# Advanced statistical methods

Frédéric Pascal

CentraleSupélec, Laboratory of Signals and Systems (L2S), France
frederic.pascal@centralesupelec.fr
`http://fredericpascal.blogspot.fr`

**MSc in Data Sciences & Business Analytics**
CentraleSupélec / ESSEC
Oct. 2nd - Dec. 20th, 2017

## Part B

On the use of classical estimation techniques
(before going on advanced ones)

# Part B: Contents

I. ML estimation with PDF

II. ML estimation with PMF

III. Non regular model

IV. Confidence interval

V. Be careful with asymptotic theory!

I. ML estimation with PDF

II. ML estimation with PMF

III. Non regular model

IV. Confidence interval

V. Be careful with asymptotic theory!

# ML estimation with PDF

## Context

Maximal height $H$ for the water in a river is observed every year because if the level is higher than 6 meters, consequences would be very important and the area surrounded the river would be flooded.

## Statistical modelling

One can model $H$ as a random variable with a Rayleigh distribution, i.e. the pdf of $H$ is given by

$$f_H(x) = \frac{x}{a} \exp\left(-\frac{x^2}{2a}\right), \qquad x \in \mathbb{R}_+,$$

where $a > 0$ is an unknown parameter.

# ML estimation with PDF

Discussion: What are we interested in?

Theoretical analysis

1. Derive the maximum likelihood estimator $\hat{a}_n$ of $a$.
2. Derive the Method of Moment estimator $\bar{a}_n$ based on the expectation.
3. Which properties does $\hat{a}_n$ verify among the following ones?
   a) Unbiased.
   b) Optimal.
   c) Efficient.
   d) Asymptotically Gaussian.

# ML estimation with PDF

## Application on real data

An insurance company estimates that a disaster appears at most once every one thousand years. The aim of this application is to decide whether this is justified or not regarding the following observations of the water level during 8 consecutive years:

$$2{,}5 \qquad 1{,}8 \qquad 2{,}9 \qquad 0{,}9 \qquad 2{,}1 \qquad 1{,}7 \qquad 2{,}2 \qquad 2{,}8.$$

1. Let $p$ the probability that a disaster happens during one year. What is $p$ (as a function of $a$)?
2. Deduce the probability that at most one disaster happens during one thousand years.
3. Give an estimation of this probability regarding the set of observations.

I. ML estimation with PDF

II. ML estimation with PMF

III. Non regular model

IV. Confidence interval

V. Be careful with asymptotic theory!

# ML estimation with PMF

## Statistical modelling and theoretical analysis

Let $X$ a random variable that takes values in $\mathbb{N}^\star$. This variable is defined as the first instant of success in a Bernouilli scheme with parameter $q \in ]0,1[$. This is generally known as the Geometry PMF:

$$P(X = k) = (1-q)^{k-1}\, q, \text{ for } k \in \mathbb{N}^\star$$

1. Verify that this PMF belongs to the exp. family. Give a sufficient stat.
2. Derive the FIM $I(q)$ for parameter $q$ (one can use only 1 r.v.).

   Let $X_1, \ldots, X_n$ a $n$-sample with same distribution as $X$.
3. Find the maximum likelihood estimator $\hat{q}_n$ of $q$.
4. Prove that this estimator is asymptotically Gaussian, and derive the asymptotic variance.
5. Give an asymptotic $(1-\alpha)$-confidence interval for $q$.

# ML estimation with PMF

## Application on real data

An urban mass transit company wants to estimate the number of passengers that do not pay their transport ticket on a given bus line. To that end, for one day in the week, the company knows the number $n_0$ of validated tickets in this bus line. Moreover, results from the following experiment are available: at each bus stop, ticket inspectors count the number of passengers alighting the bus with a validated ticket, until the exit of the first fraudster. The latter is included in the following data:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 44 | 09 | 11 | 59 | 81 | 44 | 19 | 89 | 10 | 24 |
| 07 | 21 | 90 | 38 | 01 | 15 | 22 | 29 | 19 | 37 |
| 26 | 219 | 02 | 57 | 11 | 34 | 69 | 12 | 21 | 28 |
| 34 | 05 | 07 | 15 | 06 | 129 | 14 | 18 | 02 | 156 |

1. Estimate the fraud probability. Give a 95%-confidence interval. Estimate the number of fraudsters $n_f$ if $n_0 = 20000$.

I. ML estimation with PDF

II. ML estimation with PMF

III. Non regular model

IV. Confidence interval

V. Be careful with asymptotic theory!

# Non regular model

Statistical modelling and theoretical analysis

Let $(X_1,\ldots,X_n)$ independent random variables following a uniform distribution on the interval $[0,\theta]$ where $\theta > 0$.

One aims at estimating $\theta$ from $(X_1,\ldots,X_n)$.

