Laurent DELSOL
Laurent.delsol@univ-orleans.fr.

19.10.2022.

Estimates methods in Statistics.

Part 1. Parametric estimation. $\theta \in \mathbb{R}^d$.
- Definition
- Property.

Part 2. Non parametric estimation.
- f_x r.

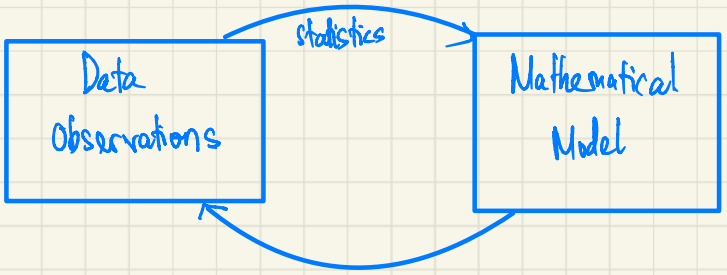
Exercises
- Practice on R software.

8h - 11h15
13h - 16h15.

Part 1 Parametric Estimation.

Introduction. Probability vs Statistics.

Example. Flip a coin.
- 1 : tail
- 0 : else.



{ Probabilistic model $(\Omega, \mathcal{A}, \mathbb{P})$ — kg chứa tham số.
{ Statistical model $(\Omega, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$

Estimation problem. Data get something about θ.

Statistical framework:

One observes n random variables $X_1, X_2, \dots, X_n$ taking value in a measurable space $(\mathcal{X}, \mathcal{F})$. We will assume in this lecture that those variables are independent with the same law $\mathbb{P}_{X,\theta}$. We will write here after $X_1, X_2 \dots X_n \overset{i.i.d}{\sim} \mathbb{P}_{X,\theta}$. $(X_1 \dots X_n \overset{i.i.d}{\sim} X)$. And $\mathbb{P}_{X,\theta}$ is unknown.

And X would represent a variable whose distribution is $\mathbb{P}_{X,\theta}$.

i.i.d: independent and identically distributed.

Aim: Is it possible to infer something on $\mathbb{P}_{X,\theta}$ or to get from the observation of the random variable $X_1 \dots X_n$?

$X = X_1 \dots X_n$: random sample.

X_1: represent the behavior of the first individual.

X_j: _____ j^{th} _____

Observation: $x_1 \dots x_n$ values taken by $X_1 \dots X_n$ in the study.

A. Definitions.

Def 1.1 . The sequences of RVs $(X_1 \_\_ X_n)$ is a random sample of size n if $X_1 \_\_ X_n \overset{i.i.d}{\sim} \mathbb{P}_X$

. The distribution of the sample is $(\mathbb{P}_X)^{\otimes n}$

. When the statistical study is performed you observe the values of these RVs $x_1 \_\_ x_n$ (DATA)

Def 1.2 . A statistical model is said to be parametric if the set of parameters $\Theta \subseteq \mathbb{R}^m$ $(m < +\infty)$

$(\Omega, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta \subseteq \mathbb{R}^m\})$

← parameter is vector

↓ biết phân phối gì nhưng có tham số chưa biết.

Def 1.3 . A parametric model is said to be dominated if there exists a σ-finite dominant measure Q in such a way that all probabilities $\mathbb{P}_\theta$ are absolutely continuous with respect to Q.

. We will then write $h(\cdot, \theta)$ the density of $\mathbb{P}_\theta$ with respect to Q and hence the density of the random sample $\underline{X} = (X_1 \_\_ X_n)$ is:

$$h(\underline{x}, \theta) = h(x_1, \theta) \_\_ h(x_n, \theta).$$

Examples. continuous RV $\longrightarrow$ Q : Lebesgue measure.

discrete RV $\longrightarrow$ Q : counting measure $\mathbb{N}$.

We will denoted hereafter $\mathbb{E}_\theta$ and Var_θ the expectation and variance to explain they refer to $\mathbb{P}_\theta$.

Examples Discrete case.

Flip a coin $\quad \mathbb{P}_\theta = \mathrm{Ber}(\theta) \qquad X \sim \mathrm{Ber}(\theta)$.

$\mathbb{P}(X=1) = \theta \qquad X$ takes value in $\{0,1\}$.

$\mathbb{P}(X=0) = 1-\theta$.

$Q = \sum_{k\in\mathbb{N}} \delta_k$: δ_k is Dirac distribution at k.

counting measure.

$$Qf = \sum_{k\in\mathbb{N}} \delta_k f = \sum_{k\in\mathbb{N}} f(k).$$

. If $\mathbb{P}_\theta = \mathrm{Ber}(\theta) \qquad \Theta =]0,1[$.

$$\mathbb{P}_\Theta = \{ \mathrm{Ber}(\theta), \theta \in \Theta =]0,1[\}$$

$$h(x,\theta) = \theta^x (1-\theta)^{1-x} = \begin{cases} \theta & \text{if } x = 1. \\ 1-\theta & \text{if } x = 0. \end{cases}$$

$$h_n(\underline{x},\theta) = \theta^{x_1} \cdots \theta^{x_n} (1-\theta)^{1-x_1} \cdots (1-\theta)^{1-x_n}$$

$$= \theta^{y_n} (1-\theta)^{n-y_n}, \text{ where } y_n = x_1 + \cdots + x_n.$$

$$Y_n = \sum_{i=1}^n X_i \sim \mathrm{Bin}(n,\theta) \quad \text{for binomial} \quad \binom{x}{n} \theta^x (1-\theta)^{n-x}.$$

Continuous case. $P_\theta = \mathcal{N}(\mu, \sigma)$, $\theta = (\mu, \sigma) \subseteq \mathbb{R} \times \mathbb{R}_+^*$

σ : standard deviation

Q : Lebesgue measure.

$$f(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

$$f_n(\underline{x}, \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}$$

. **Notation** The function $\theta \longmapsto f_n(\underline{x}, \theta)$, when $\underline{x}$ is fixed is called the likelihood of the sample (usually written $L_n(\underline{x}, \theta)$)

. Of course the likelihood depends on dominant measure.

Def 1.4. A statistic T is a function $\Omega \longrightarrow \mathbb{R}^k$ computed from the random sample $X_1 \dots X_n$

. $\omega \longmapsto (X_1(\omega), \dots, X_n(\omega))$
$\longmapsto T(X_1(\omega), \dots, X_n(\omega))$
$f : \mathbb{R}^n \to \mathbb{R}^k$
. $T(\omega) = f(X_1(\omega), \dots, X_n(\omega))$

Ex $X_1 \dots X_n$ random sample

$$T = \frac{X_1 + \dots + X_n}{n} = \overline{X}$$

$T = X_1$, $T = 1$, $T = \underline{X}$, $T = \underline{X}$, $T = \min(X_i, 1 \le i \le n)$ ..

1.2. Parametric estimators.

We consider a parametric statistical model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^m$ and a random sample $X_1, \dots, X_n \overset{i.i.d}{\sim} P_\theta \in \mathcal{P}$.

Def. <u>Estimator of $\lambda = g(\theta)$</u> (g is a known function).

