Parametric estimation

Michaël Baudin (avec l'aimable autorisation de Jean-Marc Martinez)

29 novembre 2022

Note : Sections with an asterisk mark (*) can be skipped on a first reading.
In this part, we focus on the step B of the ABC method. Using a sample of the input $\mathbf{X}$, we want to estimate the parameters of the input distribution.

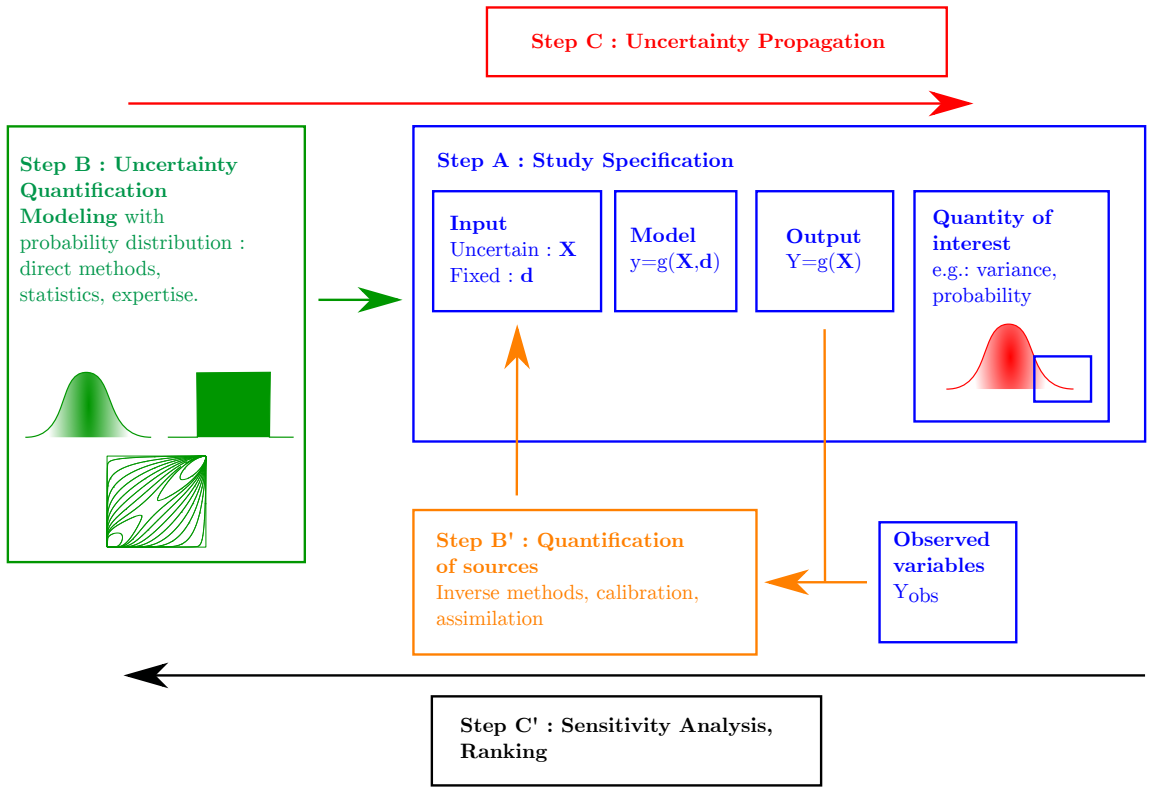

Figure 1 – Steps in the ABC method.

In some cases, some of these methods can be used in the step C as well, e.g. estimate the parameters of a model which fits the distribution of the output Y.

Two methods are often used.

- Maximum likelihood estimation (MLE) : selects the set of values of the model parameters that maximizes the likelihood function. This function measures the "agreement" of the selected model with the observed data.
- Method of moments (MoM) : finds the value of the parameters which are so that the sample moments are equal to the moments of the distribution.

Introduction

- The maximum likelihood estimator is a parametric tool.
- Let X be a random variable with probability density function $f(x, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^m$ is an unknown vector and x is an observation of X.

Définition 1 (Parametric model)

Let $\Theta \subset \mathbb{R}^m$ be the parameter space. The set $\mathcal{P}_{\boldsymbol{\theta}}$ is a parametric model if :

$$\mathcal{P}_{\boldsymbol{\theta}} = \{f(x, \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\},$$

where f is a discrete probability distribution function or a continuous probability density function.

Exemple 2 (Parameters of the normal distribution)

We have $m = 2$ parameters corresponding to the parameters of the Gaussian distribution : $\mu \in \mathbb{R}$ and $\sigma \in]0, +\infty[$.

Hypothèse 1

Let X be a random variable with probability density function $f(x, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta$ is the vector of parameters and $x \in \mathbb{R}$ is an observation of X. Let $x_1, \ldots, x_n$ be a sample of n independent observations of X.

The most likely value of $\boldsymbol{\theta}$ is the value for which the probability density of observing the random vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is the largest.
The following definition introduces the likelihood [1].

Définition 3 (Likelihood)

The likelihood is the function $\mathcal{L}$ defined as the probability density of the vector $(x_1, \ldots, x_n)$:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \tag{1}$$

for any $\boldsymbol{\theta} \in \Theta$.

The problem is *inverse* : instead of seeing $\boldsymbol{\theta}$ as known and estimate the density f at observations x_1, ..., x_n, we consider that the sample is set and we search for $\boldsymbol{\theta}$.

1. [Wasserman, 2004] page 122

Hypothèse 2

La vraisemblance est toujours strictement positive :

$$\mathcal{L}(\boldsymbol{\theta}) > 0$$

pour tout $\boldsymbol{\theta} \in \Theta$.

- Q : quand la vraisemblance est-elle nulle ?

Le logarithme de la vraisemblance va s'avérer fort utile [2].

Définition 4 (Log-vraisemblance)

Pour tout $\boldsymbol{\theta} \in \Theta$, la *log-vraisemblance* ℓ est le logarithme de la vraisemblance :

$$\ell(\boldsymbol{\theta}) = \log\left(\mathcal{L}(\boldsymbol{\theta})\right).$$

Théorème 5

Si les réalisations sont indépendantes, la log-vraisemblance est :

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(f(x_i; \boldsymbol{\theta})\right) \tag{2}$$

pour tout $\boldsymbol{\theta} \in \Theta$.

2. [Bickel and Doksum, 1977] page 101, [Greene, 2012] page 550

Définition 6 (Estimateur du maximum de vraisemblance)

La valeur de $\boldsymbol{\theta} \in \Theta$ qui maximise la densité de probabilité du vecteur $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ est l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \, \mathcal{L}(\boldsymbol{\theta}).$$

- La fonction log est croissante, donc le paramètre qui maximise la vraisemblance est également celui qui maximise la log-vraisemblance.
- Donc :

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \, \ell(\boldsymbol{\theta})$$

Estimation de la moyenne d'une Gaussienne

On considère un échantillon $x_1 = 0.6307, x_2 = 0.0623, x_3 = -1.152$ de réalisations indépendantes d'une v.a. Gaussienne d'écart-type $\sigma = 1$ connu.
On a $\bar{x} = -0.1530$.
On souhaite estimer la moyenne μ_{vraie}, inconnue.

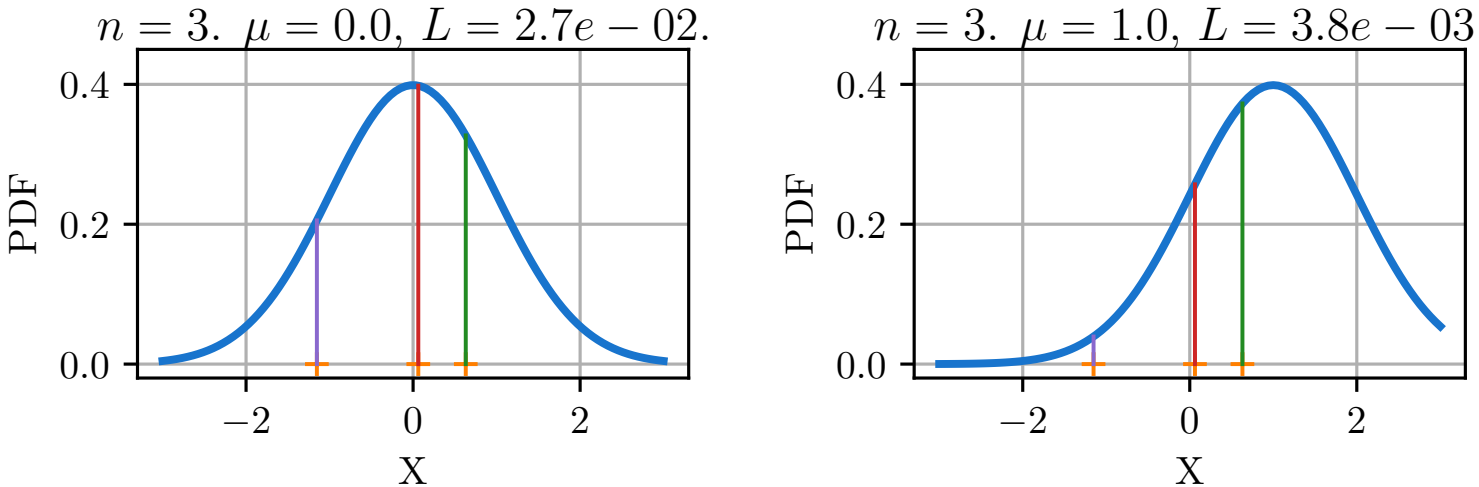

Figure 2 – Probability density function of a Gaussian distribution with $\sigma = 1$ with 3 independent realizations.

La vraisemblance est le produit des barres verticales [3].

3. `Parametric/scripts/normal-likelihood-principle.py`

Estimation de la moyenne d'une Gaussienne

A gauche, $f_{\mu=0}(x_1) = 0.3270$, $f_{\mu=0}(x_2) = 0.3982$, $f_{\mu=0}(x_3) = 0.2055$ et $\mathcal{L}(\mu = 0) = 0.02675$.
A droite, $f_{\mu=1}(x_1) = 0.3726$, $f_{\mu=1}(x_2) = 0.2570$, $f_{\mu=1}(x_3) = 0.0393$ et $\mathcal{L}(\mu = 1) = 0.003772$.
La distribution de gauche est plus vraisemblable.

Estimation de la moyenne d'une Gaussienne

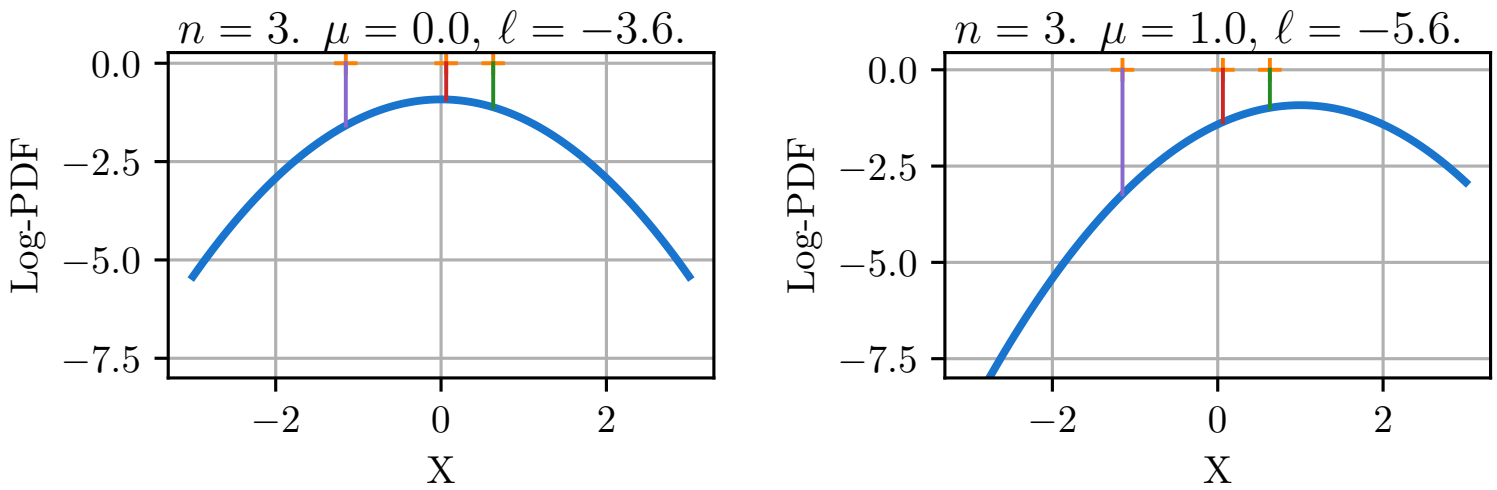

Figure 3 – Logarithm of the probability density function of a Gaussian distribution with $\sigma = 1$.

La log-vraisemblance est la somme des barres verticales.
A gauche, $\log(f_{\mu=0}(x_1)) = -1.118$, $\log(f_{\mu=0}(x_2)) = -0.9209$,
$\log(f_{\mu=0}(x_3)) = -1.5825$ et $\ell(\mu = 0) = \log(\mathcal{L}(\mu = 0)) = -3.621$.
A droite, $\log(f_{\mu=1}(x_1)) = -0.987$, $\log(f_{\mu=1}(x_2)) = -1.359$, $\log(f_{\mu=1}(x_3)) = -3.234$ et $\ell(\mu = 1) = -5.580$.
La distribution de gauche est plus vraisemblable.