1. Show that the likelihood function is given by:

$$p_n(x;\theta) = \mathbb{1}_{\{\min_{1\le i\le n} x_i \ge 0\}}\, \mathbb{1}_{\{\max_{1\le i\le n} x_i \le \theta\}}\, \theta^{-n}.$$

   Deduce that the statistic $\hat{\theta}_n = \max_{1\le i\le n} X_i$ is sufficient. One assumes in the following that this statistic is also complete.

2. Why the model is not a regular one?

How can we deal with this problem?

# Non regular model

Uniform distribution - Interest: data simulations

Statistical modelling and theoretical analysis

1. Derive the distribution of $\hat{\theta}_n$ and compute $\mathrm{E}(\hat{\theta}_n)$. Deduce that $\hat{\theta}'_n = (1+\frac{1}{n})\hat{\theta}_n$ is optimal in the class of unbiased estimators.
2. Compute $\mathrm{E}(X_1)$ and deduce the moments estimator $\hat{\theta}''_n$ of $\theta$.
3. Let us now consider estimators of the form $c_n\hat{\theta}_n$. Prove that $\hat{\theta}_n$ provides a maximal likelihood.
4. Prove that $\hat{\theta}_n \leq \theta$ and derive $\mathsf{P}(|\hat{\theta}_n - \theta| > \epsilon)$ for any $\epsilon > 0$. Deduce that $\hat{\theta}_n$ converges towards $\theta$.
5. Does $\hat{\theta}'_n$ verify the CRB inequality?

Verify previous results on simulations (Matlab or R)

I. ML estimation with PDF

II. ML estimation with PMF

III. Non regular model

IV. Confidence interval

V. Be careful with asymptotic theory!

# Confidence interval

Let $(X_1,\ldots,X_n)$ independent random variables following a uniform distribution on the interval $[0,\theta]$ where $\theta > 0$. One aims at building a confidence interval for $\theta$ from $(X_1,\ldots,X_n)$.

1. Let $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Prove that the distribution of $\bar{X}_n/\theta$ does not depend on $\theta$. Thanks to the CLT, built a interval of the following form

$$I_n = \left[2\left(1+\frac{a}{\sqrt{n}}\right)^{-1}\bar{X}_n, 2\left(1-\frac{a}{\sqrt{n}}\right)^{-1}\bar{X}_n\right]$$

   such that this is an asymptotic $(1-\alpha)$-confidence interval for $\theta$.

2. Prove that the length of $I_n$ tends to 0 at the speed $1/\sqrt{n}$.
3. Let $\hat{\theta}_n = \max_{1\leq i\leq n} X_i$, the MLE for $\theta$ (To be discussed, cf. previous application). Prove that the distribution of $\hat{\theta}_n/\theta$ does not depend on $\theta$ and compute $\mathrm{P}[a \leq \hat{\theta}_n/\theta \leq b]$.

# Confidence interval

Statistical modelling and theoretical analysis

1. Deduce that there exist intervals of the form
$$J_n = [b_n^{-1}\hat{\theta}_n, a_n^{-1}\hat{\theta}_n]$$
that are exact $(1-\alpha)$-confidence intervals for $\theta$.
2. Prove that the smaller interval $J_n$ is the following
$$J_n^\star = [\hat{\theta}_n, \alpha^{-1/n}\hat{\theta}_n].$$
3. Prove that the length of $J_n^\star$ tends to 0 at the speed $1/n$.
4. Prove that $\{n(\theta-\hat{\theta}_n)\}_{n\geq 1}$ converges to an exponentially-distributed random variable with parameter $1/\theta$. Deduce that $\hat{\theta}_n$ is asymptotically biased and that the interval
$$K_n = \left[\hat{\theta}_n, \frac{1}{1+\ln(\alpha)/n}\hat{\theta}_n\right]$$
is an asymptotic $(1-\alpha)$-confidence interval for $\theta$.

I. ML estimation with PDF

II. ML estimation with PMF

III. Non regular model

IV. Confidence interval

V. Be careful with asymptotic theory!

# Counter-example

*From Basu D.(1988) Statistical Information and Likelihood, Springer-Verlag, N.Y.*

A ballot box contains 1000 tickets: on 20 tickets, it is written a value $\theta$ and on the 980 others, it is written $10\theta$. Test the problem with a given value of $\theta = \theta_0$.

1. Derive $\hat{\theta}$ the MLE of $\theta$ when we pull only one ticket denoted with value $X$, and show that $P(\hat{\theta} = \theta) = 0.98$.
2. Let's now remunerate the tickets "$10\theta$" by $a_i\theta$, for $i = 1, \ldots, 980$ where the $a_i$ are known real numbers, all distincts, and belonging in the interval $[10; 10.1]$. Give the new MLE $\tilde{\theta}$ and show that $P(\tilde{\theta} < 10\theta) = 0.02$.

Where is the paradox?

More tricky: $a_i \sim \mathcal{U}([10; 10.1])$ ... Here, $n = 1$, but this can be easily extended to large $n$... So, be careful!!!!