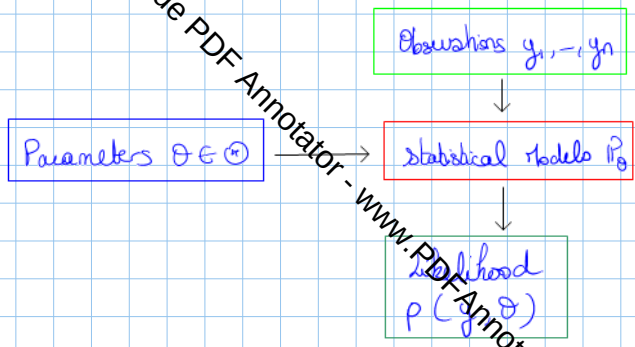
MATH 531 Chapter 2

Statistical inference

To start this chapter, we will give a schematic overview of the landscape:



The questions that arise are:

(i) Can we estimate the unknown parameter $\theta \in \Theta$ so that the model fits the data as good as possible → Point estimation

(ii) How can we take into account the uncertainty of this estimation to know for example if $\theta < \theta^*$ or $\theta \geq \theta^*$ where θ^* is a reference value that we know → hypothesis testing

(iii) Provide the precision of the estimation → Confidence intervals.

These will be the 3 topics we will see in the following

The first notion to be exhibited in the inference issue is what is called identifiability

Def: the model is said to be identifiable if $\theta_1 \neq \theta_2$ implies $P_{\theta_1} \neq P_{\theta_2}$

ex: the Exponential densities forms a family of identifiable models. So does the normals $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$

Goal of parametric inference = identify the unknown P_θ from the n sample of observations $(Y_1, \dots, Y_n)$

More generally, one may want to estimate a function $g(\theta) \in \mathbb{R}^k$ of $\theta \in \Theta$

Def: An estimator δ of $g(\theta)$ is a function of the n sample $(Y_1 \dots Y_n)$ taking its values in $g(\Theta)$
A sequence of estimators $(\delta_n)_{n>0}$ of $g(\theta)$ where δ_n is a function of $(Y_1 \dots Y_n)$ of size n is said consistent if for all $\theta \in \Theta$ we have

$$\lim_{n \to \infty} \delta_n = g(\theta) \quad P_\theta \text{ as.}$$

The question now is how do we construct consistent estimators. We will see several of them in the next sections.

Remark: Suppose we have a convergent estimator $\hat{\theta}_n$ of θ. If g is continuous then $g(\hat{\theta}_n)$ is a convergent estimator of $g(\theta)$

I Method of moments

We start this section with an illustrative example.

Example: Let us consider the beta distribution family $\mathcal{F} = \{\beta(a,b),\ a>0,\ b>0\}$ and $\theta = (a,b)$. We want to estimate θ.

Let $Y_1 \sim \beta(a,b)$, therefore $\mathbb{E}_\theta[Y_1] = \frac{a}{a+b}$ and $\mathbb{E}_\theta[Y_1(1-Y_1)] = \frac{ab}{(a+b)(a+b+1)}$

Thanks to the strong law of large numbers, we can propose the following convergent estimators of a and b:

$\hat{c}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\hat{d}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i(1-Y_i)$ where Y_i i.i.d $\beta(a,b)$

thanks to the previous remark, $\hat{a}_n = \frac{\hat{c}_n \hat{d}_n}{\hat{c}_n - \hat{d}_n - \hat{c}_n^2}$ and $\hat{b}_n = \frac{(1-\hat{c}_n)\hat{d}_n}{\hat{c}_n - \hat{d}_n - \hat{c}_n^2}$ are convergent estimators of a and b.