Estimation de la moyenne d'une Gaussienne

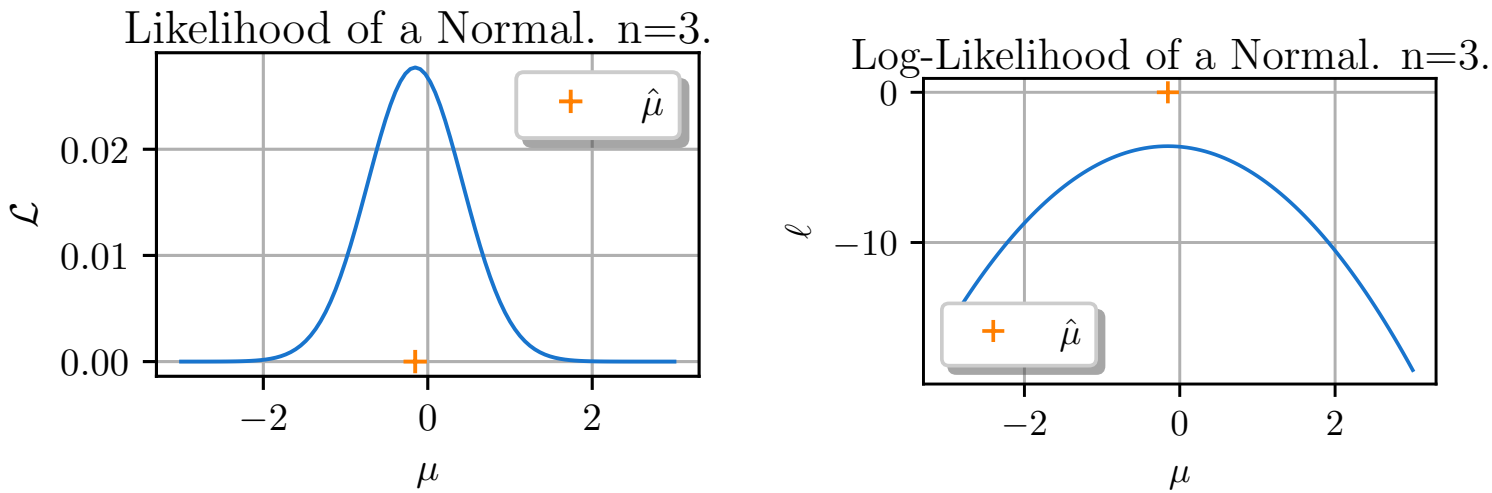

Figure 4 – Likelihood and log-likelihood of a Gaussian sample with 3 realizations.

La solution maximisant la vraisemblance peut ne pas être unique à cause de la présence éventuelle de plusieurs maxima locaux.

Estimation des paramètres d'une loi uniforme par maximum de vraisemblance

La vraisemblance peut être **nulle** si θ est un paramètre de bornes.

Exemple 7

(*Loi uniforme*) Soit X une variable uniforme sur $[a, b]$ où $a < b$.
La densité est :

$$f(x; a, b) = \frac{1}{b - a}$$

si $x \in [a, b]$ et nulle sinon.
Soient $x_1, \ldots, x_n$ des réalisations indépendantes de X.
La vraisemblance est :

$$\mathcal{L}(a, b) = \frac{1}{(b - a)^n} \tag{3}$$

si $x_i \in [a, b]$ pour $i = 1, \ldots, n$ et nulle sinon.
Si $b < \max(x_1, \ldots, x_n)$, alors $\mathcal{L}(a, b) = 0$ car une observation x_i au moins est supérieure à b, pour un certain indice i.
De même si $a > \min(x_1, \ldots, x_n)$, alors $\mathcal{L}(a, b) = 0$ car une observation x_i au moins est inférieure à a, pour un certain indice i.
Conclusion : il est possible d'avoir une vraisemblance nulle.

Estimation des paramètres d'une loi uniforme par maximum de vraisemblance

- Chaque ligne diagonale en haut à gauche représente un ensemble de points où la vraisemblance est constante : $\{(a, b) \mid b - a = c\}$ où c est une constante.
- La valeur maximale est atteinte au point $\hat{a}_{ML} = x_{min} = \min(x_1, \ldots, x_n)$ et $\hat{b}_{ML} = x_{max} = \max(x_1, \ldots, x_n)$.

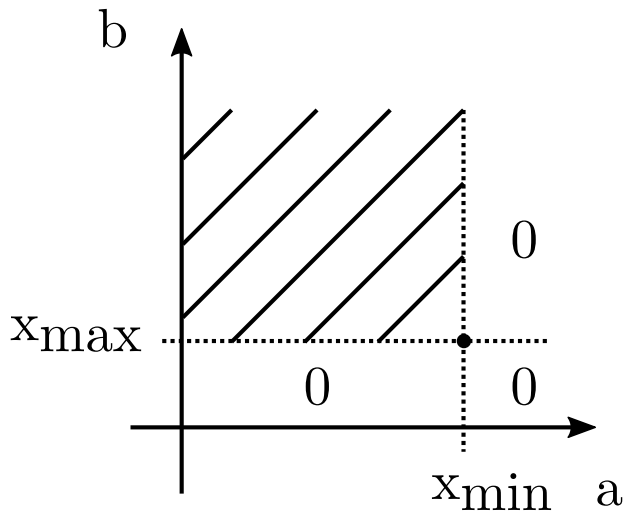

Figure 5 – Vraisemblance d'un échantillon issu de la loi uniforme de paramètres a et b, avec $a \leq b$.

Estimation des paramètres d'une loi uniforme par maximum de vraisemblance

L'équation 3 montre que si a augmente, alors la vraisemblance augmente car :

$$\frac{\partial \mathcal{L}(a,b)}{\partial a} = \frac{n}{(b-a)^{n+1}} > 0$$

pour tout $a < b$.
On maximise la vraisemblance en prenant la valeur la plus grande de a (et telle que la vraisemblance est non nulle), qui est :

$$\hat{a}_{ML} = \min(x_1, \ldots, x_n).$$

Au contraire, si b augmente, alors la vraisemblance diminue car :

$$\frac{\partial \mathcal{L}(a,b)}{\partial b} = -\frac{n}{(b-a)^{n+1}} < 0$$

pour tout $a < b$.
Donc on maximise la vraisemblance en utilisant la plus petite valeur de b (et telle que la vraisemblance est non nulle), qui est :

$$\hat{b}_{ML} = \max(x_1, \ldots, x_n).$$

Supposons que la log-vraisemblance ℓ est de classe $C^2(\mathbb{R}^m)$.
Le maximum est atteint lorsque le gradient est nul :

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

où le gradient de la log-vraisemblance est :

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_m} \end{pmatrix} \in \mathbb{R}^m$$

pour tout $\boldsymbol{\theta} \in \Theta$.
De plus, la solution $\hat{\boldsymbol{\theta}}_{ML}$ est un maximum local si les valeurs propres de la matrice Hessienne sont négatives. Ainsi la fonction ℓ est bien concave au point $\hat{\boldsymbol{\theta}}_{ML}$.
C'est pourquoi on s'intéressera à la matrice Hessienne de la log-vraisemblance [4] :

$$H(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_m \partial \theta_1} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_m \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_m^2} \end{pmatrix} \in \mathbb{R}^{m \times m} \quad (4)$$

~~pour tout $\boldsymbol{\theta} \in \Theta$.~~

4. [Wasserman, 2004] page 133

Exemple : Loi normale avec moyenne et variance inconnues

Exemple 8

(*Loi normale avec moyenne et variance inconnues*) Soit X une variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$ où μ et σ sont inconnus.
On considère un échantillon $\{x_1, \ldots, x_n\}$ de n réalisations indépendantes de X.
Le vecteur des paramètres est $\boldsymbol{\theta} = (\mu, \sigma^2)^T$.
La vraisemblance est :

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Théorème 9 (Estimateur du maximum de vraisemblance des paramètres de la loi Gaussienne)

Les estimateurs par MV sont :

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2$$

On observe que la moyenne et la variance empiriques $\bar{x}, \hat{\sigma}^2$ sont les estimateurs du MV.

Proof. (*)
The log-likelihood is [5] :

$$\begin{aligned}
\ell(\mu, \sigma^2) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right) \\
&= -\sum_{i=1}^{n} \log\left(\sqrt{2\pi\sigma^2}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}
\end{aligned}$$

5. [Greene, 2012] page 553

The partial derivative of the log-likelihood function with respect to μ is :

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2}.$$

We set $\frac{\partial \ell(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu} = 0$, which implies :

$$\sum_{i=1}^{n} \frac{(x_i - \hat{\mu})}{\hat{\sigma}^2} = 0.$$

By hypothesis, we have $\hat{\sigma}^2 > 0$, which implies :

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0.$$

Therefore $\sum_{i=1}^{n} x_i - n\hat{\mu} = 0$ which implies :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The partial derivative of the log-likelihood function with respect to σ^2 (not with respect to σ) is :

$$\begin{aligned}\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2.\end{aligned}$$

We set $\frac{\partial \ell(\hat{\mu}, \hat{\sigma}^2)}{\partial \sigma^2} = 0$, which implies :

$$\frac{n}{2\hat{\sigma}^2} = \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Therefore $n = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$ which implies :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

and concludes the proof. □

Exemple : Loi exponentielle

L'estimateur du maximum de vraisemblance peut être **biaisé**, comme le montre l'estimation par M.V. :

- de la variance de la loi gaussienne ;
- du paramètre de la loi exponentielle.

Consider the exponential distribution [6] :

$$f(x) = \lambda \exp(-\lambda x)$$

for any $x > 0$, where $\lambda > 0$ is the rate parameter.

Théorème 10

L'estimateur du maximum de vraisemblance est [7] *:*

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

6. [Hoel, 1971], p.87
7. [Delmas, 2010], section VIII.6.3 Estimateurs du maximum de vraisemblance, page 224

Exemple : Loi exponentielle

Preuve.

La vraisemblance est

$$\begin{aligned}\mathcal{L}(\lambda) &= \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) \\ &= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \\ &= \lambda^n \exp\left(-\lambda n \overline{x}\right),\end{aligned}$$

La log-vraisemblance est :

$$\log(\mathcal{L}(\lambda)) = n \log(\lambda) - \lambda n \overline{x}.$$

La dérivée de la log-vraisemblance par rapport à λ est :

$$\frac{d}{d\lambda} \log \mathcal{L}(\lambda) = \frac{n}{\lambda} - n\overline{x}$$

Cette dérivée est nulle si et seulement si :

$$\frac{n}{\lambda} - n\overline{x} = 0,$$

ce qui mène à l'estimateur. □

Exemple : Loi exponentielle

L'estimateur du M.V. est convergent [8]. Il est asymptotiquement Gaussien de variance asymptotique λ^2. L'estimateur $\hat{\lambda}$ est un estimateur biaisé de λ, comme le montre le théorème suivant.

Théorème 11

The bias is :

$$b = \mathbb{E}\left[\widehat{\lambda}_{MLE} - \lambda\right] = \frac{\lambda}{n-1}$$

which yields the bias-corrected maximum likelihood estimator :

$$\hat{\lambda}^*_{MLE} = \hat{\lambda}_{MLE} - \hat{b}.$$

8. [Delmas, 2010], page 224

Exemple : Loi exponentielle

Preuve. (*). La loi gamma de paramètres k, λ et γ a pour densité de probabilité :

$$f(x) = \frac{\lambda}{\Gamma(k)} (\lambda(x-\gamma))^{k-1} \exp(-\lambda(x-\gamma))$$

pour tout $x \geq \gamma$ où k est un paramètre de forme, λ est un taux et γ est un paramètre de position. Si $k = 1$, alors

$$f(x) = \lambda \exp(-\lambda(x-\gamma))$$

pour tout $x \geq \gamma$ car $\Gamma(1) = 1$.
Ainsi, si $k = 1$, il s'agit de la densité de probabilité exponentielle de paramètres λ et γ.
Si $X_1, \ldots, X_n$ sont des variables aléatoires de loi gamma de paramètres k, λ et $\gamma = 0$, alors la variable aléatoire $Y = \sum_{i=1}^{n} X_i$ est une variable aléatoire de loi gamma de paramètres nk et λ (et $\gamma = 0$).
En particulier, si $X_1, \ldots, X_n$ sont des variables aléatoires de loi exponentielle, alors elles sont de loi gamma de paramètres $k = 1$, λ et $\gamma = 0$, alors la variable aléatoire $Y = \sum_{i=1}^{n} X_i$ est une variable aléatoire de loi gamma de paramètres n et λ (et $\gamma = 0$).