θ: natural parameter
λ: interest parameter

. An estimator of λ usually denoted $\hat{\lambda}$, is a statistic computed from the sample $X_1 \_\_ X_n$ (and taking values in a $g(\Theta)$)

Remark. This definition is very general and does not ensure at all that $\hat{\lambda}$ will be closed to λ.

Ex Flip a coin $P_\theta = Ber(\theta)$, $\theta \in]0,1[$

$\lambda = \theta$.

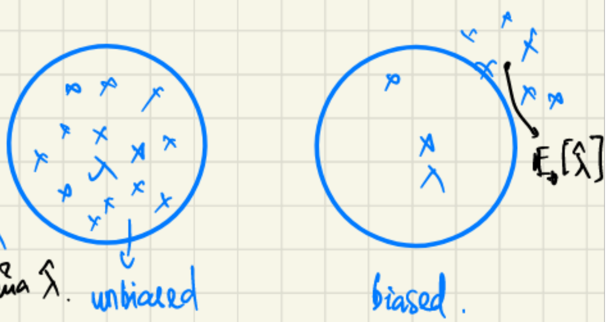
$\hat{\lambda} = 1/4$. or $\hat{\lambda} = X_1$ or $\hat{\lambda} = \overline{X}$ → đều ok theo def, but $\underline{o}$ gần λ → ƯL $\underline{o}$ tốt.

$\overline{X}$ → ƯL tốt nhất vì thu thập về $\theta = \lambda$.

$\overline{X} \xrightarrow{a.s} \mathbb{E}[X] = \theta = \lambda$.

1.2.1 Expected properties

a) Bias



. The bias of an estimator $\hat{\lambda}$ of λ is the quantity:

(trung bình của $\hat{\lambda}$.)

$$\boxed{Bias_\theta(\hat{\lambda}) = \mathbb{E}_\theta[\hat{\lambda}] - \lambda}$$

. An estimator is said to be unbiased if $Bias_\theta(\hat{\lambda}) = 0$

________ biased if $Bias_\theta(\hat{\lambda}) \neq 0$.

________ asymptotically unbiased if $Bias_\theta(\hat{\lambda}) \xrightarrow[n \to +\infty]{} 0$

Example $X \sim Ber(\theta)$, $X_1, \dots, X_n \overset{i.i.d}{\sim} X$

$\hat{\theta}_1 = X_1$

$\hat{\theta}_2 = \bar{X}$

$\hat{\theta}_3 = 1/4$

Biased?

- $\mathbb{E}_\theta[\hat{\theta}_1] = \mathbb{E}_\theta[X_1] = \theta$
- $\mathbb{E}_\theta[\hat{\theta}_2] = \mathbb{E}_\theta\left[\frac{1}{n}\cdot\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i] = \frac{1}{n}\cdot\sum_{i=1}^{n}\mathbb{E}[X] = \mathbb{E}[X] = \theta$

↳ **Note:** $\boxed{\mathbb{E}_\theta[\bar{X}] = \mathbb{E}_\theta[X] \quad \text{for all cases.}}$

$\mathbb{E}_\theta[1/4] = 1/4 \neq \theta$ → $\hat{\theta}_1, \hat{\theta}_2$: unbiased estimator

$\hat{\theta}_3$: biased estimator.

Example. $X \sim \mathcal{N}(\mu, \sigma)$, $\lambda = \sigma^2$.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$\hat{\sigma}^2 = \frac{1}{n}\cdot\sigma^2\cdot Y_n$

→ $\mathbb{E}_\theta[\hat{\sigma}^2] = \frac{1}{n}\cdot\sigma^2\cdot\mathbb{E}(Y_n) = \frac{n-1}{n}\sigma^2$

$$\mathbb{E}_\theta[\hat{\sigma}^2] = \frac{n-1}{n}\cdot\sigma^2.$$

$\mathbb{E}(Y_n) = n-1$.

since $Y_n = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1)$: χ^2 chi bình phương.

- $\text{Bias}(\hat{\sigma}^2) = \frac{n-1}{n}\cdot\sigma^2 - \sigma^2 = -\frac{1}{n}\cdot\sigma^2 \xrightarrow{n\to+\infty} 0$.

↳ $\hat{\sigma}^2$ is an asymptotically unbiased.

$\Rightarrow \hat{\sigma}^2_{n-1} = \frac{n}{n-1}\hat{\sigma}^2$: để ra unbiased $\rightarrow \hat{\sigma}^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Then $E_\theta[\hat{\sigma}^2_{n-1}] = E_\theta[\frac{n}{n-1}\hat{\sigma}^2]$

$= \frac{n}{n-1} E_\theta[\hat{\sigma}^2]$

$= \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \sigma^2 = \sigma^2 \rightarrow$ unbiased.

b) Compare variances.

If $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are 2 unbiased estimators of λ. Which is the best?

small variance

Better

great variance

Worse.

- Quadratic risk.

$$E_\theta[(\hat{\lambda}-\lambda)^2] = Var_\theta(\hat{\lambda}) + (Bias_\theta(\hat{\lambda}))^2$$

- For unbiased estimator, the quaratic risk is equal to the variance of $\hat{\lambda}$

• Hence the best estimator will be the one with smallest variance.

Def. If $\lambda = g(\theta) \in \mathbb{R}$

λ is square integrable.

Then the quadratic risk $R_\theta(\hat{\lambda})$ is $R_\theta(\hat{\lambda}) = \mathbb{E}_\theta[(\hat{\lambda}-\lambda)^2]$

Property. $R_\theta(\hat{\lambda}) = \mathrm{Var}(\hat{\lambda}) + (\mathrm{Bias}(\hat{\lambda}))^2$.

• $R_\theta(\hat{\lambda}) \xrightarrow{n\to+\infty} 0$ then $\mathbb{L}^2$ consistant.

New unbiased then $R_\theta(\hat{\lambda}) = \mathrm{Var}(\hat{\lambda})$.

Come back to our example

$X \sim \mathrm{Ber}(\theta)$

$\hat{\theta}_1 = X_1$ } unbiased

$\hat{\theta}_2 = \bar{X}$ }

• $\mathrm{Var}_\theta(\hat{\theta}_1) = \theta(1-\theta) \not\to 0$ ⇒ not $\mathbb{L}^2$ consistant.

• $\mathrm{Var}_\theta(\hat{\theta}_2) = \mathrm{Var}\left(\frac{1}{n}\cdot\sum_{i=1}^n X_i\right)$

$\overset{\text{independent}}{=} \frac{1}{n^2}\cdot\sum_{i=1}^n \mathrm{Var}(X_i)$

$= \frac{1}{n^2}\cdot\sum_{i=1}^n \mathrm{Var}(X) = \frac{\mathrm{Var}(X)}{n} = \frac{\theta(1-\theta)}{n}$.

Remarks: $X_1, \dots, X_n \overset{i.i.d}{\sim} X$

• $\mathbb{E}_\theta[\bar{X}] = \mathbb{E}_\theta[X]$

• $\mathrm{Var}_\theta(\bar{X}) = \frac{\mathrm{Var}_\theta(X)}{n}$

$R(\hat{\theta}_1) = \theta(1-\theta)$

$R(\hat{\theta}_2) = \frac{\theta(1-\theta)}{n} \xrightarrow[n\to+\infty]{} 0$

↳ $\hat{\theta}_2 = \bar{X}$ is better.