The method of moments consists in finding a function m, invertible and continuous, a measurable function g s.t $\mathbb{E}_\theta[|g(Y)|] < \infty$ and $m(\theta) = \mathbb{E}_\theta[g(Y)]$ $\forall \theta \in \Theta$

def: An estimator based on the method of moments is $\hat{\theta}_n = m^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i)\right)$

Remark: In most of the cases, the function g is polynomial

Proposition: If $\mathbb{E}_\theta[|g(Y)|] < \infty$ and m^{-1} is continuous, $\hat{\theta}_n \to \theta$ $\mathbb{P}_\theta$ ae

Moreover if $\forall \theta \in \Theta$, $\mathbb{E}[|g(Y)|^2] < \infty$ and if m is differentiable then

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{m'(\theta)^2}\mathrm{Var}_\theta[g(Y_i)]\right)$$

Proposé en exercice les cas lois exponentielle et Cauchy

This definition can be generalized to multidimensional parameters: $\theta \in \Theta \subset \mathbb{R}^d$, $d \geq 1$

In this case, it is most of the time impossible to identify θ with a single function g.

Let $g_\ell : \mathbb{R} \to \mathbb{R}$, $\ell \in \{1, \dots, d\}$ and consider $y \mapsto (g_1(y), \dots, g_d(y))$, $y \in \mathbb{R}$

such that $m_\ell(\theta) = \mathbb{E}_\theta[g_\ell(Y)] = \int_{\mathbb{R}} g_\ell(y) f_\theta(y)\, dy$; $\ell = 1, \dots, d$ has a unique solution

def: An estimator by the method of moments associated to the $(g_\ell)_\ell$ s is the solution of

$$m_\ell(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} g_\ell(Y_i), \quad \ell = 1, \dots, d$$

Notation: $\vec{m}(\theta) = \mathbb{E}_\theta[\vec{g}(Y)] = (\mathbb{E}_\theta[g_1(Y)], \dots, \mathbb{E}_\theta[g_d(Y)])$

and $\theta = \vec{m}^{-1}(m_1(\theta), \dots, m_d(\theta))$

To estimate θ we set $\hat{\theta}_n = \vec{m}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_1(Y_i); \dots; \frac{1}{n}\sum_{i=1}^{n} g_d(Y_i)\right)$

Proposition: If $\vec{m}$ is continuous, invertible with a continuous inverse, then the estimator by the method of moments is well defined and

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\text{Pas}} \theta$$

Moreover, if $\vec{m}^{-1}$ is differentiable and $\mathbb{E}_\theta[g_\ell(Y)^2] < \infty$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, V(\theta))$$

where $V(\theta) = J_{\vec{m}^{-1}} \, \Sigma_{\vec{m}}(\theta) \, J_{\vec{m}^{-1}}^{t}$ with $\Sigma_{\vec{m}}(\theta)$ the covariance matrix of $(g_1(Y), \dots, g_k(Y))^T$ and $J_{\vec{m}^{-1}}$ is the differential matrix of $\vec{m}^{-1}$ (Jacobian)

$$\left(\Sigma_{\vec{m}}(\theta)\right)_{ll'} = \mathbb{E}_\theta\left[g_l(Y) g_{l'}(Y)\right] - \mathbb{E}_\theta\left[g_l(Y)\right]\mathbb{E}_\theta\left[g_{l'}(Y)\right]$$

→ Faire l'exemple de la loi normale si on a le temps en cours sinon le faire en exercice (cf Nacc p83)

II Maximum likelihood estimator

Let us start with an example as above

Example: For an investigation, one is looking for a person of 1,80 meter high. the first question to ask is: Is this person a male of a female?

Thanks to the data from the "National Center for Health Statistics" (USA), the population is modeled as follows: males $\sim \mathcal{N}(m_H = 1{,}76, \sigma_H^2 = 0{,}07^2)$ and females $\sim \mathcal{N}(m_F = 1{,}62, \sigma_F^2 = 0{,}06^2)$

this looks like the following repartition:

this leads us to consider a male suspect rather than a female

this writes: $p(1{,}80 ; \theta_H) \geq p(1{,}80; \theta_F)$ where $\theta_H = (m_H, \sigma_H^2)$ and $\theta_F = (m_F, \sigma_F^2)$

We have chose the set of parameters which maximises the quantity $\theta \mapsto p(1{,}80; \theta)$