Exemple : Loi exponentielle

L'espérance de l'inverse est :

$$\begin{aligned}\mathbb{E}\left(Y^{-1}\right) &= \int_0^\infty y^{-1} f_Y(y) dy \\ &= \int_0^\infty y^{-1} \frac{\lambda}{\Gamma(n)} (\lambda y)^{n-1} \exp(-\lambda y) dy \\ &= \int_0^\infty \frac{\lambda}{\Gamma(n)} (\lambda y)^{n-2} \exp(-\lambda y) dy\end{aligned}$$

Or $\Gamma(n) = (n-1)\Gamma(n-1)$, ce qui implique :

$$\mathbb{E}\left(Y^{-1}\right) = \frac{\lambda}{n-1} \int_0^\infty \frac{1}{\Gamma(n-1)} (\lambda y)^{n-2} \exp(-\lambda y) dy$$

Cela fait apparaître l'espérance d'une variable aléatoire de loi gamma de paramètres $n-1$ et λ. Son intégrale est donc égale à 1 :

$$\int_0^\infty \frac{1}{\Gamma(n-1)} (\lambda y)^{n-2} \exp(-\lambda y) dy = 1.$$

Exemple : Loi exponentielle

Par conséquent,

$$\mathbb{E}\left(Y^{-1}\right) = \frac{\lambda}{n-1}.$$

Par conséquent,

$$\begin{aligned}
\mathbb{E}\left(\hat{\lambda}_{MLE}\right) &= \mathbb{E}\left(\frac{1}{\bar{X}}\right) \\
&= \mathbb{E}\left(\frac{n}{\sum_{i=1}^{n} X_i}\right) \\
&= \mathbb{E}\left(\frac{n}{Y}\right) \\
&= n\mathbb{E}\left(Y^{-1}\right) \\
&= n\frac{\lambda}{n-1}.
\end{aligned}$$

Exemple : Loi exponentielle

Par conséquent, le biais est :

$$\begin{aligned} b &= \mathbb{E}(\hat{\lambda}_{MLE} - \lambda) \\ &= n\frac{\lambda}{n-1} - \lambda \\ &= \frac{n\lambda - (n-1)\lambda}{n-1} \\ &= \frac{\lambda}{n-1}, \end{aligned}$$

ce qui conclut la preuve. □

Exemple 12

(*Taille d'un homme*) Let X be the height of a man with age between 20 and 79 years old [9].
This has a Gaussian distribution with parameters [10] $\mu = 1.763$ (m) and $\sigma = 0.0680$ (m).

```
import openturns as ot
import openturns.viewer as otv
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import cm

mu = 1.763
sigma = 0.0680
N = ot.Normal(mu, sigma)
graph = N.drawPDF()
otv.View(graph)
```

9. normal-likelihood.py
10. [U.S. Census Bureau, 2012]

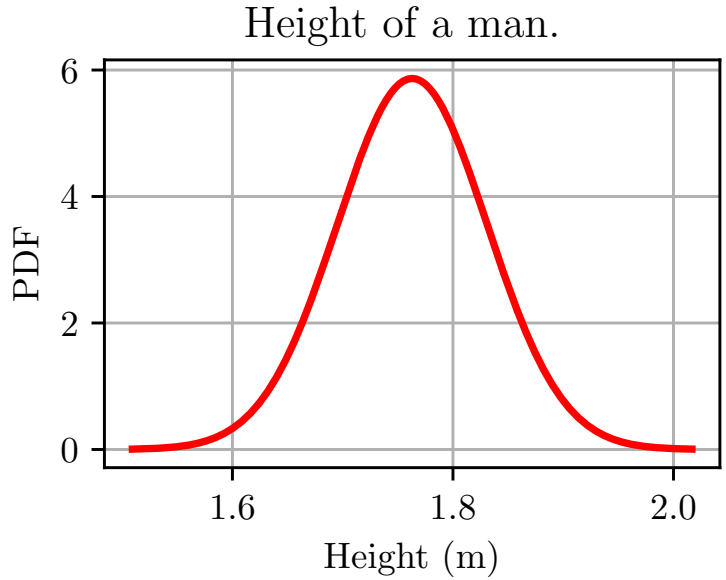

Figure 6 – Probability density function of the height of a man with age between 20 and 79 years old.

- Then we generate a sample with size equal to 20 and compute the likelihood by combining the `computePDF` method and the Numpy `prod` function.
- Finally, we compute the log-likelihood with the `computeLogPDF` method and the Numpy `sum` function.

```
# Generate a sample
sample_size = 20
sample = N.getSample(sample_size)
# Compute likelihood
likelihood = np.prod(N.computePDF(sample))
print("Likelihood=", likelihood)
# Compute log-likelihood
log_pdf = N.computeLogPDF(sample)
log_likelihood = np.sum(log_pdf)
print("Log-Likelihood=", log_likelihood)
```

- The previous script produces the following output.

```
Likelihood= 3.176e11
Log-Likelihood= 24.18
```

- Another way of computing the log-likelihood is to use the `computeMean` method.

```
log_likelihood = log_pdf.computeMean()[0] * \
    sample_size
```

- In order to plot the likelihood, we define the `likelihood_gauss` function which takes into account two input arguments `mu` and `sigma2` and returns the likelihood value.

```
def likelihood_gauss(mu, sigma2):
    X = ot.Normal(mu, np.sqrt(sigma2))
    likelihood = np.prod(X.computePDF(sample))
    return likelihood
```

- We then evaluate the likelihood on a grid produced by the Numpy function `meshgrid`.
- In order to evaluate the likelihood on the grid without any `for` loop, we use the `vectorize` function.

```
n_points = 50
delta_mu = 0.05
mu_min = mu - delta_mu
mu_max = mu + delta_mu
sigma2_min = 0.04 ** 2
sigma2_max = 0.15 ** 2
mu_array = np.linspace(
    mu_min, mu_max, n_points)
sigma2_array = np.linspace(
    sigma2_min, sigma2_max, n_points)
mu_array, sigma2_array = np.meshgrid(
    mu_array, sigma2_array)
likelihood = np.vectorize(likelihood_gauss)(
    mu_array, sigma2_array)
```

- It is then straightforward to use the `plot_surface` function from Matplotlib.

```
fig = plt.figure()
ax = fig.gca(projection="3d")
surf = ax.plot_surface(
    mu_array, sigma2_array, likelihood,
    cmap=cm.coolwarm, linewidth=0,
    antialiased=False)
```

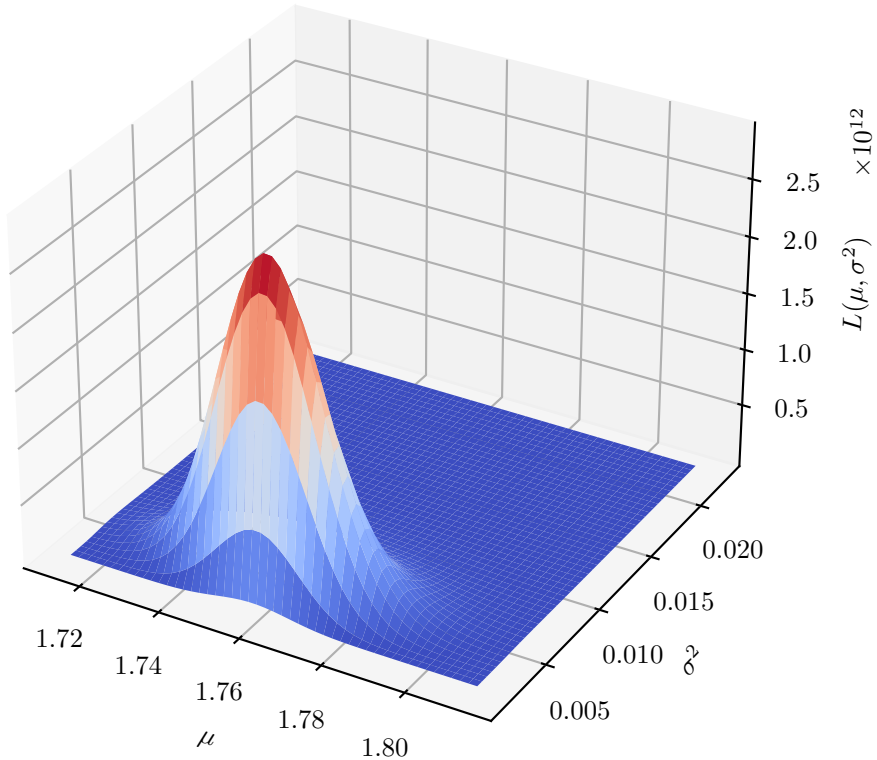

Figure 7 – Likelihood of a sample with size 20 from a Gaussian distribution.

- Pic atteint au point correspondant à la moyenne et à la variance empirique (biaisée).
- On voit que l'ordre de grandeur de la vraisemblance est proche de 10^{10}, ce qui est grand, même lorsque nous utilisons des nombres à virgule flottante.
- Puisque la vraisemblance dépend du nombre d'observations, elle peut être très grande ou très proche de zéro.

It is easy to modify body of the function `likelihood_gauss` in order to evaluate the log-likelihood.

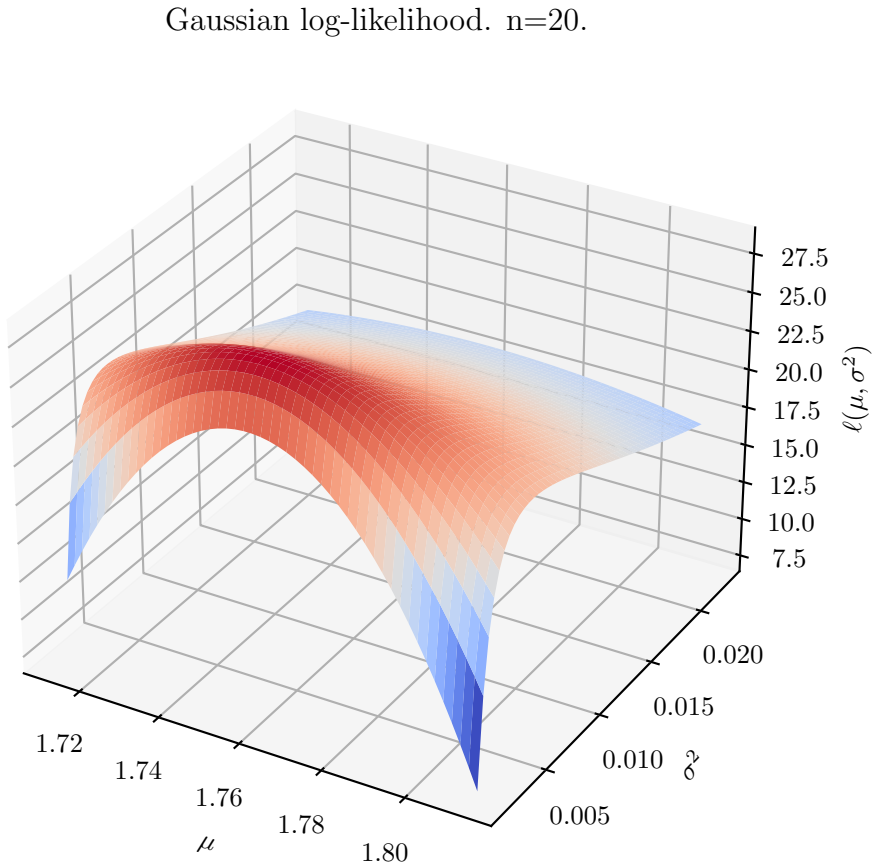

Figure 8 – Log-Likelihood of a sample with size 20 from a Gaussian distribution.

- We see that the log-likelihood is much smaller in magnitude than the likelihood.
- We also see that the maximum of the log-likelihood is achieved at the same parameter value $(\hat{\mu}_{ML}, \hat{\sigma}^2_{ML})$ as the maximum of the likelihood.

It is easier to analyse the likelihood in a contour plot. In order to create it, we define a function which takes a dimension 2 vector (μ, σ^2).

```
def loglikelihood_gauss_point(X):
    mu, sigma2 = X
    N = ot.Normal(mu, np.sqrt(sigma))
    log_pdf = N.computeLogPDF(sample)
    log_likelihood = np.sum(log_pdf)
    return [log_likelihood]
```

Then we create a `PythonFunction` which evaluates the log-likelihood and use the `draw` method. In order to plot 5 contour lines (the default is 10), we use the `"Contour-DefaultLevelsNumber"` key of the `ResourceMap`.

```
logLikelihoodFunction = ot.PythonFunction(
    2, 1, loglikelihood_gauss_point)
ot.ResourceMap_SetAsUnsignedInteger(
    "Contour-DefaultLevelsNumber", 5)
graph = logLikelihoodFunction.draw(
    [mu_min, sigma2_min],
    [mu_max, sigma2_max], [50] * 2)
```

Finally, we add the true (but unknown) mean and variance and the sample mean and variance to the plot.

```
cloud = ot.Cloud([mu], [sigma ** 2])
cloud.setPointStyle("bullet")
graph.add(cloud)
sample_mean = sample.computeMean()[0]
sample_variance = sample.computeVariance()[0]
cloud = ot.Cloud(
    [sample_mean], [sample_variance])
cloud.setPointStyle("plus")
graph.add(cloud)
```

- We see that the maximum likelihood estimators $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$ indeed achieves the maximum of the log-likelihood function.
- The true mean and variance is not very far for the estimate : the estimator converges to the true parameter value when the sample size increases.