. $\hat{\lambda}$ is said to be consistant in quadratic mean if $R(\hat{\lambda}) \xrightarrow[n\to+\infty]{} 0$
(converge in L^2 sense)

c) **Consistant estimator**

Our aim $\hat{\lambda} \xrightarrow[n\to+\infty]{?} \lambda$ hội tụ kiểu gì cũng được

- L^p sense
- a.s : hội tụ hầu chắc chắn.
- P : hội tụ theo xs
- $\mathcal{L}$: ——— p^2

Def. Let $\hat{\lambda}$ be (a sequence of) estimators of λ. Then $\hat{\lambda}$ is a strongly consistant estimator of λ if (P_θ a.s)

$\forall \theta \in \Theta$ $\quad \hat{\lambda} \xrightarrow[n\to+\infty]{P_\theta \text{ a.s}} \lambda$ ← dùng SLLN.

Remark. Strong Law of Large Numbers (SLLN)

If $\begin{cases} X_1, \dots, X_n \overset{i.i.d}{\sim} X, & X \in \mathbb{R}^k \\ \varphi : \mathbb{R}^k \to \mathbb{R}^n \end{cases}$

with $\mathbb{E}[|\varphi(X)|] < +\infty$

$$\overline{\varphi(X)} = \frac{\varphi(X_1) + \dots + \varphi(X_n)}{n} \xrightarrow[n\to+\infty]{P_\theta \text{ a.s}} \mathbb{E}[|\varphi(X)|]$$

Come back to our example $X \sim Ber(\theta)$ ← dùng $P_\theta(\hat{\lambda})$. (có dính n mới nói đến hội tụ)

	Unbiased	L^2 consistant	P_θ a.s consistant.
$\hat{\theta}_1 = X_1$	OK	No	No.
$\hat{\theta}_2 = \overline{X}$	OK	OK	ok (dùng SLLN)
$\hat{\theta}_3 = 1/4$	No	No	No

- If $\lambda = \theta(1-\theta) = g(\theta)$ then $\hat{\lambda} = \overline{X}(1-\overline{X})$ is strongly consistant

because | g is continuous
| $\overline{X}$ is strongly consistant

from the continuous mapping theorem.

↳ $\overline{X}$ hội tụ về θ } $\Rightarrow g(\overline{X}) \xrightarrow{\text{a.c}} g(\theta)$.
g liên tục

d) Asymptotic distribution

Asymptotic normality.

Def Let $\hat{\lambda}$ be a sequence of estimators of λ. The sequence is said asymptotically normally distributed if for all $\theta \in \Theta$

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

↳ variance - covariance matrix.

or $r_n(\hat{\lambda} - \lambda) \xrightarrow[n \to +\infty]{} \mathcal{N}(0, \Sigma(\theta))$

Example. $X \sim Ber(\theta)$

$\hat{\theta}_2 = \overline{X}$, $E_\theta[X] = \theta$, $Var(X) = \theta(1-\theta)$

$$\frac{\overline{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

↳ Gaussian.

(Central Limit Theorem) CLT

$$\sqrt{n}(\overline{X} - \theta) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta))$$

Main idea. CLT + Δ method

Slutsky theorem.

1.3. Comparing estimators and optimization.

The aim is to compare estimators and get the best of them.

Def A loss function L is an application $L: \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ such that

$\forall x \in \mathbb{R}^d, L(x,x) = 0$ and (when in addition

$\forall x,y \in \mathbb{R}^d, x \neq y \Longrightarrow L(x,y) > 0$ | we talk about a contrast function

Example. $L_p(x,y) = \sum_{i=1}^{d} |x_i - y_i|^p$

. If $p=2$, the quadratic loss function. More generally L is convex and gives zero when $x=y$.

. If $\hat{\lambda}$ is an estimator of λ then $L(\hat{\lambda}, \lambda)$ is the loss coming from the estimator of λ.

. **Remark.** It is usually difficult to compare directly $L(\hat{\lambda}, \lambda)$ and $L(\tilde{\lambda}, \lambda)$ because both of them are RVs.

. for instance if $\lambda \in \mathbb{R}$: $L_2(\hat{\lambda}, \lambda) = (\hat{\lambda} - \lambda)^2$

$L_2(\tilde{\lambda}, \lambda) = (\tilde{\lambda} - \lambda)^2$

Def. A risk is the expectation of a loss function

$$R(\hat{\lambda}) = \mathbb{E}_\theta \left[L(\hat{\lambda}, \lambda) \right]$$

Remarks

1) Quadratic risk is linked to the quadratic loss function.

2) $R(\hat{\lambda}) = Var_\theta(\hat{\lambda}) + (Bias(\hat{\lambda}))^2$

Def: 1) $\hat{\lambda}$ is better than $\tilde{\lambda}$ for R if and only if $\forall \theta \in \Theta : R(\hat{\lambda}) \leq R(\tilde{\lambda})$

2) —— strictly better —— $<$ ——

Usually it is not easy to optimize a risk function.

In this lecture we mainly consider quadratic risk.

2. How to construct interesting estimators?

2.1. Unbiased linear estimator

Let $X_1, \dots, X_n \overset{i.i.d}{\sim} P_\theta$, $\theta \in \Theta$ is unknown.

Assume that for all $\theta \in \Theta$, the moment $m_\theta = \mathbb{E}_\theta[X]$ and $\mathbb{E}_\theta[X^2]$ are finite. $\sigma_\theta^2 = Var_\theta(X) = \mathbb{E}_\theta[X^2] - (\mathbb{E}_\theta[X])^2$

We consider linear estimator of $\lambda = m_\theta$:

$$\hat{\lambda} = b + \sum_{i=1}^{n} a_i X_i$$

Proposition For a n-sample with at least second order moment, the arithmetical empirical mean $\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ is the best unbiased linear estimator of m_θ for the quadratic risk. (if there exists $\theta_1, \theta_2 : m_{\theta_1} \neq m_{\theta_2}$)

Proof To get an unbiased estimator, we want $\mathbb{E}_\theta[\hat{\lambda}] = \lambda$.

. Now $\mathbb{E}_\theta[\hat{\lambda}] = \mathbb{E}_\theta\left[b + \sum_{i=1}^n a_i X_i\right]$

$= b + \sum_{i=1}^n a_i \, \mathbb{E}_\theta[X_i]$

$= b + \sum_{i=1}^n a_i \, \mathbb{E}_\theta[X]$

$= b + m_\theta \cdot \sum_{i=1}^n a_i$

. For our estimator to be unbiased we want:

$$b + m_\theta \cdot \sum_{i=1}^n a_i = m_\theta \quad \text{for any } \theta \in \Theta \quad (*)$$

. Now take θ_1, θ_2 such $m_{\theta_1} \neq m_{\theta_2}$, we have:

$$\begin{cases} m_{\theta_1} = b + m_{\theta_1} \cdot \sum_{i=1}^n a_i \\ m_{\theta_2} = b + m_{\theta_2} \cdot \sum_{i=1}^n a_i \end{cases} \Leftrightarrow \begin{cases} m_{\theta_1} = b + m_{\theta_1} \cdot \sum_{i=1}^n a_i \\ m_{\theta_2} - m_{\theta_1} = (m_{\theta_2} - m_{\theta_1}) \cdot \sum_{i=1}^n a_i \end{cases}$$

$$\Leftrightarrow \begin{cases} b = 0 \\ \sum_{i=1}^n a_i = 1 \end{cases}$$

. Then $\hat{\lambda} = \sum_{i=1}^n a_i X_i$ where $\sum_{i=1}^n a_i = 1$.

