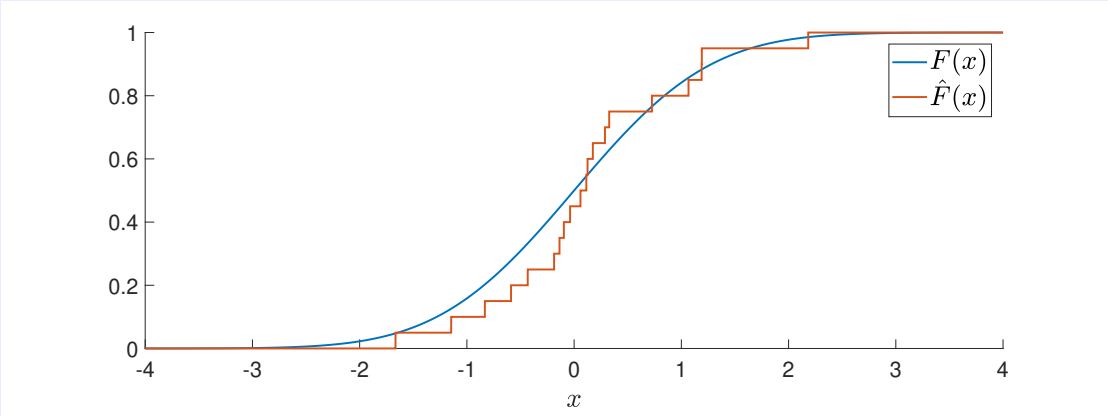


Figure – ECDF for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathscr{N}(0,1)$ and $n = 20$.

back to slide ??