1) Definition of the maximum likelihood estimator (MLE)

def: we likelihood (function) associated to a statistical experiment $\hat{\mathcal{E}}$ the following function:

$$\theta \in \Theta \longmapsto \mathcal{L}_n(\theta; Y_1, Y_2, \dots, Y_n) = \prod_{i=1}^{n} f(\theta; Y_i) \quad \text{(noted } \mathcal{L}_n(\theta)\text{)}$$

where $f(\theta; Y)$ is the probability density function of Y (w.r.t the Lebesgue measure)

Note that the likelihood function is a random function, depending on the observations

Example: For the normal distribution above: $\mathcal{P}_\theta = \{ \mathcal{P}(\mu, \sigma^2) ; \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^* \}$

$$P_\theta(dy) = f(\theta, y)\, dy = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) dy$$

and $\mathcal{L}_n(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right)$

Faire en exercice le cas cauchy, Poisson (loi discrète) et melange (cf Nacc p82-93)

def: we call a maximum likelihood estimator, any estimator $\hat{\theta}_n^{MLE}$ such that

$$\mathcal{L}_n(\hat{\theta}_n^{MLE}; Y_1, \dots, Y_n) = \max_{\theta \in \Theta} \mathcal{L}_n(\theta; Y_1, \dots, Y_n)$$

that is to say $\hat{\theta}_n^{MLE} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathcal{L}_n(\theta; Y_1, \dots, Y_n)$

Note that the MLE may not exist and may not be unique

def: application : $\theta \in \Theta \longmapsto l_n(\theta; Y_1, \dots, Y_n) = \frac{1}{n} \log \mathcal{L}_n(\theta; Y_1, \dots, Y_n)$

$= \frac{1}{n} \sum_{i=1}^{n} \log f(\theta; Y_i)$

is well defined if $f(\theta, Y) > 0$ and is called the log likelihood function. Imposing $\log 0 = 0$, one can generalise this definition.

Note that $\hat{\theta}_n^{MLE} \in \operatorname{argmax}_{\Theta} l_n(\theta, Y_1, \dots, Y_n)$

2) Likelihood equation

If the maximum of $\theta \mapsto \mathcal{L}_n(\theta)$ (or $l_n(\theta)$) is not reached on the boundary of Θ and if $\mathcal{L}_n$ is continuously differentiable, then a necessary condition for $\hat{\theta}_n^{MLE}$ to be the MLE is

$$\nabla_\theta \mathcal{L}_n(\theta, Y_1, \dots, Y_n)\Big|_{\theta=\hat{\theta}_n^{MLE}} = 0 \quad \text{or equivalently} \quad \nabla_\theta l_n(\theta, Y_1, \dots, Y_n)\Big|_{\theta=\hat{\theta}_n^{MLE}} = 0$$

this is the likelihood equation.

3) Convergence:

We define the entropy of P_θ as $H_\theta = -\mathbb{E}_\theta[\log p(Y, \theta)]$ which write for a r.v with pdf $f(\theta, x)$: $H_\theta = -\int f(\theta, x) \log f(\theta, x)\, dx$

This quantity measure the "disorder" of a probability measure.

Theorem: We assume that:

(i) the model is identifiable, the set Θ is compact and $\Delta_\theta = \sup_\theta f(\theta, Y)$ does not depend on θ

(ii) $(\theta, y) \mapsto f(\theta, y)$ is bounded

(iii) $\theta \mapsto f(\theta, y)$ is continuous

(iv) H_θ is well defined for all $\theta \in \Theta$

then the MLE $\hat{\theta}_n^{MLE}$ is a convergent estimator of θ

Comments: this theorem is really restrictive and of course there exist other theorems that generalises this one. The thing to know is that the smoother the likelihood, the better the convergence.