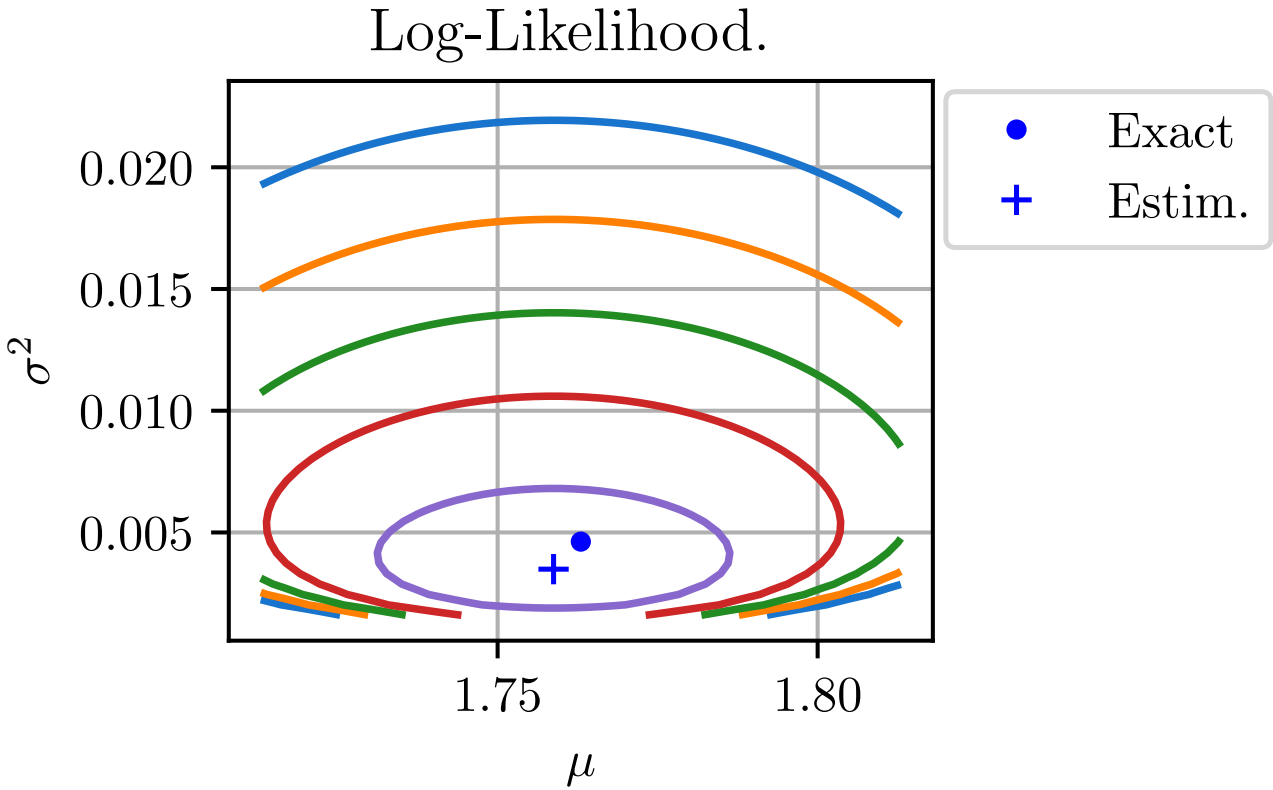

Figure 9 – Contours of the log-Likelihood of a sample with size 20 from a Gaussian distribution.

Score

Le score sera utile par la suite [11].

Définition 13 (Score)

Le score est le gradient de la log-vraisemblance :

$$\mathbf{s}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^m \tag{5}$$

pour tout $\boldsymbol{\theta} \in \Theta$.

Remarque 1

Dans certains textes [12]*, le score est la dérivée de la log-densité :*

$$\frac{\partial \log(f(x, \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}.$$

Il s'agit du score d'une seule observation. Nous noterons plus loin $\mathbf{y}$ *le vecteur associé à la dérivée de la log-densité de l'échantillon.*

11. Voir [Greene, 2012] page 557.
12. [Wasserman, 2004] page 128

Score

Exemple 14

(*Score en dimension 1*) Si $\theta \in \Theta \subset \mathbb{R}$, alors le score est la dérivée de la log-vraisemblance :

$$s(\theta) = \ell'(\theta)$$

pour tout $\theta \in \Theta$. Donc :

$$s(\theta) = \ell'(\theta) = \frac{d}{d\theta} \log(\mathcal{L}(\theta))$$

ce qui implique :

$$s(\theta) = \frac{\mathcal{L}'(\theta)}{\mathcal{L}(\theta)} \tag{6}$$

pour tout $\theta \in \Theta$.

Score

En général :

$$\mathbf{s}(\boldsymbol{\theta}) = \frac{1}{\mathcal{L}(\boldsymbol{\theta})} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{7}$$

pour tout $\boldsymbol{\theta} \in \Theta$ tel que $\mathcal{L}(\boldsymbol{\theta}) > 0$.
The likelihood equation is the necessary condition for maximising the likelihood [13].

Définition 15 (Likelihood equation)

The parameter $\boldsymbol{\theta}_0 \in \Theta$ is the solution of the likelihood equation if :

$$\mathbf{s}(\boldsymbol{\theta}_0) = \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{8}$$

13. [Greene, 2012] page 553

Score

La figure 10 montre que l'équation de vraisemblance est résolue si le score est nul.

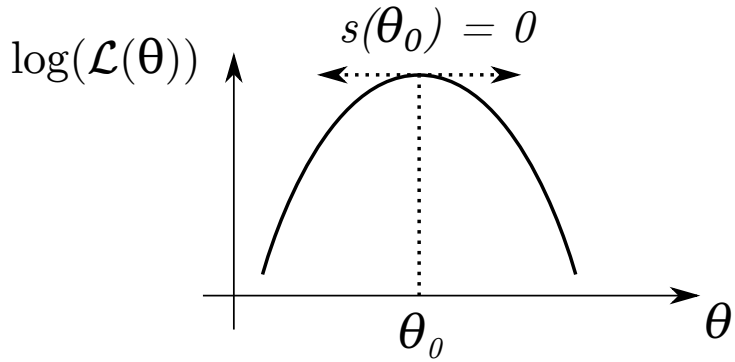

Figure 10 – Résoudre l'équation de vraisemblance consiste à annuler la dérivée de la log-vraisemblance.

Score

L'hypothèse suivante est requise pour obtenir des résultats avec le maximum de vraisemblance [14].

Hypothèse 3

(Regularity of the density - 1) Consider the hypothesis 1.
We assume that the probability density function f is so that :

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathbb{R}} f(x;\boldsymbol{\theta})dx = \int_{\mathbb{R}} \frac{\partial}{\partial \boldsymbol{\theta}} f(x;\boldsymbol{\theta})dx \tag{9}$$

for any $\boldsymbol{\theta} \in \Theta$.

Contre-exemple : paramètres de bornes d'une loi, par exemple les paramètres d'une loi uniforme ou d'une loi tronquée.

14. [Papoulis and Pillai, 2002] page 327, [Bickel and Doksum, 1977] page 126

Score

L'espérance du score est nulle [15].

Théorème 16 (Expectation of the score)

Assume that the hypotheses 2 and 3 are satisfied. Assume that the observations are independent. Let $\mathbf{Y}_i(\boldsymbol{\theta}) \in \mathbb{R}^m$ *be the random vector defined by the equation :*

$$\mathbf{Y}_i(\boldsymbol{\theta}) = \frac{\partial \log [f(X_i, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \tag{10}$$

for $i = 1, \ldots, n$. *Therefore,*

$$\mathbb{E}(\mathbf{Y}_i(\boldsymbol{\theta})) = \mathbf{0}, \tag{11}$$

for $i = 1, \ldots, n$ *and*

$$\mathbb{E}(\mathbf{s}(\boldsymbol{\theta})) = \mathbf{0}, \tag{12}$$

for any $\boldsymbol{\theta} \in \Theta$.

15. [Wasserman, 2004] page 128

Score

Preuve. (*)
By definition of a probability density function, we have :

$$\int_{\mathbb{R}} f(x;\boldsymbol{\theta})dx = 1$$

for any $\boldsymbol{\theta} \in \Theta$. Hence, the equation 9 implies :

$$\int_{\mathbb{R}} \frac{\partial}{\partial \boldsymbol{\theta}} f(x;\boldsymbol{\theta})dx = \mathbf{0} \tag{13}$$

for any $\boldsymbol{\theta} \in \Theta$.

Score

Hence,

$$\begin{aligned}
\mathbb{E}(\mathbf{Y}_i(\boldsymbol{\theta})) &= \int_{\mathbb{R}} \mathbf{y}_i(\boldsymbol{\theta}) f(x;\boldsymbol{\theta}) dx \\
&= \int_{\mathbb{R}} \frac{\partial \log\left[f(x,\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}} f(x;\boldsymbol{\theta}) dx \\
&= \int_{\mathbb{R}} \frac{1}{f(x,\boldsymbol{\theta})} \frac{\partial f(x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(x;\boldsymbol{\theta}) dx \\
&= \int_{\mathbb{R}} \frac{\partial f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dx \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathbb{R}} f(x;\boldsymbol{\theta}) dx
\end{aligned}$$

which implies the equation 11. Here, the last equality comes from the equation 13.

Score

The random variable $\mathbf{Y}_i(\boldsymbol{\theta})$ defined by the equation 10 is associated with the score function since, if the observations are independent, then :

$$\begin{aligned}\mathbf{s}(\boldsymbol{\theta}) &= \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} \log(f(x_i, \boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log(f(x_i, \boldsymbol{\theta}))\end{aligned}$$

which implies :

$$\mathbf{s}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{y}_i(\boldsymbol{\theta}). \tag{14}$$

In other words, the random variable $\mathbf{Y}_i(\boldsymbol{\theta})$ is equal to the score when there is $n = 1$ observation.

Score

Hence,

$$\begin{aligned}\mathbb{E}(\mathbf{s}(\boldsymbol{\theta})) &= \mathbb{E}\left(\sum_{i=1}^{n}\mathbf{Y}_i(\boldsymbol{\theta})\right)\\ &= \sum_{i=1}^{n}\mathbb{E}\left(\mathbf{Y}_i(\boldsymbol{\theta})\right)\\ &= \mathbf{0}\end{aligned}$$

where the last equality comes from the equation 11. □

Matrice de Fisher

Fisher's information matrix measures the amount of information brought by an observation or a sample [16].

Définition 17 (Fisher information matrix)

The Fisher information matrix $\mathcal{I}(\boldsymbol{\theta}) \in \mathbb{R}^{m \times m}$ is :

$$\mathcal{I}(\boldsymbol{\theta})_{jk} = \mathbb{E}\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j}\right)\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_k}\right)\right] \tag{15}$$

for $j, k = 1, \ldots, m$.
Let $\mathcal{I}_1(\boldsymbol{\theta})$ be the Fisher matrix for one single observation :

$$\mathcal{I}_1(\boldsymbol{\theta})_{jk} = \mathbb{E}\left[\left(\frac{\partial \log(f(X;\boldsymbol{\theta}))}{\partial \theta_j}\right)\left(\frac{\partial \log(f(X;\boldsymbol{\theta}))}{\partial \theta_k}\right)\right] \tag{16}$$

for any $j, k = 1, \ldots, m$ and any $\boldsymbol{\theta} \in \Theta$.

16. [Papoulis and Pillai, 2002] page 343, [Wasserman, 2004] page 128, [Bickel and Doksum, 1977] page 127, [Vaart, 2000] page 39

Matrice de Fisher

The equation 15 is sometimes written depending on the partial derivatives of ℓ :

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T\right) \in \mathbb{R}^{m \times m}$$

where $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T$ is an outer product.

Matrice de Fisher

Théorème 18 (Covariance of the score)

Fisher's matrix is the covariance of the score :

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{Cov}(\mathbf{s}(\boldsymbol{\theta})) \tag{17}$$

for any $\boldsymbol{\theta} \in \Theta$.

Démonstration.

Indeed, by definition of the score,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}(\mathbf{s}(\boldsymbol{\theta})\mathbf{s}(\boldsymbol{\theta})^T)$$

where $\mathbf{s}(\boldsymbol{\theta})\mathbf{s}(\boldsymbol{\theta})^T$ is the outer product of the two vectors $\mathbf{s}(\boldsymbol{\theta})$ and $\mathbf{s}(\boldsymbol{\theta})^T$. The properties of the covariance imply :

$$\mathrm{Cov}(\mathbf{s}(\boldsymbol{\theta})) = \mathbb{E}\left[\mathbf{s}(\boldsymbol{\theta})\mathbf{s}(\boldsymbol{\theta})^T\right] - \mathbb{E}\left[\mathbf{s}(\boldsymbol{\theta})\right]\mathbb{E}\left[\mathbf{s}(\boldsymbol{\theta})\right]^T.$$

Since $\mathbb{E}\left[\mathbf{s}(\boldsymbol{\theta})\right] = \mathbf{0}$ by theorem 16 this concludes the proof. □

Matrice de Fisher

Théorème 19 (Hessian of the score and Fisher matrix - part I)

Assume that the hypotheses 2 and 3 are satisfied. Let X be a random variable with probability density function $f(x;\boldsymbol{\theta})$. Assume that the observations $x_1,\dots,x_n$ are independent realizations of X. Let $\mathbf{Y}_i$ be the random vector defined by the equation 11.
Therefore, for any $j,k=1,\dots,m$ and any $\boldsymbol{\theta}\in\Theta$,

$$\mathrm{Cov}\left(\mathbf{Y}_i(\boldsymbol{\theta})\right)_{jk} = -\mathbb{E}\left[\frac{\partial^2 \log(f(X_i,\boldsymbol{\theta}))}{\partial\theta_j\partial\theta_k}\right], \tag{18}$$

Matrice de Fisher

Preuve. (*)
We have :

$$\frac{\partial^2 \log(f(X_i, \boldsymbol{\theta}))}{\partial\theta_j \partial\theta_k} = \frac{\partial}{\partial\theta_j}\left(\frac{\partial \log(f(X_i, \boldsymbol{\theta}))}{\partial\theta_k}\right)$$
$$= \frac{\partial}{\partial\theta_j}\left(\frac{\frac{\partial f(X_i, \boldsymbol{\theta})}{\partial\theta_k}}{f(X_i, \boldsymbol{\theta})}\right)$$

which implies :

$$\frac{\partial^2 \log(f(X_i, \boldsymbol{\theta}))}{\partial\theta_j \partial\theta_k} = \frac{\frac{\partial^2 f(X_i, \boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} f(X_i, \boldsymbol{\theta}) - \frac{\partial f(X_i, \boldsymbol{\theta})}{\partial\theta_j} \frac{\partial f(X_i, \boldsymbol{\theta})}{\partial\theta_k}}{f(X_i, \boldsymbol{\theta})^2}. \tag{19}$$

for any $\boldsymbol{\theta} \in \Theta$, any $i = 1, \ldots, n$ and any $j, k = 1, \ldots, m$.