. The quadratic risk is then:

$$\mathbb{E}_\theta\left[(\hat{\lambda} - \lambda)^2\right] = \mathrm{Var}_\theta(\hat{\lambda}) + \underbrace{\left(\mathrm{Bias}_\theta(\hat{\lambda})\right)^2}_{=0 \text{ (by unbiased)}}$$

$$= \mathrm{Var}_\theta(\hat{\lambda})$$

$$\mathrm{Var}_\theta(\hat{\lambda}) = \mathrm{Var}_\theta\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \, \mathrm{Var}_\theta(X) = \sum_{i=1}^n a_i^2 \cdot \sigma_\theta^2$$

Since

$$1 = 1^2 = \left(\sum a_i\right)^2 \overset{\text{(Cauchy-Schwarz)}}{\leq} \left(\sum a_i^2\right) \cdot \left(\sum 1^2\right) = n \cdot \sum a_i^2$$

$\rightarrow \sum a_i^2 \geq \frac{1}{n}$.

Then the quadratic risk is $R_\theta(\lambda) \geq \frac{1}{n} \cdot \sigma_\theta^2$

If we take $a_i = 1/n$ for $0 \leq i \leq n$ then $\sum a_i^2 = 1/n$

↳ Mean is the best unbiased estimator (linear) □

. **Remark.** In Cauchy Schwarz inequality, we have the equality iff

$(a_1, a_2, \dots, a_n) = k.(1, \dots, 1)$ for some value $k \in \mathbb{R}$,

which means

$\forall i : 1 \leq i \leq n \; : \; a_i = k$.

There is no other minimizer.

2.2. Method of moments.

- Starting point: Strong Law of Large Numbers
- We will use this with $\varphi : x \longmapsto x^p$

if $\mathbb{E}[|X|^p] < +\infty$

p^{th} empirical moment (fonction) $\overline{X^p} = \frac{X_1^p + \dots + X_n^p}{n} \xrightarrow[n \to +\infty]{P_\theta \text{ a.s}} \mathbb{E}[X^p] =: \mu_p(\theta)$ p^{th} theoretical moment

- **Algorithm**

1) Compute the first moment μ_1, μ_2 — (how much time we can till the final boss which μ_k) in function of λ.
2) Solve the equation to get λ in functions of $\mu_1, \mu_2 \dots$
3) Get $\hat{\lambda}$ by replacing μ_p by $\overline{X^p}$

Come back to examples

1) $X \sim \text{Ber}(\theta)$.

1) $\mu_1 = \mathbb{E}_\theta[X] = \theta$.

2) $\theta = \mu_1 = \mathbb{E}_\theta[X]$

3) $\hat{\theta} = \overline{X}$

2) $X \sim G(p)$ — đếm số lần thất bại trước khi thành công

$\mathbb{P}(X = k) = (1-p)^k \cdot p$

1) $\mu_1 = \mathbb{E}_\theta[X] = \frac{1-p}{p} = \frac{1}{p} - 1$.

2) $p = \frac{1}{\mu_1 + 1} = \frac{1}{\mathbb{E}_\theta[X] + 1}$

3) $\hat{p} = \frac{1}{\overline{X} + 1}$.

3) $X \sim \mathcal{N}(\mu, \sigma)$ $\quad \lambda = (\mu, \sigma^2)$

1) $\begin{cases} \mu_1 = \mathbb{E}_\theta[X] = \mu \\ \mu_2 = \mathbb{E}_\theta[X^2] = \mu^2 + \sigma^2 \end{cases}$

2) $\begin{cases} \mu = \mu_1 \\ \sigma^2 = \mu_2 - \mu_1^2 \end{cases}$

3) $\begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = \overline{X^2} - (\overline{X})^2 \end{cases}$

4) $X \sim \mathcal{E}(\theta)$: phân phối $f_X(x) = \theta . e^{-\theta x} \mathbb{1}_{x \geq 0}$.

1) $\mu_1 = \mathbb{E}_\theta[X] = \frac{1}{\theta}$

2) $\theta = \frac{1}{\mu_1}$

3) $\hat{\theta} = \frac{1}{\overline{X}}$

↳ nhanh nhưng ≈ tốt bằng maximum likelihood.

2.3. Maximum likelihood

We assume to have a dominated model.

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

in which for all $\theta \in \Theta$, $\mathbb{P}_\theta \ll Q$. (with Q : σ - finite positive measure)

Examples 1) $\Theta = \{\theta_1, \theta_2\}$. $\quad Q = \mathbb{P}_{\theta_1} + \mathbb{P}_{\theta_2}$.

then $\mathbb{P}_{\theta_1} \ll Q$ (đọc thuyết thôi: $Q(A) = 0 \Rightarrow \mathbb{P}_{\theta_1}(A) = 0$)

$\mathbb{P}_{\theta_2} \ll Q$.

2) $\Omega_n (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$.

$Q = \lambda$ Lebesgue measure.

$$d\mathbb{P}_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} \, d\lambda(x).$$

Def. The likelihood of a random sample $X_1, \dots, X_n$ is given by

$$\theta \longmapsto L(\theta, X_1, \dots, X_n) = \prod_{i=1}^{n} h(x_i, \theta)$$

hàm mật độ xs f_X

Remark. For all θ, $L(\theta, X_1, \dots, X_n)$ is a RV.

Def. The Maximum likelihood estimator of λ is the value of λ maximizing

$$\lambda \longmapsto L(\theta, X_1, \dots, X_n)$$

$$\hat{\lambda}_{MLE} = \arg\max_{\lambda} L(\theta, X_1, \dots, X_n)$$

- **Remark.**

  1) $\hat{\lambda}$ may be non unique or may not exist.

  2) In practice, when for all $\theta \in (H)$, almost surely $L(\theta, X_1, \dots, X_n) > 0$, we take the logarithm:

  $$LL(\theta, X_1, \dots, X_n) = \ln\left(L(\theta, X_1, \dots, X_n)\right) = \sum_{i=1}^{n} \ln\left(h(x_i, \theta)\right)$$

  . Then we usually search λ such that: $(\lambda \in \mathbb{R})$

  $$\frac{\partial}{\partial \lambda} LL(\theta, X_1, \dots, X_n) = 0$$

  $$\frac{\partial^2}{\partial \lambda^2} LL(\theta, X_1, \dots, X_n) < 0$$

20. 10. 2022.

Examples

1) $X \sim Ber(\theta)$, $\theta \in [0;1]$

$f(x;\theta) = \theta^x (1-\theta)^{1-x}$.