4) Examples:

a) The gaussian model

We recall that $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$

$$l_n(\theta, Y_1, \dots, Y_n) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2$$

the likelihood equation writes: $\partial_\mu l_n(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)$

$$\partial_{\sigma^2} l_n(\theta) = -\frac{n}{2\sigma^2} + \frac{1}{4\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2$$

this yields : $\hat{\theta}_n^{MLE} = \left(\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \; ; \; \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 \right)$

We check that this critical point is the unique maximum

En exercice la loi log normale, des Bernoulli, des uniformes pour introduire le cas des indicatrices et l'exemple où il n'y a pas de MLE. → Tsac p 100-101

the maximum likelihood is given by one of the maximum of a specific function - this scheme is however more general and one can define other estimators of the same form. this is what we will see in the sequel

In the same spirit, an estimator can be one zero of a given function. this is another class of estimators.

the following section introduces these 2 categories : M and Z estimators.

III Z- and M- estimators

Problem of the method of moments : Requires an invertible function which may not be easy to find. the following classes of estimators do not depend on such a restrictive function. Moreover there exist methods to automatically select the best one (in an asymptotic way)

1) Z - estimators :

Let first assume that $\Theta \subset \mathbb{R}$, let us denote $m_g(\theta) = \int_{\mathbb{R}} g(x) P_\theta(x) dx$, then the method of moments can be written as :

$$\int_{\mathbb{R}} \left(m_g(\theta) - g(x) \right) P_\theta(x) dx = 0 \quad \text{where } g \text{ has to be chosen.}$$

One can generalize this definition.

Let now $\Theta \subset \mathbb{R}^d$, $d \geq 1$ and $\phi : \Theta \times \mathbb{R} \to \mathbb{R}^d$ such that $\int_{\mathbb{R}} \phi_\ell(\theta, x) P_\theta(x) dx = 0 \; , \forall 1 \leq \ell \leq d$

def : Given $\phi : \Theta \times \mathbb{R} \to \mathbb{R}$ as above, we call Z estimator associated to ϕ any estimator which solve the empirical form of the equality : that is to say :

$$\frac{1}{n} \sum_{i=1}^{n} \phi_\ell(\hat{\theta}_n, Y_i) = 0 \qquad \forall \; 1 \leq \ell \leq d$$

2) M- estimators :

Let now $\Psi : \Theta \times \mathbb{R} \to \mathbb{R}_+$ such that $\forall \theta \in \Theta \subset \mathbb{R}^d$, $d \geq 1$ the function

$$a \mapsto \mathbb{E}_\theta [\Psi(a, Y)] = \int_{\mathbb{R}} \Psi(a, y) P_\theta(y) dy \text{ has a maximum in } \theta \in \Theta$$

def : An M-estimator associated to Ψ (called contrast function) is given by

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \Psi(\theta, Y_i)$$

Note that an M estimator (up to regularity conditions) is also a Z - estimator by differentiation.

3) Convergence

We will give basic arguments on $\{P_\theta, \theta \in \Theta\}$ to ensure convergence of Z and M estimators. You can find more complex theorem with lighter conditions in the litterature.

See in particular the book of Van der Vaart on "Asymptotic Statistics".

i) M estimators

Let $M_n(a) = \frac{1}{n}\sum_{i=1}^{n} \Psi(a, Y_i)$, $a \in \Theta$ and $M(a,\theta) = \mathbb{E}_\theta[\Psi(a, Y)]$

Proposition: Assume $\Theta \subset \mathbb{R}^d$, $d \geq 1$ and $\hat{\theta}_n$ the M estimator is well defined. Assume in addition that

(i) $\sup_{a\in\Theta} |M_n(a) - M(a,\theta)| \xrightarrow{P_\theta} 0$

(ii) $\forall \varepsilon > 0 \quad \sup_{|a-\theta|>\varepsilon} M(a,\theta) < M(\theta,\theta)$ (condition of maximum)

(iii) $M_n(\hat{\theta}_n) \geq M_n(\theta) - \varepsilon_n$ with $\varepsilon_n \xrightarrow{P_\theta} 0$

then the M-estimator $\hat{\theta}_n$ is consistent (or convergent): $\hat{\theta}_n \xrightarrow{P_\theta} \theta$

ii) Z-estimators

Let $Z_n(a) = \frac{1}{n}\sum_{i=1}^{n} \phi(a, Y_i)$, $a \in \Theta$ and $Z(a,\theta) = \mathbb{E}_\theta(\phi(a, Y))$, $a \in \Theta$