Matrice de Fisher

Preuve. (suite)
Two successive derivatives of the equation 9 imply :

$$\int_{\mathbb{R}} \frac{\partial^2 f(x,\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} dx = 0 \tag{20}$$

for any $\boldsymbol{\theta} \in \Theta$. Hence, for any $\boldsymbol{\theta} \in \Theta$, any $i = 1, \ldots, n$ and any $j, k = 1, \ldots, m$

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial^2 \log(f(X_i,\boldsymbol{\theta}))}{\partial\theta_j \partial\theta_k}\right) &= \int_{\mathbb{R}} \frac{\frac{\partial^2 f(x,\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} f(x,\boldsymbol{\theta})}{f(x,\boldsymbol{\theta})^2} f(x,\boldsymbol{\theta}) dx \\
&\quad - \int_{\mathbb{R}} \frac{\frac{\partial f(x,\boldsymbol{\theta})}{\partial\theta_j} \frac{\partial f(x,\boldsymbol{\theta})}{\partial\theta_k}}{f(x,\boldsymbol{\theta})^2} f(x,\boldsymbol{\theta}) dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 f(x,\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} dx - \int_{\mathbb{R}} \frac{\frac{\partial f(x,\boldsymbol{\theta})}{\partial\theta_j}}{f(x,\boldsymbol{\theta})} \frac{\frac{\partial f(x,\boldsymbol{\theta})}{\partial\theta_k}}{f(x,\boldsymbol{\theta})} f(x,\boldsymbol{\theta}) dx.
\end{aligned}$$

Matrice de Fisher

Preuve. (suite)
The equation 20 imply that the first integral in the previous equation is zero, which implies :

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial^2 \log(f(X_i,\boldsymbol{\theta}))}{\partial\theta_j\partial\theta_k}\right) &= -\mathbb{E}\left(\frac{\frac{\partial f(X_i,\boldsymbol{\theta})}{\partial\theta_j}}{f(X_i,\boldsymbol{\theta})}\ \frac{\frac{\partial f(X_i,\boldsymbol{\theta})}{\partial\theta_k}}{f(X_i,\boldsymbol{\theta})}\right) \\
&= -\mathbb{E}\left(\frac{\partial \log(f(X_i,\boldsymbol{\theta}))}{\partial\theta_j}\ \frac{\partial \log(f(X_i,\boldsymbol{\theta}))}{\partial\theta_k}\right) \\
&= -\mathbb{E}\left(\mathbf{Y}_i(\boldsymbol{\theta})_j\mathbf{Y}_i(\boldsymbol{\theta})_k\right)
\end{aligned}$$

which implies the equation 18. □

Matrice de Fisher dans le cas indépendant

The Hessian matrix of the log-likelihood indicates the curvature of the log-likelihood function. A high curvature indicates that the maximum is easy to identify because the log-likelihood is sharply curved [17].

Théorème 20 (Hessian of the score and Fisher matrix - part II)

Consider the hypotheses of the theorem 19.
Fisher's matrix is the opposite of the expected value of the Hessian of the log-likelihood :

$$\mathcal{I}(\boldsymbol{\theta})_{jk} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k}\right]. \tag{21}$$

17. [Millar, 2011] page 6

Matrice de Fisher dans le cas indépendant

Remarque 2

The equation 21 implies :

$$\mathcal{I}_1(\boldsymbol{\theta})_{jk} = -\mathbb{E}\left[\frac{\partial^2 \log(f(X_1;\boldsymbol{\theta}))}{\partial\theta_j \partial\theta_k}\right] \tag{22}$$

for any $j, k = 1, \ldots, m$ *and any* $\boldsymbol{\theta} \in \Theta$.

Remarque 3

The equations 16 and 11 imply that the covariance of $\mathbf{Y}_i$ *is the Fisher matrix of one observation :*

$$\mathcal{I}_1(\boldsymbol{\theta})_{jk} = \mathrm{Cov}(\mathbf{Y}_i) \tag{23}$$

for any $j, k = 1, \ldots, m$, $i = 1, \ldots, n$ *and any* $\boldsymbol{\theta} \in \Theta$.

Matrice de Fisher dans le cas indépendant

Preuve. (*)
Let us prove the equation 21. Moreover, for any $j, k = 1, \ldots, m$, the equations 14 and 17 imply :

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\theta})_{jk} &= \mathbb{E}\left[\mathbf{s}(\boldsymbol{\theta})_j \mathbf{s}(\boldsymbol{\theta})_k\right] \\
&= \mathbb{E}\left[\left(\sum_{i_1=1}^{n} \mathbf{Y}_{i_1}(\boldsymbol{\theta})_j\right)\left(\sum_{i_2=1}^{n} \mathbf{Y}_{i_2}(\boldsymbol{\theta})_k\right)\right] \\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \mathbb{E}\left[\mathbf{Y}_{i_1}(\boldsymbol{\theta})_j \mathbf{Y}_{i_2}(\boldsymbol{\theta})_k\right].
\end{aligned}$$

Matrice de Fisher dans le cas indépendant

Preuve. (suite)
By definition of the random vector $\mathbf{Y}_i$, the equation 10 implies :

$$\mathbb{E}\left[\mathbf{Y}_{i_1}(\boldsymbol{\theta})_j \mathbf{Y}_{i_2}(\boldsymbol{\theta})_k\right] = \mathbb{E}\left[\frac{\partial \log(f(X_{i_1}, \boldsymbol{\theta}))}{\partial \theta_j} \frac{\partial \log(f(X_{i_2}, \boldsymbol{\theta}))}{\partial \theta_k}\right] = 0$$

if $i_1 \neq i_2$, because the observations X_{i_1} and X_{i_2} are, by hypothesis, independent. Hence,

$$\mathcal{I}(\boldsymbol{\theta})_{jk} = \sum_{i=1}^{n} \mathbb{E}\left[\mathbf{Y}_i(\boldsymbol{\theta})_j \mathbf{Y}_i(\boldsymbol{\theta})_k\right]$$

which implies :

$$\mathcal{I}(\boldsymbol{\theta})_{jk} = \sum_{i=1}^{n} \mathrm{Cov}\left[\mathbf{Y}_i(\boldsymbol{\theta})\right]_{jk}. \tag{24}$$

Matrice de Fisher dans le cas indépendant

Preuve. (suite) The previous equation implies :

$$\mathcal{I}(\boldsymbol{\theta})_{jk} = -\sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial^2 \log(f(X_i, \boldsymbol{\theta}))}{\partial\theta_j \partial\theta_k}\right]$$

by the equation 18. This implies :

$$\mathcal{I}(\boldsymbol{\theta})_{jk} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} \sum_{i=1}^{n} \log(f(X_i, \boldsymbol{\theta}))\right]$$

which implies the equation 21. □

Matrice de Fisher dans le cas indépendant

La figure 11 montre le lien entre la courbure de la log-vraisemblance et la variance de l'estimateur : lorsque la courbure est forte, l'information de Fisher est grande et la variance de l'estimateur est faible.

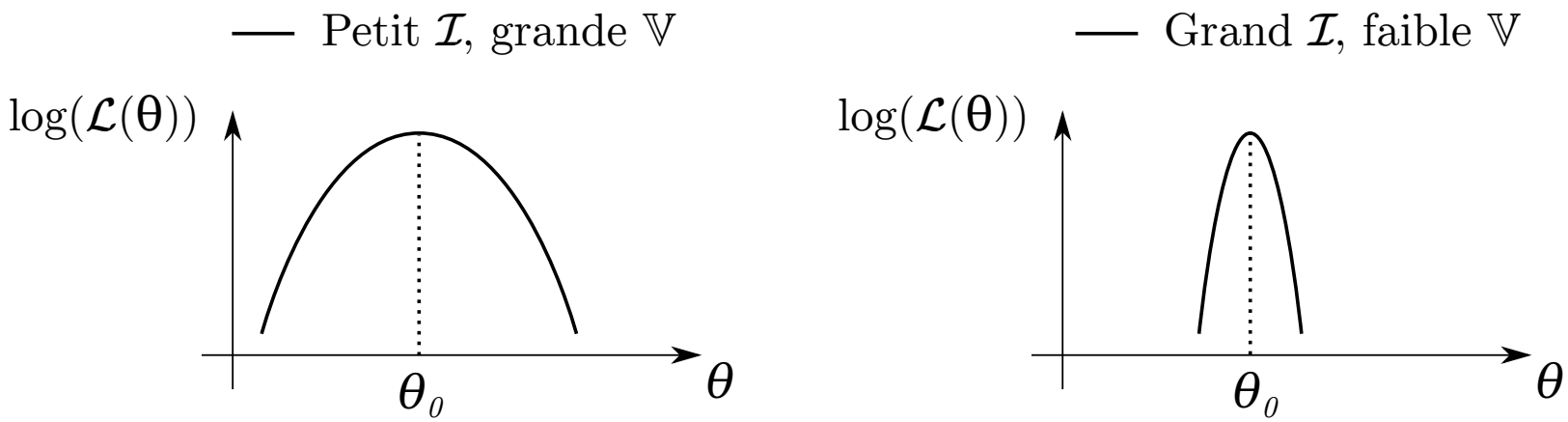

Figure 11 – Lien entre la vraisemblance et la courbure de la log-vraisemblance.

Matrice de Fisher dans le cas indépendant

The following theorem states that the Fisher matrix of a sample is a multiple of the Fisher matrix of one single observation [18].

Théorème 21 (Fisher matrix - part III)

Consider the hypotheses of the theorem 19. The Fisher information matrix of the sample is equal to a multiple of the Fisher matrix of one observation :

$$\mathcal{I}(\boldsymbol{\theta}) = n\mathcal{I}_1(\boldsymbol{\theta}). \tag{25}$$

where $\mathcal{I}_1(\boldsymbol{\theta})$ is the Fisher matrix of a single observation.

Preuve.
The equations 16 and 24 imply :

$$\mathcal{I}(\boldsymbol{\theta})_{jk} = \sum_{i=1}^{n} \mathcal{I}_1(\boldsymbol{\theta})_{jk}$$

which imply the equation 25. □

18. [Bickel and Doksum, 1977] page 129

Matrice de Fisher

Distribution	Fisher matrix
$\mathcal{N}(\mu, \sigma)$	$\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$
$\mathcal{N}(\mu, \sigma^2)$	$\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$
$\mathcal{U}(a, b)$	Undefined
$\mathcal{E}(\lambda)$	$\frac{1}{\lambda}^2$

Table 1 – Fisher matrix $\mathcal{I}_1$ of several distributions.

Voir le théorème 22 pour la loi gaussienne (μ, σ^2), et l'exercice 5 pour la loi gaussienne (μ, σ).
Voir l'exercice 6 pour la loi exponentielle.
Pour la loi Beta, l'information de Fisher dépend de la fonction trigamma [19].

19. [Aryal and Nadarajah, 2004, Nagar et al., 2015]

Matrice de Fisher dans le cas Gaussien

Le théorème suivant présente la matrice de Fisher dans le cas de la loi gaussienne [20].

Théorème 22 (Fisher matrix for a Gaussian sample)

Let X be a Gaussian random variable $\mathcal{N}(\mu, \sigma^2)$ where μ and σ are known. Let $x_1, \ldots, x_n$ be n independent realizations of X.
Therefore, the information matrix is :

$$\mathcal{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}. \tag{26}$$

Remarque 4

The Fisher matrix depends on the parametrization. The theorem 22 presents the the Fisher matrix with parametrization $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. In the exercise 5, we present the Fisher matrix with parametrization $\boldsymbol{\theta} = (\mu, \sigma)^T$ that is, where the second parameter is the standard deviation instead of the variance.