$L(\theta, X_1, \dots, X_n) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$

$= \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}$

$= \theta^{n\bar{x}} \cdot (1-\theta)^{n(1-\bar{x})}$

$\Rightarrow LL(\theta, X_1, \dots, X_n) = \ln\left(\theta^{n\bar{x}}\right) + \ln\left((1-\theta)^{n(1-\bar{x})}\right)$

$= n\bar{x} \cdot \ln\theta + n(1-\bar{x}) \cdot \ln(1-\theta)$.

$\frac{\partial}{\partial\theta} LL(\theta, X_1, \dots, X_n) = \frac{n\bar{x}}{\theta} - \frac{n(1-\bar{x})}{1-\theta}$.

$= \frac{n}{\theta(1-\theta)} \left[\bar{x}(1-\theta) - \theta(1-\bar{x})\right]$

$= \frac{n(\bar{x}-\theta)}{\theta(1-\theta)}$

$\frac{\partial}{\partial\theta} LL(\theta, X_1, \dots, X_n) = 0 \hookrightarrow \bar{x} = \theta$.

$\frac{\partial^2}{\partial\theta^2} LL \text{ ——— } = \frac{-n\bar{x}}{\theta^2} - \frac{n(1-\bar{x})}{(1-\theta)^2} < 0$

Hence $\hat{\theta}_{MLE} = \bar{X}$

⚠ MLE is better than moment estimator.
(thường ra kquả giống nhau).

2) $X \sim \mathcal{E}(\lambda)$ $\quad f(x,\lambda) = \lambda . e^{-\lambda x} \mathbb{1}_{x \geq 0}$

$$L(\lambda, x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i,\lambda)$$

$$= \lambda^n . e^{-\lambda \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \mathbb{1}_{x_i \geq 0}$$

$$= \lambda^n . e^{-n\lambda \bar{x}} \boxed{\prod_{i=1}^{n} \mathbb{1}_{x_i \geq 0}} \quad .$$

$$= 1 \text{ a.s.}$$

$$= \lambda^n . e^{-n\lambda \bar{x}} . \mathbb{1}_{[\min x_i \geq 0]}$$

$$. LL(\lambda, x_1, \dots, x_n) = n . \ln\lambda - n\lambda \bar{x} .$$

$$\frac{\partial}{\partial \lambda} LL \text{ ——————} = \frac{n}{\lambda} - n\bar{x}$$

$$\text{——————} = 0 \iff \lambda = 1/\bar{x}$$

$$\frac{\partial^2}{\partial \lambda^2} \text{ ——————} = \frac{-n}{\lambda^2} < 0$$

$$\Rightarrow \hat{\lambda}_{MLE} = 1/\bar{x} .$$

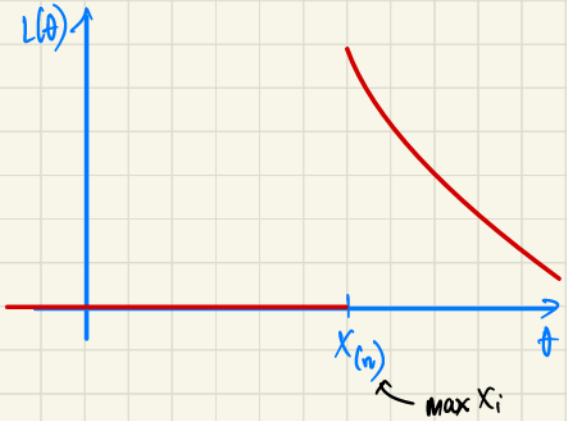
3) $X \sim \mathcal{U}[0,\theta]$ $\quad , f(x,\theta) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x) .$

$$L(\theta, x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x_i)$$

$$= \frac{1}{\theta^n} \cdot \prod_{i=1}^{n} \mathbb{1}_{\{x_i \geq 0\}} \cdot \mathbb{1}_{\{x_i \leq \theta\}} \cdot \longrightarrow \mathbb{1}_{[A \cap B]} = \mathbb{1}_A \cdot \mathbb{1}_B$$

$$= \frac{1}{\theta^n} \cdot \underbrace{\prod_{i=1}^{n} \mathbb{1}_{\{x_i \geq 0\}}}_{= 1 \text{ a.s.}} \cdot \prod_{i=1}^{n} \mathbb{1}_{\{x_i \leq \theta\}}$$

$$= \frac{1}{\theta^n} \cdot \mathbb{1}_{\{\max x_i \leq \theta\}}$$



$\Rightarrow \hat{\theta}_{MLE} = X_{(n)} = \max_i X_i$

But Moments Methods. 1) $\mu_1 = \mathbb{E}_\theta[X] = \frac{\theta}{2}$.

2) $\theta = 2\mu_1$

3) $\hat{\theta} = 2 \cdot \bar{X}$

→ Moments Methods give the different estimator from MLE.

4) $X \sim \mathcal{N}(\mu, \sigma)$. , $\theta = (\mu, \sigma)$ → ML μ and σ.

$$f(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$L(\theta, X_1, \dots, X_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-(X_i-\mu)^2/2\sigma^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}$$

$$\underbrace{LL(\theta, X_1, \dots, X_n)}_{f(\theta)} = n \cdot \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2$$

$$= -n\ln\sigma - n\ln(\sqrt{2\pi}) - \sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2$$

∇f : $\frac{\partial f}{\partial \mu}(\theta) = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n}(X_i-\mu) = \frac{1}{\sigma^2}(n\bar{X} - n\mu) = \frac{n}{\sigma^2} \cdot (\bar{X}-\mu)$.

$$\frac{\partial f}{\partial \sigma}(\theta) = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\cdot\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{n}{\sigma}\left[-1 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\cdot\frac{1}{\sigma^2}\right]$$

- $\nabla f(\theta) = 0 \Leftrightarrow \begin{cases} \mu = \bar{X} \\ -1 + \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\mu)^2\cdot\frac{1}{\sigma^2} = 0 \end{cases}$

$$\Leftrightarrow \begin{cases} \mu = \bar{X} \\ \sigma = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{cases}$$

- $Hf(\theta)$: $\frac{\partial^2 f}{\partial\mu\,\partial\sigma}(\theta) = \frac{-2}{\sigma^3}n\cdot(\bar{X}-\mu)$

$$\frac{\partial^2 f}{\partial\mu^2}(\theta) = \frac{-n}{\sigma^2}$$

$$\frac{\partial^2 f}{\partial\sigma^2}(\theta) = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2. \longrightarrow \text{Cm các trị riêng} < 0$$

Claim: $Hf\left(\bar{x}, \underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}}_{\hat{\sigma}}\right) < 0$ ← Chỉ cần check tại $\hat{\theta}$.