Proposition: We assume that the Z-estimator is well defined and:

(i) $\sup_{a\in\Theta} \|Z_n(a) - Z(a,\theta)\| \xrightarrow{P_\theta} 0$ ($\|.\|$ is the Euclidean norm)

(ii) $\forall \varepsilon > 0 \quad \inf_{|a-\theta|\geq\varepsilon} \|Z(a,\theta)\| > 0 = \|Z(\theta,\theta)\|$

(iii) $Z_n(\hat{\theta}_n) \xrightarrow{P_\theta} 0$

then $\hat{\theta}_n$ is consistant (or convergent) ie $\hat{\theta}_n \xrightarrow{P_\theta} \theta$

One can also prove Central limit theorems to have the speed of convergence. → analyse asymptotique?

IV Which estimator is the best?

The standard approach to making such judgments is called decision theory.

For estimation, this approach begins with the choice of a loss function L such that $L(\theta, d)$ quantifies the loss associated with estimating $g(\theta)$ by the value d. It is natural to assume that $L(\theta, g(\theta)) = 0$ so that there is no loss with the correct answer. Moreover we assume also $L(\theta, d) \geq 0$ $\forall \theta$ and d.

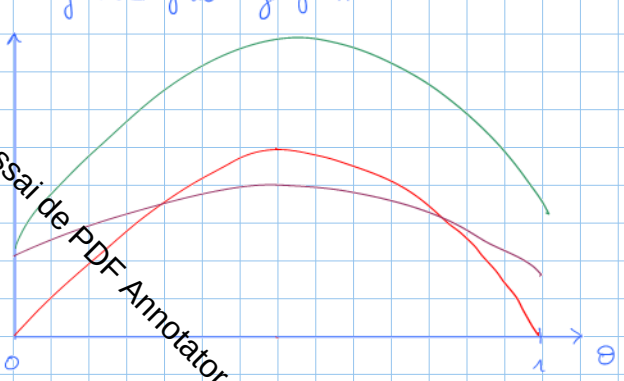
Because Y is random, $L(\theta, \delta(Y))$ is also random and can be large even if the estimator is excellent. Therefore, to judge the optimality of an estimator, one have to think in average.

def: Let $R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(Y))]$. R is called the risk function

Example: $Y \sim Bin(100, \theta)$, $\theta \in [0,1]$
this is equivalent of n coin tosses with a coin with $\theta = P(\text{head})$
therefore an estimator of θ is $\delta(Y) = \frac{Y}{100}$

If we consider the loss function given by $L(\theta, d) = (\theta - d)^2$ which is the quadratic loss, then $R(\theta, \delta) = \mathbb{E}_\theta\left[\left(\theta - \frac{Y}{100}\right)^2\right] = \frac{\theta(1-\theta)}{100}$

The risk is of the following form



$R(\theta, \delta_1)$

$R(\theta, \delta_0)$

$R(\theta, \delta_2)$

Now the question is: can we compare 2 estimators?

Following our example, one can invent other estimators: let $\delta_1(Y) = \frac{Y+3}{100}$ and $\delta_2(Y) = \frac{Y+3}{106}$

Then $R(\theta, \delta_0) = \frac{\theta(1-\theta)}{100}$; $R(\theta, \delta_1) = \frac{9 + 100\theta(1-\theta)}{100^2}$ and $R(\theta, \delta_2) = \frac{(9-8\theta)(1+8\theta)}{106^2}$

If we plot these risks, one can see that δ_0 and δ_2 are better than δ_1. However comparing δ_0 and δ_2 depends on the value of θ. (→ explain the plot)

However, as one does not know θ, one cannot choose the best estimator. This leads us to the following definition:

def: δ_1 is better to δ_2 if: for all $\theta \in \Theta$, $R(\theta, \delta_1) \leq R(\theta, \delta_2)$
If in addition $R(\theta, \delta_1) \neq R(\theta, \delta_2)$ for at least one $\theta \in \Theta$, we say that δ_2 is inadmissible.