20. [Greene, 2012] page 560, [Wasserman, 2004] page 134

Matrice de Fisher dans le cas Gaussien

Proof.
The density probability function of X is :

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for any $x \in \mathbb{R}$. The log-density is [21] :

$$\log(f(x;\mu,\sigma^2)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

for any $\mu \in \mathbb{R}$ and any $\sigma > 0$. The partial derivative with respect to μ is :

$$\frac{\partial \log(f(x;\mu,\sigma^2))}{\partial \mu} = \frac{x-\mu}{\sigma^2}.$$

21. [Greene, 2012] page 553

Matrice de Fisher dans le cas Gaussien

In order to derive the log-density with respect to σ^2, we write it :

$$\log(f(x;\mu,\sigma^2)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x-\mu)^2}{2}(\sigma^2)^{-1}$$

for any $\mu \in \mathbb{R}$ and any $\sigma > 0$. The partial derivative with respect to σ^2 is :

$$\begin{aligned}\frac{\partial \log(f(x;\mu,\sigma^2))}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2}(\sigma^2)^{-2} \\ &= -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}.\end{aligned}$$

Matrice de Fisher dans le cas Gaussien

The second partial derivative with respect to μ is :

$$\frac{\partial^2 \log(f(x;\mu,\sigma^2))}{\partial \mu^2} = -\frac{1}{\sigma^2}.$$

The partial derivative with respect to μ and σ^2 is :

$$\frac{\partial^2 \log(f(x;\mu,\sigma^2))}{\partial \mu \partial \sigma^2} = -\frac{x-\mu}{\sigma^4}.$$

The second partial derivative with respect to σ^2 is :

$$\frac{\partial^2 \log(f(x;\mu,\sigma^2))}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}.$$

Matrice de Fisher dans le cas Gaussien

We apply the equations 25 and 22. Hence, the first diagonal component is :

$$\begin{aligned}\mathcal{I}(\mu,\sigma^2)_{11} &= -n\mathbb{E}\left[\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial\mu^2}\right] \\ &= n\frac{1}{\sigma^2}.\end{aligned}$$

The off-diagonal component is :

$$\begin{aligned}\mathcal{I}(\mu,\sigma^2)_{12} &= -n\mathbb{E}\left[\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial\mu\partial\sigma^2}\right] \\ &= n\mathbb{E}\left[\frac{X-\mu}{\sigma^4}\right] \\ &= 0,\end{aligned}$$

Matrice de Fisher dans le cas Gaussien

The second diagonal component is :

$$\begin{aligned}
\mathcal{I}(\mu, \sigma^2)_{22} &= -n\mathbb{E}\left[\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial(\sigma^2)^2}\right] \\
&= n\mathbb{E}\left[-\frac{1}{2\sigma^4} + \frac{(X-\mu)^2}{\sigma^6}\right] \\
&= n\left[-\frac{1}{2\sigma^4} + \frac{\mathbb{E}[(X-\mu)^2]}{\sigma^6}\right] \\
&= n\left[-\frac{1}{2\sigma^4} + \frac{\sigma^2}{\sigma^6}\right] \\
&= n\frac{1}{2\sigma^4}.
\end{aligned}$$

Gathering the previous expressions leads to the equation 26. □

Estimation de la matrice de Fisher

Théorème 23 (Estimateur de la matrice de Fisher)

Soit $\boldsymbol{y}_i$ la dérivée de la log-densité d'une observation :

$$\boldsymbol{y}_i = \frac{\partial \log(f(x_i, \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}$$

pour $i = 1, ..., n$. L'estimateur suivant est un estimateur non biaisé de la matrice de Fisher :

$$\hat{\mathcal{I}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^T. \tag{27}$$

Preuve. (*) On a :

$$\boldsymbol{y}_i = \frac{1}{f(x_i, \boldsymbol{\theta})} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

pour $i = 1, ..., n$.
L'équation 16 page 52 indique que la matrice de Fisher d'une seule observation est une espérance. On l'estime par l'équation :

$$\hat{\mathcal{I}}_1(\boldsymbol{\theta})_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \log(f(x_i; \boldsymbol{\theta}))}{\partial \theta_j} \right) \left(\frac{\partial \log(f(x_i; \boldsymbol{\theta}))}{\partial \theta_k} \right).$$

Estimation de la matrice de Fisher

La matrice de Fisher d'une observation peut également être estimée en réalisant une somme matricielle de produits tensoriels :

$$\hat{\mathcal{I}}_1(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^T.$$

On utilise l'équation 25, ce qui mène à l'estimateur de la matrice de Fisher de l'échantillon :

$$\hat{\mathcal{I}}(\boldsymbol{\theta}) = n\hat{\mathcal{I}}_1(\boldsymbol{\theta}).$$

En substituant l'estimateur de la matrice de Fisher d'une observation dans l'expression précédente, on obtient l'équation 27. □

Matrice de Fisher

Exemple 24

(*Compute Fisher matrix of a gaussian*) Consider the random variable [22] $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu = 1$ and $\sigma = 2$. This particular choice of σ avoids confusions between σ and σ^2. We generate a sample of size $n = 1000$. The MLE estimators of μ and σ^2 are :

$$\hat{\mu} = 1.086, \quad \hat{\sigma}^2 = 3.864.$$

The associated estimator of σ is :

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 1.966.$$

(this is not necessarily the MLE for σ). We use the equation 26 to compute Fisher's matrix with parametrization $\boldsymbol{\theta} = (\mu, \sigma^2)$ evaluated at the estimated parameter :

$$\mathcal{I}(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} 258.8 & 0 \\ 0 & 33.48 \end{pmatrix}.$$

22. `gaussian_fisher_matrix.py`

Matrice de Fisher

We use the equation 70 page 145 to compute Fisher's matrix with parametrization $\boldsymbol{\theta}(\mu, \sigma)$ evaluated at the estimated parameter :

$$\mathcal{I}(\hat{\mu}, \hat{\sigma}) = \begin{pmatrix} 258.8 & 0 \\ 0 & 517.5 \end{pmatrix}.$$

We use the equation 27 to estimate Fisher's matrix on this sample and get :

$$\hat{\mathcal{I}}(\hat{\mu}, \hat{\sigma}) = \begin{pmatrix} 259.8 & -5.799 \\ -5.799 & 556.3 \end{pmatrix}.$$

Vraisemblance de la distribution de Bernoulli

Exemple 25

Assume that X is a random variable with Bernoulli distribution with parameter p. We generate $x_1, \ldots, x_n$ independent realizations of X and want to estimate the parameter p from this sample.

On a déjà vu que l'estimateur de la probabilité est

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

La vraisemblance est :

$$\mathcal{L}(p) = p^{n\hat{p}}(1-p)^{n-n\hat{p}} \tag{28}$$

pour tout $p \in [0, 1]$ et la log-vraisemblance est :

$$\ell(p) = n\hat{p}\log(p) + n\left(1-\hat{p}\right)\log(1-p) \tag{29}$$

pour $p \in]0, 1[$.

Vraisemblance de la distribution de Bernoulli

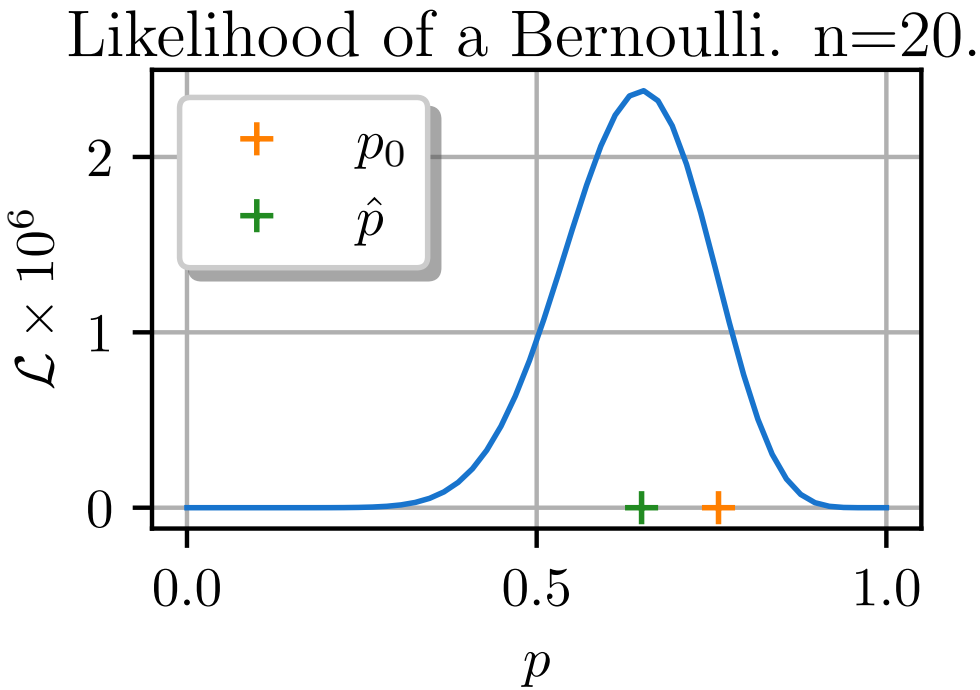

Figure 12 – Likelihood of the Bernoulli distribution with sample size $n = 20$.

On représente $10^6 \times \mathcal{L}(p)$ pour faciliter la lecture.

Vraisemblance de la distribution de Bernoulli

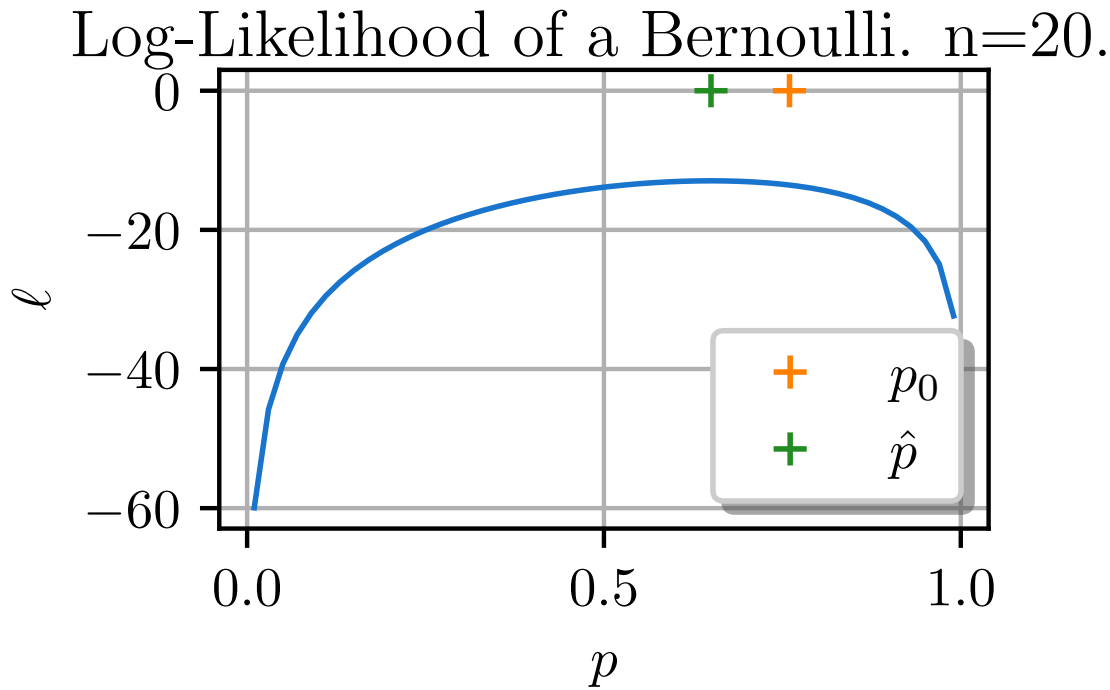

Figure 13 – Log-Likelihood of the Bernoulli distribution with sample size $n = 20$.

Vraisemblance de la distribution de Bernoulli

Différents échantillons mènent à différents $\hat{p}$.

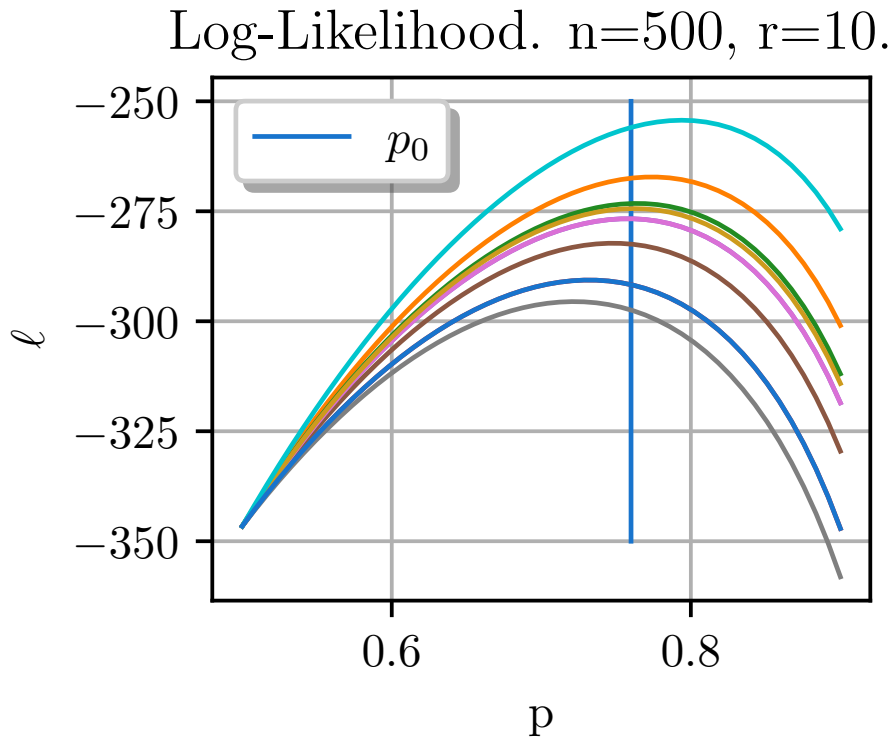

Figure 14 – Log-Likelihood of the Bernoulli distribution with sample size $n = 500$, with $r = 10$ independent samples.