- If $\mu = \bar{X}$, $\sigma = \hat{\sigma}$

$$\frac{\partial^2}{\partial\sigma^2}f(\theta) = \frac{n}{\hat{\sigma}^2} - \frac{3}{\hat{\sigma}^4}n\cdot\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}_{\hat{\sigma}^2}$$

$$= \frac{n}{\hat{\sigma}^2} - \frac{3n}{\hat{\sigma}^2} = \frac{-2n}{\hat{\sigma}^2}$$

$$\Rightarrow \quad Hf\left((\bar{X}, \hat{\sigma})\right) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{2n}{\hat{\sigma}^2} \end{pmatrix} \quad : \text{negative definite}$$

Hence $\hat{\theta}_{MLE} = (\bar{X}, \hat{\sigma})$

- **Theorem** In an identifiable and dominated statistical model, if:

  1) $\theta \longmapsto \ln\left(f(x,\theta)\right)$ is continuous for all x.

  2) (H) is a compact subset of $\mathbb{R}^k$

  Then for all $\theta \in (H)$, $\hat{\theta}_{MLE} \xrightarrow[n\to+\infty]{P_\theta \text{ a.s}} \theta$.

- **Remarks**

  1) There exists a kind of CLT for MLE.

  2) In practice, if (H) is not compact, it is possible to apply this Theorem to an increasing sequence of compact sets increasing to (H) and finally, get the same result

  3) Usually MLE are biased.

  But under some assumptions it is possible to show that MLE are asymptotically unbiased.

4. Fisher information, Cramer Rao Lower Bound and Efficiency.

4.1) Fisher information.

. Let $P_{(\Theta)} = \{ P_\theta, \theta \in (\Theta) \}$ be a parametric and dominated statistical model

. Let $\hat{\lambda}$ be an estimator of $\lambda = g(\theta)$.

Def. A model is said to be regular iff

1) for almost all x, $\theta \longmapsto h(x,\theta)$ is differentiable.

2) It is possible to invert $\frac{\partial}{\partial\theta}$ and $\int$ in $\frac{\partial}{\partial\theta} \mathbb{E}_\theta[h(\theta, X)]$

3) $I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \ln h(x,\theta)\right)^2\right]$ exists and has an inverse.

↑ Nếu θ là vector thì đây là ma trận.

$I_{ij}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i} \ln(h(x,\theta)) \cdot \frac{\partial}{\partial\theta_j} \ln(h(x,\theta))\right] \rightarrow I_{ij}(\theta)$ has an inverse

↑ từng phần của matrận I.

Def. When the integral exists, the Fisher Information of the model is

- for one value X

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \ln(h(x,\theta))\right)^2\right]$$

- for a random sample $X_1, X_2, \ldots, X_n$:

$$I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \ln L(\theta, X_1, \ldots, X_n)\right)^2\right]$$

. **Remarks :**

1) $I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} LL(\theta, X_1, \dots, X_n)\right)^2\right]$

2) When θ is a vector $I(\theta)$ is a matrix

3) $I_n(\theta) = n.I(\theta)$.

Example $X \sim \mathcal{E}(\lambda)$

$$LL(\theta, X) = \ln(\lambda) - \lambda.X$$

$$\frac{\partial}{\partial\lambda} LL(\theta, X) = \frac{1}{\lambda} - X.$$

$$\begin{aligned} . I(\lambda) &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\lambda} LL(\theta, X)\right)^2\right] \\ &= \mathbb{E}_\theta\left[\left(\frac{1}{\lambda} - X\right)^2\right] \\ &= \mathbb{E}_\theta\left[\frac{1}{\lambda^2} - 2.\frac{1}{\lambda}.X + \bar{X}^2\right] \\ &= \frac{1}{\lambda^2} - \frac{2}{\lambda}\mathbb{E}_\theta[X] + \mathbb{E}_\theta[\bar{X}^2] \\ &= \frac{1}{\lambda^2} - \frac{2}{\lambda}.\frac{1}{\lambda} + \frac{2}{\lambda^2} = \frac{1}{\lambda^2} \end{aligned}$$

$\Rightarrow$ Hence $I_n(\lambda) = n.I(\lambda) = \frac{n}{\lambda^2}$.

• **Proposition.** Fisher information does not depend on dominating measure Q.

• **Proposition.** Under the following assumptions:

1) $\frac{\partial}{\partial\theta}\int h(x,\theta)\,dQ(x) = \int \frac{\partial}{\partial\theta} h(x,\theta)\,dQ(x)$

2) $\frac{\partial^2}{\partial\theta^2}\int h(x,\theta)\,dQ(x) = \int \frac{\partial^2}{\partial\theta^2} h(x,\theta)\,dQ(x)$.

} this we can check the.

We get

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ln\left(h(X,\theta)\right)\right] = 0$$

and

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln\left(h(X,\theta)\right)\right]$$

4.2. Cramer–Rao Blackwell Lower Bound.

Let T be a statistic such that

1) $\mathbb{E}_\theta[T^2] < +\infty$

2) T is an unbiased estimator of $\lambda = g(\theta)$: $\mathbb{E}_\theta[T] = \lambda$.

Assume the model is regular then we get the following theorem:

• **Theorem** Cramer–Rao inequality.

Under previous assumptions:

$$\mathrm{Var}_\theta(T) \geq \frac{|g'(\theta)|^2}{I_n(\theta)} \quad \text{if } \theta \in \mathbb{R},\ \lambda \in \mathbb{R}.$$

$$\cdot \mathrm{Var}_\theta(T) \geq J_g(\theta) \cdot I_n^{-1}(\theta) \cdot J_g(\theta)^T \quad \text{if } \theta, \lambda \text{ is vector.}$$

Rang VT – VP de matrice est ↑.

where $g : \mathbb{R}^k \longrightarrow \mathbb{R}^m$ $\quad g(\theta) = (g_1(\theta), \dots, g_n(\theta))$

$$J_g(\theta) = \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \vdots \\ \vdots & \vdots \\ \cdots & \frac{\partial g_i(\theta)}{\partial \theta_j} \end{pmatrix}$$

· Remarks

1) When $g(\theta) = \theta$ then we get $\mathrm{Var}_\theta(T) \geq \frac{1}{I_n(\theta)}$

2) Because $I_n(\theta) = n \cdot I(\theta)$, we get:

$$\mathrm{Var}(T) \geq \frac{|g'(\theta)|^2}{n \cdot I(\theta)} = \frac{\text{Constant}}{n}.$$

4.3. Efficiency.

· **Def.** When the assumptions of the previous theorem hold the estimator T is said to be **efficient** iff the lower bounded is archieved:

$$\mathrm{Var}_\theta(T) = \frac{|g'(\theta)|^2}{I_n(\theta)}.$$

- T is hence an **unbiased estimator** of $\lambda = g(\theta)$ with **minimal variance**
- **Best estimator** of λ (with **respect** to the **quadratic risk**).

. **Theorem** Under the previous regularity assumptions the statistic T is an efficient estimator of $\lambda = g(\theta)$ if:

$$\frac{\partial}{\partial \theta} LL_n(\theta, X_1, \dots, X_n) \overset{\text{as}}{=} c(\theta) \cdot [T(X) - g(\theta)]$$

Back to example: 1) $X \sim \text{Ber}(\theta)$.

$$\frac{\partial}{\partial \theta} LL_n(\theta, X_1, \dots, X_n) = \underbrace{\frac{n}{\theta(1-\theta)}}_{c(\theta)} [\underbrace{\bar{X}}_{T(X)} - \underbrace{\theta}_{g(\theta)}] \rightarrow \bar{X} \text{ : efficient.}$$

. **Corollary**. If the model is regular and the MLE exists, then if T is efficient, unbiased estimator of λ then it is the MLE.