As one can see in our example, there are not necessary better estimators in the whole family. However restricting to a subfamily, one may get a better estimator.

Remark: In 1-D : $R(\theta, \delta) = \mathrm{Var}_\theta(\delta) + \left(\mathbb{E}_\theta[\delta] - g(\theta)\right)^2$ for the quadratic case.

Therefore, as we aim at minimising $R(\theta, \delta)$ for all θ, it looks natural to choose an estimator δ such that $\mathbb{E}_\theta(\delta) = g(\theta)$ and with a minimal variance.

def: An estimator, δ, of $g(\theta)$ is integrable if $\mathbb{E}_\theta[\delta] < \infty$ for all $\theta \in \Theta$. It is unbiased if it is integrable and $\mathbb{E}_\theta[\delta] = g(\theta)$ $\forall \theta \in \Theta$.

Note that if δ is unbiased and $\mathbb{E}_\theta[\delta^2] < \infty$ $\forall \theta \in \Theta$, then $R(\theta, \delta) = \mathrm{Var}_\theta(\delta)$

Examples:
- If $X \in L^1(P_\theta)$ $\forall \theta \in \Theta$ then $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of $\mathbb{E}_\theta[X]$ (thanks to the strong law of large numbers)
- $\mathcal{B} = \{\mathcal{U}_{[0,\theta]} ; \theta > 0\}$ then $\hat{\theta}_n = \max_{i=1,\dots,n} X_i$ is a convergent estimator of θ. However it is biased.

Indeed, since the X_i's are independent: $\forall x \geq 0$

$$P_\theta(\hat{\theta}_n \leq x) = P_\theta(\forall i \in \{1,\dots,n\}\ X_i \leq x) =$$

$$= \prod_{i=1}^{n} P_\theta(X_i \leq x) = P(X_1 \leq x)^n$$

$$= \frac{1}{\theta^n} x^n \mathbb{1}_{[0,\theta]}(x)$$

this yields $\hat{\theta}_n$ has a density function given by $\frac{n}{\theta^n} t^{n-1} \mathbb{1}_{[0,\theta]}(t)$

and $\mathbb{E}_\theta[\hat{\theta}_n] = \frac{n}{n+1}\theta$

To construct a convergent unbiased estimator of θ, one can use $\tilde{\hat{\theta}}_n = \frac{n+1}{n} \hat{\theta}_n$

Remark: In general cases, the MLE is biased

→ example of the estimation of the support of a uniform distribution as exercice (cf Eric poly p 52)

A way to compare two estimators δ_1 and δ_2 is therefore to study their risk at the "hardest" point

def: An estimator δ_0 is called minimax estimator of $g(\theta)$ on $\Theta_1 \subset \Theta$ for the loss function L if:

$$\sup_{\theta \in \Theta_1} R_L(\delta_0, \theta) = \inf_{\delta} \sup_{\theta \in \Theta_1} R_L(\delta, \theta)$$

where the infimum is taken over all estimators of $g(\theta)$

Example: Let $Z = (X_1, \dots, X_n)$ with $X_1, \dots, X_n$ iid real r.v. and $Z \sim P_\theta^{\otimes n}$ for some $\theta \in \Theta$

We assume that $\forall \theta \in \Theta \quad \mathbb{E}_\theta[|X_1|] = \int |x| P_\theta(dx) < \infty$ and $\mathbb{E}_\theta[X_1^2] < \infty$

We want to estimate $\mathbb{E}_\theta(X_1) = g(\theta)$

We introduce the statistic $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Therefore $\overline{X}_n$ is an unbiased estimator of $g(\theta)$.

By definition X_1 is also an unbiased estimator of $g(\theta)$.