Vraisemblance de la distribution de Bernoulli

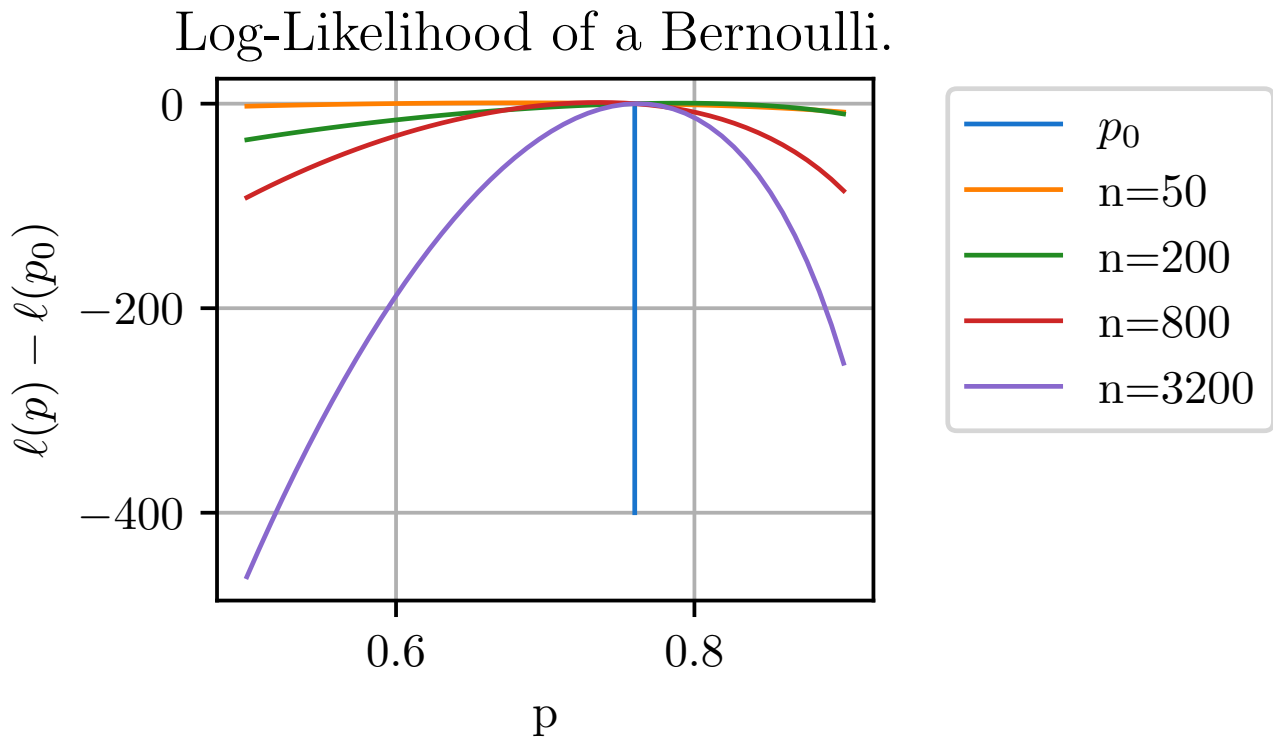

Figure 15 – Log-Likelihood difference of the Bernoulli distribution with sample sizes $n = 50, 200, 800, 1600$.

On présente $\ell(p) - \ell(p_0)$ ce qui décale verticalement la fonction de vraisemblance de la constante $\ell(p_0)$.
Quand la taille de l'échantillon augmente, la fonction devient plus concave : la variance d'estimation dépend de la dérivée seconde de la vraisemblance.

Inégalité de Cauchy-Schwartz pour la covariance (*)

Le coefficient de corrélation est une mesure normalisée de la covariance [23].

Définition 26 (Correlation coefficient)

Let X and Y be random variables. The correlation coefficient is :

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}}. \tag{30}$$

23. [Bickel and Doksum, 1977] page 453, [Ross, 2004] page 126, [Hamilton, 2020] page 745

Inégalité de Cauchy-Schwartz pour la covariance (*)

La valeur absolue du coefficient de corrélation est inférieure ou égale à 1 [24].

Théorème 27 (The probabilistic Cauchy-Schwartz inequality)

Let X and Y be random variables with variances σ_X^2 and σ_Y^2. Therefore,

$$|\rho(X,Y)| \leq 1. \tag{31}$$

If $\rho(X,Y) = \pm 1$, therefore there exists two real constants $a, b \in \mathbb{R}$ such that $Y = aX + b$. Moreover, if $\rho(X,Y) = 1$, therefore :

$$Y = \frac{\sigma_Y}{\sigma_X} X + b \tag{32}$$

and if $\rho(X,Y) = -1$, therefore :

$$Y = -\frac{\sigma_Y}{\sigma_X} X + b. \tag{33}$$

24. [Bickel and Doksum, 1977] page 453, [Ross, 2004] page 138

The Cramér-Rao bound

The Cramér-Rao bound gives the lower bound of an estimator [25].

Théorème 28 (Univariate Cramér-Rao bound of a general estimator)

Let $T(\mathbf{X})$ be an estimator of the real parameter $\theta \in \Theta \subset \mathbb{R}$. Let ψ be the expectation of $T(\mathbf{X})$:

$$\psi(\theta) = \mathbb{E}(T(\mathbf{X})) \tag{34}$$

for any $\theta \in \Theta$. Assume that the following regularity condition holds :

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} T(\mathbf{x}) \mathcal{L}(\boldsymbol{\theta}) d\mathbf{x} = \int_{\mathbb{R}^n} T(\mathbf{x}) \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta} d\mathbf{x} \tag{35}$$

for any $\theta \in \Theta$. If ψ is differentiable, therefore,

$$\mathbb{V}(T(\mathbf{X})) \geq \frac{\left(\frac{\partial \psi(\theta)}{\partial \theta}\right)^2}{\mathcal{I}(\theta)} \tag{36}$$

for any $\theta \in \Theta$.

25. [Bickel and Doksum, 1977] page 127, [Papoulis and Pillai, 2002] page 327

The Cramér-Rao bound

Démonstration (partie 1) (*)
In order to prove 36, we prove the equivalent inequality :

$$\left(\frac{\partial \psi(\theta)}{\partial \theta}\right)^2 \leq \mathbb{V}(T(\mathbf{X}))\mathcal{I}(\theta) \tag{37}$$

for any $\theta \in \Theta$. The equation 34 implies :

$$\frac{\partial \psi(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\mathbb{E}(T(\mathbf{X})) \tag{38}$$

for any $\theta \in \Theta$. The properties of the covariance imply :

$$\mathrm{Cov}(s(\theta), T(\mathbf{X})) = \mathbb{E}(s(\theta)T(\mathbf{X})) - \mathbb{E}(s(\theta))\mathbb{E}(T(\mathbf{X}))$$

for any $\theta \in \Theta$.

The Cramér-Rao bound

Démonstration (partie 2)
The theorem 16 states that the expectation of the score is zero, which implies :

$$\mathrm{Cov}(s(\theta), T(\mathbf{X})) = \mathbb{E}(s(\theta)T(\mathbf{X}))$$

for any $\theta \in \Theta$. Expanding the expectation in the right hand side, we get :

$$\begin{aligned}
&\mathrm{Cov}(s(\theta), T(\mathbf{X})) \\
&= \int_{\mathbb{R}^n} T(\mathbf{X}) s(\theta) \mathcal{L}(\boldsymbol{\theta}) d\mathbf{x} \\
&= \int_{\mathbb{R}^n} T(\mathbf{X}) \frac{\partial}{\partial \theta} \ell(\theta) \mathcal{L}(\boldsymbol{\theta}) d\mathbf{x} \text{ (by the eq. 5)} \\
&= \int_{\mathbb{R}^n} T(\mathbf{X}) \frac{\mathcal{L}'(\theta)}{\mathcal{L}(\theta)} \mathcal{L}(\boldsymbol{\theta}) d\mathbf{x} \text{ (by the eq.6)} \\
&= \int_{\mathbb{R}^n} T(\mathbf{X}) \mathcal{L}'(\theta) d\mathbf{x}
\end{aligned}$$

after the simplification of the probability density function.

The Cramér-Rao bound

Démonstration (partie 3)
Hence,

$$\begin{aligned}\mathrm{Cov}(s(\theta), T(\mathbf{X})) &= \int_{\mathbb{R}^n} T(\mathbf{X}) \frac{\partial}{\partial \theta} \mathcal{L}(\theta) d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} T(\mathbf{X}) \mathcal{L}(\theta) d\mathbf{x}\end{aligned}$$

by the eq. 35. Hence,

$$\mathrm{Cov}(s(\theta), T(\mathbf{X})) = \frac{\partial}{\partial \theta} \mathbb{E}(T(\mathbf{X})) = \frac{\partial \psi(\theta)}{\partial \theta}$$

by the equation 38. The probabilistic Cauchy-Schwarz theorem implies :

$$\left(\frac{\partial \psi(\theta)}{\partial \theta} \right)^2 = \mathrm{Cov}(s(\theta), T(\mathbf{X}))^2 \leq \mathbb{V}(s(\theta)) \mathbb{V}(T(\mathbf{X})).$$

The equation 17 implies $\mathbb{V}(s(\theta)) = \mathcal{I}(\theta)$, which leads to the equation 37 and concludes the proof. □

The Cramér-Rao bound

Si l'estimateur est non biaisé, on peut établir un résultat plus simple [26].

Théorème 29 (Univariate Cramér-Rao bound of an unbiased estimator)

Let $T(\mathbf{X})$ be an unbiased estimator of the real parameter $\theta \in \Theta \subset \mathbb{R}$. Assume that the regularity equation 35 holds. Therefore,

$$\mathbb{V}(T(\mathbf{X})) \geq \frac{1}{\mathcal{I}(\theta)} \tag{39}$$

for any $\theta \in \Theta$.

Preuve : exercice.

26. [Bickel and Doksum, 1977] page 129

The Cramér-Rao bound

Théorème 30 (Multivariate Cramér-Rao bound of a general estimator)

Let $\mathbf{T}(\mathbf{X})$ *be an estimator of the parameter* $\boldsymbol{\theta} \in \Theta$. *Let* $\boldsymbol{\psi}$ *be the expectation of* $\mathbf{T}(\mathbf{X})$ *:*

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbb{E}(\mathbf{T}(\mathbf{X})) \tag{40}$$

for any $\boldsymbol{\theta} \in \Theta$. *Assume that the following regularity condition holds :*

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathbb{R}^n} \mathbf{T}(\mathbf{x}) \mathcal{L}(\boldsymbol{\theta}) d\mathbf{x} = \int_{\mathbb{R}^n} T(\mathbf{x}) \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} d\mathbf{x} \tag{41}$$

for any $\boldsymbol{\theta} \in \Theta$. *If* $\boldsymbol{\psi}$ *is differentiable, therefore,*

$$\mathrm{Cov}(\mathbf{T}(\mathbf{X})) \geq \left[\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T [\mathcal{I}(\theta)]^{-1} \left[\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \tag{42}$$

for any $\boldsymbol{\theta} \in \Theta$.

The right side of the equation 42 is the product of three $m - by - m$ matrices.

The Cramér-Rao bound

The equation 42 is a matrix inequality [27]. If A and B are two real square matrices, we say that $A \geq B$ if the matrix $A - B$ is positive definite.
The vector estimator is :

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\psi}(\boldsymbol{\theta})_1 \\ \vdots \\ \boldsymbol{\psi}(\boldsymbol{\theta})_m \end{pmatrix} \in \mathbb{R}^m.$$

The matrix $\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the Jacobian matrix of $\boldsymbol{\psi}$:

$$\left(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)_{ij} = \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})_i}{\partial \theta_j}$$

for $i, j = 1, \ldots, m$. Then

$$\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})_1}{\partial \theta_1} & \cdots & \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})_1}{\partial \theta_m} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})_m}{\partial \theta_1} & \cdots & \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})_m}{\partial \theta_m} \end{pmatrix} \in \mathbb{R}^{m \times m}.$$

27. [Papoulis and Pillai, 2002] page 343

The Cramér-Rao bound

In the special case where the estimator is unbiased, the multivariate Cramér-Rao bound is simpler [28].

Théorème 31 (Multivariate Cramér-Rao bound of an unbiased estimator)

Let $\mathbf{T}(\mathbf{X})$ *be an unbiased estimator of the parameter* $\boldsymbol{\theta} \in \Theta$. *Assume that the following regularity condition 41 holds. Therefore,*

$$\text{Cov}(\mathbf{T}(\mathbf{X})) \geq [\mathcal{I}(\boldsymbol{\theta})]^{-1} \tag{43}$$

for any $\boldsymbol{\theta} \in \Theta$.

An efficient estimator reaches the Cramér-Rao lower bound [29].

Définition 32 (Efficient estimator)

Let $\mathbf{T}(\mathbf{X})$ be an unbiased estimator of the parameter $\boldsymbol{\theta} \in \Theta$. If its variance attains the Cramér-Rao lower bound, that is, if :

$$\text{Cov}(\mathbf{T}(\mathbf{X})) = [\mathcal{I}(\boldsymbol{\theta})]^{-1} \tag{44}$$

for any $\boldsymbol{\theta} \in \Theta$, therefore we say that this estimator is efficient.

28. [Papoulis and Pillai, 2002] page 343
29. [Bickel and Doksum, 1977] page 138, [Greene, 2012] page 554

The following theorem gathers the main properties of the MLE [30].

Théorème 33 (Properties of the MLE)

Assume that the hypotheses 2 and 3 are satisfied. Assume that the density is differentiable and that the Fisher matrix is finite and nonsingular. The maximum likelihood estimator has the following properties.

1. *The estimator converges in probability :*

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}. \tag{45}$$

2. *The estimator is asymptotically Gaussian :*

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}, [\mathcal{I}(\boldsymbol{\theta})]^{-1}\right). \tag{46}$$

Ce théorème a trois conséquences immédiates.

1. The estimator is asymptotically efficient.
2. The estimator is asymptotically unbiased.
3. The estimator of $\mathbf{c}(\boldsymbol{\theta})$ is $\mathbf{c}\left(\hat{\boldsymbol{\theta}}_{ML}\right)$ if $\mathbf{c} \in C^1(\Theta)$.