(dli)

5. **Exhaustive statistics** (sufficient statistic).

Def. A statistic T is said to be exhaustive if the conditional distribution of $(X_1, \dots, X_n)$ given T does not depend on θ.

Example. Let $(X_1, \dots, X_n) \overset{\text{i.i.d}}{\sim} \mathcal{P}(\theta)$.

$$T = \sum_{i=1}^{n} X_i$$

We know that $T \sim \mathcal{P}(n\theta)$.

$$P_\theta((X_1, \_, X_n) = (x_1, \_, x_n) \mid T = t)$$

$$= P_\theta(\_\_\_\_\_, T = t) / P_\theta(T = t).$$

$$= \begin{cases} \dfrac{P_\theta((X_1,\dots,X_n)=(x_1,\dots,x_n)}{P(T=t)} & \text{if } t=\sum_{i=1}^n x_i \\ 0 \quad \text{if } t \neq \sum_{i=1}^n x_i \end{cases}$$

$$= \begin{cases} \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} \cdot e^{-\theta} \Big/ \frac{(n\theta)^{\sum_{i=1}^n x_i}}{(\sum x_i)!} \cdot e^{-n\theta} & \text{if } t=\sum_{i=1}^n x_i \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} \frac{\theta^{\sum_{i=1}^n x_i}}{\prod x_i!} \cdot e^{-n\theta} \Big/ \frac{(n\theta)^{\sum x_i}}{(\sum x_i)!} e^{-n\theta} & \text{if } t=\sum x_i \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} \dfrac{(\sum x_i)!}{\prod x_i! \cdot n^{\sum x_i}} & \text{if } t=\sum x_i \\ 0 & \text{else} \end{cases}$$

does not depend on θ $\longrightarrow$ exhaustive.

21. 10. 2022.

. Theorem: Let T be a statistic computed from the sample $(X_1, \dots, X_n)$.

T is said to be exhaustive if and only if the likelihood of the sample may be factorized in the following way.

$$L(\theta, X_1, \dots, X_n) = \Psi(T, \theta) \cdot \ell(X_1, \dots, X_n)$$

. Remarks

1) An other way to define an exhaustive statistic is $\forall h$ measurable function $\mathbb{E}_\theta[h(X_1, \dots, X_n) | T]$ does not depend on θ.

2) If there exists an exhaustive statistic T given by the factorization theorem then the MLE estimator is a function of T.

. Examples 1) $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\theta, 1)$, $T = \sum_{i=1}^{n} X_i$

$$. L[\theta, X_1, X_2, \dots, X_n] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot e^{-(x_i - \theta)^2/2}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot e^{-\sum_{i=1}^{n} (x_i^2 + \theta^2 - 2x_i\theta)/2}.$$

$$= (\sqrt{2\pi})^{-n} \cdot e^{-1/2\left[\sum_{i=1}^n x_i^2 + n\theta^2 - 2\theta \cdot \sum x_i\right]}.$$

$$= \underbrace{(\sqrt{2\pi})^n \cdot e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}}_{g(x_1 - x_n)} \cdot \underbrace{e^{-\frac{1}{2}n\theta^2 + \theta T}}_{\psi(T,\theta)}.$$

Hence T is exhaustive.

2) $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma)$, (θ, σ)

$$L(\theta; X_1, \dots, X_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-(x_i - \theta)^2/2\sigma^2}$$

$$= \left(\sqrt{2\pi}\cdot\sigma\right)^{-n} \cdot e^{-\sum (x_i^2 - 2x_i\theta + \theta^2)/2\sigma^2}$$

$$= (\sqrt{2\pi})^{-n} \cdot \sigma^{-n} \cdot e^{-\left(\sum x_i^2 - 2\theta\sum x_i + n\theta^2\right)/2\sigma^2}$$

$$= (\sqrt{2\pi})^{-n} \cdot \sigma^{-n} \cdot e^{-n\theta^2/2\sigma^2} \cdot e^{n\theta\bar{X}/\sigma^2} \cdot e^{-n\cdot\overline{X^2}/2\sigma^2}$$

$$= \psi\left((\theta, \sigma), T\right);$$

where $T = \begin{pmatrix} \bar{x} \\ \overline{x^2} \end{pmatrix}$. We have seen that: $\hat{\theta}_{MLE} = \bar{X}$

$$\hat{\sigma}_{MLE} = \sqrt{\overline{X^2} - (\bar{X})^2}$$

↳ function of T.

3) $X_1, \dots, X_n \overset{i.i.d}{\sim} U[0, \theta]$, $\theta > 0$.

$$L(\theta, X_1, \dots, X_n) = \prod_{i=1}^n \frac{1}{\theta} \cdot \mathbb{1}_{[0,\theta]}(x_i)$$

$$= \underbrace{\frac{1}{\theta^n} \cdot \mathbb{1}_{\{\max x_i \le \theta\}}}_{\psi(T,\theta)} \cdot \mathbb{1}_{\{\min x_i \ge 0\}}$$

$T = \max X_i$ is an exhaustive statistic.

. **Theorem** Rao – Blackwell.

Let $\hat{\lambda}$ be an unbiased estimator of λ and assume $\mathbb{E}_\theta[\hat{\lambda}] < +\infty$.

If there exists an exhaustive T then $\hat{\lambda}_T = \mathbb{E}_\theta[\hat{\lambda}|T]$ is a better estimator of λ than $\hat{\lambda}$ for the quadratic risk

. **Remarks:**

- $\hat{\lambda}_T$ is the projection (= limit distribution) of $\hat{\lambda}$ on the set of measurable functions of T ($\mathbb{E}[\hat{\lambda}|T] = g(T)$). If $\hat{\lambda}$ is already a measurable function T then $\hat{\lambda}_T = \mathbb{E}[\hat{\lambda}|T] = \hat{\lambda}$.
- $\mathbb{E}[\hat{\lambda}_T] = \mathbb{E}[\mathbb{E}[\hat{\lambda}|T]] = \mathbb{E}[\hat{\lambda}] = \lambda$ if $\hat{\lambda}$ is unbiased.

. **Example.**

1) $X_1, \dots, X_n \overset{i.i.d}{\sim} \text{Ber}(\theta)$, $\theta \in]0,1[$.

$\hat{\theta}_1 = X_1$: unbiased estimator of θ.

. $L(\theta, X_1, \dots, X_n) = \theta^{n\bar{x}} \cdot (1-\theta)^{n-n\bar{x}}$

Hence $T = \bar{X}$ is an exhaustive statistic.