Choosing the quadratic loss function $L(x,y) = (x-y)^2$ we get:

$$\begin{aligned}
R_L(\overline{X}_n, \theta) &= \mathbb{E}_\theta\left[\left(\overline{X}_n - g(\theta)\right)^2\right] \\
&= \mathbb{E}_\theta\left[\frac{1}{n^2}\left(\sum_{i=1}^{n}(X_i - g(\theta))\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}_\theta\left[\sum_{i=1}^{n}(X_i - g(\theta))^2 + \sum_{\substack{i=1 \\ j \neq i}}^{n}\sum_{j=1}^{n}(X_i - g(\theta))(X_j - g(\theta))\right] \\
&= \frac{1}{n^2} n\, \mathbb{E}_\theta\left[(X_i - g(\theta))^2\right] + \frac{1}{n^2}\sum_{\substack{i=1 \\ j \neq i}}^{n}\sum_{j=1}^{n} \underbrace{\mathbb{E}[X_i - g(\theta)]}_{=0}\underbrace{\mathbb{E}[X_j - g(\theta)]}_{=0} \quad (\text{iid var}) \\
&= \frac{1}{n}\mathbb{E}_\theta\left[(X_1 - g(\theta))^2\right] \\
&= \frac{1}{n} R_L(X_1, \theta)
\end{aligned}$$

Therefore for $n > 1$ and $\mathbb{E}[(X_1 - g(\theta))^2] > 0$, for at least one θ, X_1 (as an estimator of $g(\theta)$) is inadmissible

Vérifier que la notion d'admissibilité a bien été définie au dessus

Remark: loss functions usually satisfy the following properties:

(i) $L(x,y) = f(x-y)$

(ii) $f : \mathbb{R}^d \to \mathbb{R}^+$ (with $\Theta \subset \mathbb{R}^d$). $f(0) = 0$ or f is continuous and $f \neq 0$

(iii) f is symmetric ie $f(-x) = f(x)$

(iv) the level sets of f : $\{x : f(x) < c\}$ are convex for all $c > 0$ and bounded for c small

the most common loss functions are:

- the quadratic one : $f(x,y) = \|x-y\|_2^2$
- the absolute loss : $f(x,y) = \|x-y\|_1$
- $f(x,y) = 1_{\{x \neq y\}}$

* Bias - Variance Tradeoff:

If you look closely to the quadratic loss then:

$$\mathbb{E}_\theta\left[\|\delta(z) - g(\theta)\|^2\right] = \mathbb{E}_\theta\left[\|\delta(z) - \mathbb{E}_\theta[\delta(z)] + \mathbb{E}_\theta[\delta(z)] - g(\theta)\|^2\right]$$

$$= \mathbb{E}_\theta\left[\|\delta(z) - \mathbb{E}_\theta[\delta(z)]\|^2\right] + \mathbb{E}_\theta\left[\|\mathbb{E}_\theta[\delta(z)] - g(\theta)\|^2\right]$$

$$- 2\,\mathbb{E}_\theta\Big[\underbrace{\langle \delta(z) - \mathbb{E}[\delta(z)] ;\ \underbrace{\mathbb{E}_\theta(\delta(z)) - g(\theta)}_{\text{indep of } z}\rangle}_{=0}\Big]$$

$$= \underbrace{\mathbb{E}_\theta\left[\|\delta(z) - \mathbb{E}_\theta[\delta(z)]\|^2\right]}_{\text{variance of } \delta} + \underbrace{\|\mathbb{E}_\theta[\delta(z)] - g(\theta)\|^2}_{\text{bias of } \delta}$$

To get a good estimator, one can sometimes prefer a δ with a small bias and small variance towards an unbiased δ but with high variance.

Example of { survey → TD of Eric p 54 - 55 - 56
{ empirical variance

Demander à Eric pertinence de ce §

VI Bayesian estimation :

- Intérêt : utilise données pour ajouter info
  Débruitage la suite
- Introduire priors : $L^1 \rightarrow$ MAP
  $L^2 \rightarrow$ $\mathbb{E}$ a posteriori

V Unbiased Estimation

If time (ask Eric)

1) Minimum Variance Unbiased estimators

- UMVU
- Uniqueness of UMVU when T is a complete sufficient statistic.
- ex

2) Why would not considering biased estimators?

- Explain thanks to an example the tradeoff bias-variance.