. Now the distribution for all t of $X_1 | T = t$ is $\text{Ber}(t)$

. Then $\mathbb{E}[X_1|T] = T = \bar{X}$. Indeed:

$$\mathbb{P}(X_1 = x_1 | T = t) = \frac{\mathbb{P}(X_1 = x_1,\ T = t)}{\mathbb{P}(T = t)}$$

$$= \frac{\mathbb{P}(X_1 = x_1, \sum X_i = nt)}{\mathbb{P}(\sum X_i = nt)}$$

$$= \sum_{x_2, \dots, x_n} \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, \sum X_i = nt)}{\mathbb{P}(\sum X_i = nt)} \quad \sum X_i \sim \text{Bin}(\theta).$$

$$= \sum_{\substack{x_2, \dots, x_n \\ \sum_{i=1}^n x_i = nt}} \frac{\theta^{nt} \cdot (1-\theta)^{n-nt}}{\binom{n}{nt} \cdot \theta^{nt} \cdot (1-\theta)^{n-nt}}$$

$$= \begin{cases} \binom{n-1}{nt} \Big/ \binom{n}{nt} & \text{if } x_1 = 0 \\ \binom{n-1}{nt-1} \Big/ \binom{n}{nt} & \text{if } x_1 = 1. \end{cases}$$

$$\frac{(n-1)!}{(nt)!\,(n-1-nt)!} \Big/ \frac{n!}{(nt)!\,(n-nt)!} = \frac{n-nt}{n} = 1-t$$

$$\frac{(n-1)!}{(nt-1)!\,(n-nt)!} \Big/ \frac{n!}{(nt)!\,(n-nt)!} = \frac{nt}{n} = t.$$

$$= \begin{cases} 1-t & \text{if } x_1 = 0 \\ t & \text{if } x_1 = 1. \end{cases}$$

Hence $\mathbb{P}(X_1 = x \mid T = t) = \begin{cases} 1-t & \text{if } x_1 = 0 \\ t & \text{if } x_1 = 1. \end{cases}$

$\Rightarrow X_1 \mid T = t \sim \text{Bar}(t).$

2) $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\theta, 1)$

. X_1 is an unbiased estimator of θ.

$\mathbb{E}[X_1] = \theta.$

. $\bar{X}$ exhaustive

↳ $\mathbb{E}[X_1 | \bar{X}] = \varphi(\bar{X}) = \mathbb{E}[X_i | \bar{X}]$ (Idea)

. $\varphi(\bar{X}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}[X_i | \bar{X}] = \mathbb{E}\left[\frac{1}{n} \cdot \sum X_i \mid \bar{X}\right] = \mathbb{E}[\bar{X} | \bar{X}] = \bar{X}$

(vì tính chất $\mathbb{E}[X_i | \bar{X}] = \text{chung}$)

$\Rightarrow \mathbb{E}[X_1 | \bar{X}] = \bar{X}$ is better estimator of θ.

6) Exponential families

. **Def.** A family $\mathcal{P}_{(\Theta)} = \{P_\theta, \theta \in (\Theta)\}$ is an exponential family if the density of P_θ is the following form:

$$p(x, \theta) = e^{\langle \alpha(\theta), T_0(x) \rangle + \beta(\theta)} \cdot h(x)$$

where $\begin{cases} \alpha : (\Theta) \to \mathbb{R}^k \\ \beta : (\Theta) \to \mathbb{R} \\ T_0 : \mathcal{X} \to \mathbb{R}^k \\ \langle \cdot, \cdot \rangle : \mathbb{R}^k \text{ inner product.} \end{cases}$

. **Property:** If $X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} P_\theta$, with $P_\theta \in \mathcal{P}_{(\Theta)}$ and $\mathcal{P}_{(\Theta)}$ is an exponential family then $T = \frac{1}{n} \cdot \sum_{i=1}^{n} T_0(X_i)$ is an exhaustive statistic.

Proof

$$L(\theta, X_1, \dots, X_n) = \prod_{i=1}^{n} e^{\langle \alpha(\theta), T_0(X_i) \rangle + \beta(\theta)} \cdot h(X_i)$$

$$= \left[\prod_{i=1}^{n} \ell(x_i) \right] \cdot e^{\sum_{i=1}^{n} [\langle \alpha(\theta), T(x_i) \rangle + \beta(\theta)]}$$

$$= \prod_{i=1}^{n} \ell(x_i) \cdot e^{\langle \alpha(\theta), \sum T(x_i) \rangle + n \cdot \beta(\theta)}$$

$$= \prod_{i=1}^{n} \ell(x_i) \cdot \underbrace{e^{\langle \alpha(\theta), nT \rangle + n\beta(\theta)}}_{\Psi(T, \theta)}$$

↳ T is exhaustive. □

- **Theorem. Lehman Scheffé.**

If the model is exponential and $\alpha(\Theta)$ contains an open subset of $\mathbb{R}^k$, then for any function $f: \mathbb{R}^k \to \mathbb{R}^l$ such that $\lambda = \mathbb{E}_\theta[\|f(T)\|] < +\infty$, the statistic $f(T)$ is the best unbiased estimator of $\mathbb{E}_\theta[f(T)]$.

- **Comment**

If you want to estimate $\lambda = g(\theta)$ under the assumption of the previous theorem, if you define an estimator $\hat{\lambda}$ of λ that is unbiased and is written as a function of T (with $\mathbb{E}[\|\hat{\lambda}\|] < +\infty$) then $\hat{\lambda}$ is the best estimator of λ.

- **Def.** A statistic T is said to be complete iff.

$\forall \theta \in \Theta$, $\forall g$ borel measurable function

$$\left(\mathbb{E}_\theta[g(T)] = 0\right) \Rightarrow \mathbb{P}_\theta\left(g(T) = 0\right) = 1.$$

• **Property.** In an exponential family, if $\alpha(\Theta)$ contains a non empty open subset of $\mathbb{R}^k$, then T is complete.

↳ T exhaustive si vrai.

Idea:

$$\mathbb{E}_\theta[g(T)] = \int g(T) \cdot e^{\langle \alpha(\theta), T\rangle + \beta(\theta)} \cdot \ell(x_1, \dots, x_n)\, dQ(x_1)\, dQ(x_2) \dots dQ(x_n)$$

$$\overset{"}{=}" \int g(t) \cdot e^{\langle \alpha(\theta), t\rangle + \beta(\theta)} \cdot \int \frac{\ell(x_1, \dots, x_n) \cdot d\ell(x_1, \dots, x_n \mid T = t)}{dQ_\theta(t)}$$

$$= \int \underbrace{g(t) \cdot p(t)}_{m(t)} \cdot e^{\langle \alpha(\theta), t\rangle + \beta(\theta)}\, dQ_\theta(t) \qquad p(t)$$

$$= \int m(t) \cdot e^{\langle \alpha(\theta), t\rangle}\, dQ_\theta(t).$$

Property of Laplace transform, if Λ has non empty open subset that:

$$\left[\text{for all } \lambda \in \Lambda,\ \int e^{\langle \lambda, t\rangle} m(t)\, dQ_\theta(t) = 0\right] \Rightarrow m = 0 \quad Q_\theta \text{ a.s.}$$

• **Theorem.** Lehman - Scheffé.

If T is an exhaustive and complete statistic

$\hat{\lambda}$ is an unbiased estimator of λ.

then $\hat{\lambda}_T = \mathbb{E}[\hat{\lambda} \mid T]$ is a uniform minimum variance unbiased estimator of λ.