30. [Greene, 2012] page 554, [Wasserman, 2004] page 126

Principales propriétés

If the realizations $x_1, \ldots, x_n$ are independent, the equation 25 implies :

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}, [n\mathcal{I}_1(\boldsymbol{\theta})]^{-1}\right)$$

The previous equation is often expressed as :

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, [\mathcal{I}_1(\boldsymbol{\theta})]^{-1}\right). \tag{47}$$

Théorème 34 (Asymptotic distribution of the MLE of a Gaussian sample)

Let X be a Gaussian random variable $\mathcal{N}(\mu, \sigma^2)$ where μ and σ are known. Let $x_1, \ldots, x_n$ be n independent realizations of X.
Therefore, the MLE has asymptotically a Gaussian distribution :

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}, \text{Cov}\left(\hat{\boldsymbol{\theta}}_{ML}\right)\right) \tag{48}$$

where

$$\text{Cov}\left(\hat{\boldsymbol{\theta}}_{ML}\right) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}. \tag{49}$$

The diagonal terms of the equation 49 are :

$$\mathbb{V}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}, \quad \mathbb{V}(\hat{\sigma}^2_{ML}) = \frac{2\sigma^4}{n}.$$

Asymptotic distribution

Exemple 35

(*Compute asymptotic distribution of the parameters of a gaussian distribution*) We consider the same data as in the example 24. The inverse of the estimate of Fisher's matrix is :

$$\left[\hat{\mathcal{I}}(\hat{\mu}, \hat{\sigma})\right]^{-1} = \begin{pmatrix} 0.003851 & 4.014 \times 10^{-5} \\ 4.014 \times 10^{-5} & 0.001798 \end{pmatrix}.$$

The estimated asymptotic distribution of $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})^T$ is :

$$(\hat{\mu}, \hat{\sigma})^T \sim \mathcal{N}\left(\boldsymbol{\theta}, \begin{pmatrix} 0.003851 & 4.014 \times 10^{-5} \\ 4.014 \times 10^{-5} & 0.001798 \end{pmatrix}\right).$$

Using this distribution, we compute marginal 95% confidence intervals for the parameters, and get :

$$\mu \in [0.9647, 1.208], \quad \sigma \in [1.883, 2.049], \quad \text{with 95\% confidence.}$$

Using the Fisher matrix for a Gauss distribution, we get :

$$\mu \in [0.9644, 1.208], \quad \sigma \in [1.880, 2.052], \quad \text{with 95\% confidence.}$$

Asymptotic distribution

Using bootstrap, we get :

$$\mu \in [0.9314, 1.239], \quad \sigma \in [1.873, 2.058], \quad \text{with 95\% confidence.}$$

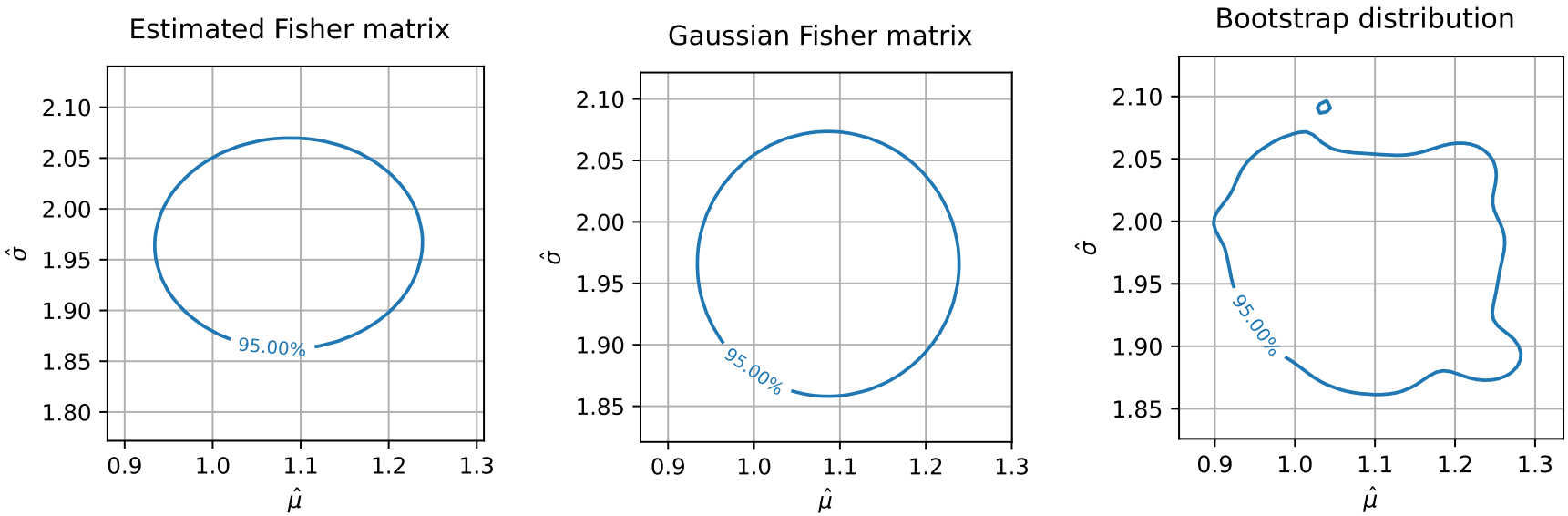

Figure 16 – Ellipsoïde de confiance contenant 95% du vecteur $\boldsymbol{\theta} = (\mu, \sigma)^T$.

The Kullback-Leibler divergence [31] or number [32].

Définition 36 (Kullback-Leibler divergence)

Let p and q be two density probability functions on Ω. The Kullback-Leibler divergence of q to p is :

$$D_{KL}(p,q) = \int_{\Omega} \log\left[\frac{p(x)}{q(x)}\right] p(x)dx \tag{50}$$

The Kullback-Leibler can be expressed has an expectation of the logarithm of a ratio, since :

$$D_{KL}(p,q) = \mathbb{E}_p\left[\log\left(\frac{p(X)}{q(X)}\right)\right], \tag{51}$$

where X has distribution p.

31. [Wasserman, 2004] page 126
32. [Bickel and Doksum, 1977] page 226

Notice that exchanging p and q cannot be done in general, since :

$$D_{KL}(q,p) = \int_{\Omega} \log\left[\frac{q(x)}{p(x)}\right] q(x)dx.$$

Hence, in general, we have that $D_{KL}(p,q)$ is not equal to $D_{KL}(q,p)$. This is why the KL divergence is *not* a distance.
We will use an expression of the divergence in terms of expectation. Indeed,

$$D_{KL}(p,q) = \int_{\Omega} \log\left[\frac{p(x)}{q(x)}\right] \frac{p(x)}{q(x)} q(x)dx$$

which implies :

$$D_{KL}(p,q) = \mathbb{E}_q\left[\frac{p(X)}{q(X)} \log\left[\frac{p(X)}{q(X)}\right]\right] \tag{52}$$

where X has density q.

Gibbs' inequality implies that the Kullback-Leibler divergence is non negative [33].

Théorème 37 (Gibbs' inequality)

Let p and q be two density probability functions on Ω. Therefore,

$$D_{KL}(p, q) \geq 0. \tag{53}$$

Furthermore, we have

$$D_{KL}(p, q) = 0$$

if and only if $p(X) = q(X)$ almost surely.

The fact that $p(X) = q(X)$ almost surely means that :

$$\mathbb{P}(p(X) = q(X)) = 1.$$

33. [Brémaud, 2012] page 68

The following definitions introduces indentifiability of a model [34].

Définition 38 (Identifiability)

Let $\mathcal{P}_{\boldsymbol{\theta}}$ be a statistical model for any $\boldsymbol{\theta} \in \Theta$. We say that the model is identifiable if

$$\mathcal{P}_{\boldsymbol{\theta}_1} \neq \mathcal{P}_{\boldsymbol{\theta}_2} \tag{54}$$

for any $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$.

For example, if we consider the set of probability density function $f(x, \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$:

$$\mathcal{P}_{\boldsymbol{\theta}} = \{f_{\boldsymbol{\theta}} : x \to f(x; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta\}$$

therefore the equation 54 writes :

$$f_{\boldsymbol{\theta}_1} \neq f_{\boldsymbol{\theta}_2} \tag{55}$$

for any $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. The previous condition is necessary if we aim at estimating the parameter $\boldsymbol{\theta}$ using a sample $\mathbf{x}$.

34. [Vaart, 2000] page 62, [Bickel and Doksum, 1977] page 60, [Greene, 2012] page 550

The following theorem introduces the convergence of the MLE [35].

Théorème 39 (Convergence of the MLE (part I))

Let X be a random variable with probability density function $f(\cdot, \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0 \in \Theta$ is the true parameter. Let $X_1, \ldots, X_n$ be n independent random variables with the same distribution as X. For any integer $n \geq 1$, let M_n be defined by the equation :

$$M_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{f(x_i, \boldsymbol{\theta})}{f(x_i, \boldsymbol{\theta}_0)} \right] \tag{56}$$

for any $\boldsymbol{\theta} \in \Theta$. Let M be defined by the equation

$$M(\boldsymbol{\theta}) = -D_{KL}(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}}) \tag{57}$$

for any $\boldsymbol{\theta} \in \Theta$.

i *Maximising the likelihood is equivalent to maximising M_n.*

ii *We have $M_n(\boldsymbol{\theta}) \xrightarrow{\mathbb{P}} M(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Theta$.*

iii *If the parametric model is identifiable, therefore the function M reaches its unique maximum at point $\boldsymbol{\theta}_0$.*

35. [Wasserman, 2004] page 127, [Vaart, 2000] page 62

Démonstration (*)
Let us prove (i). For any $\boldsymbol{\theta} \in \Theta$, the equation 56 implies :

$$\begin{aligned} M_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(x_i,\boldsymbol{\theta})}{f(x_i,\boldsymbol{\theta}_0)}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \log\left[f(x_i,\boldsymbol{\theta})\right] - \frac{1}{n}\sum_{i=1}^{n} \log\left[f(x_i,\boldsymbol{\theta}_0)\right] \\ &= \frac{1}{n}\ell(\boldsymbol{\theta}) - \frac{1}{n}\ell(\boldsymbol{\theta}_0), \end{aligned}$$

where we have used the equation 2. But $\ell(\boldsymbol{\theta}_0)$ is a constant which is independent from $\boldsymbol{\theta}$, so that maximising ℓ amounts to maximising M_n. Subtracting $\frac{1}{n}\ell(\boldsymbol{\theta}_0)$ changes (vertically) the value of the function, but does not change (horizontally) the value of $\boldsymbol{\theta}$ which maximises M_n or ℓ. Notice that we have already considered this method in the example 25, where we plot the log-likelihood of the Bernoulli distribution.

Démonstration
We can now prove (ii). Notice that the equation 56 shows that M_n is defined as the sample mean of a random variable. Since, by hypothesis, the random variables $X_1, \ldots, X_n$ are independent, the weak law of large numbers (theorem 50 page 149) implies :

$$M_n(\boldsymbol{\theta}) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\log\left(\frac{f(X,\boldsymbol{\theta})}{f(X,\boldsymbol{\theta}_0)}\right)\right].$$

Moreover,

$$\begin{aligned}
\mathbb{E}\left[\log\left(\frac{f(X,\boldsymbol{\theta})}{f(X,\boldsymbol{\theta}_0)}\right)\right] &= \mathbb{E}\left[\log\left(f(X,\boldsymbol{\theta})\right) - \log\left(f(X,\boldsymbol{\theta}_0)\right)\right] \\
&= -\left(\mathbb{E}\left[-\log\left(f(X,\boldsymbol{\theta})\right) + \log\left(f(X,\boldsymbol{\theta}_0)\right)\right]\right) \\
&= -\mathbb{E}\left[\log\left(\frac{f(X,\boldsymbol{\theta}_0)}{f(X,\boldsymbol{\theta})}\right)\right] \\
&= -D_{KL}(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}})
\end{aligned}$$

where we have used the equation 51. The previous equation leads to the equation 57.

Démonstration
We can finally prove (iii). We mainly use the theorem 37, i.e. Gibb's inequality. Let $\boldsymbol{\theta} \in \Theta$. The first part of Gibb's inequality implies that $D_{KL}(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}}) \geq 0$, which implies $M(\boldsymbol{\theta}) \leq 0$. Furthermore $D_{KL}(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}}) = 0$ if and only if $f_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}_0}$ almost surely. By hypothesis, the parametric model is identifiable hence the definition 38 of identifiability implies that $f_{\boldsymbol{\theta}} \neq f_{\boldsymbol{\theta}_0}$. This proves that $M(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and concludes the proof.

The following theorem shows that the MLE converges in probability [36].

Théorème 40 (Convergence of the MLE (part II))

We consider the same hypotheses as the theorem 39. Assume that

$$\sup_{\boldsymbol{\theta}\in\Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow{\mathbb{P}} 0. \tag{58}$$

Suppose that, for any $\delta > 0$*, we have*

$$\sup_{\boldsymbol{\theta}\in\Theta \mid |\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\delta} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0). \tag{59}$$

Therefore,

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0. \tag{60}$$

L'équation 58 suppose la convergence uniforme en probabilité de $M_n(\boldsymbol{\theta})$ vers $M(\boldsymbol{\theta})$: c'est une condition plus forte que la convergence en probabilité "ordinaire".

36. [Wasserman, 2004] page 127, [Vaart, 2000] page 45

L'équation 59 impose que le maximum de M doit être *bien séparé*[37].
Cette hypothèse n'est pas satisfaite pour la fonction suivante.

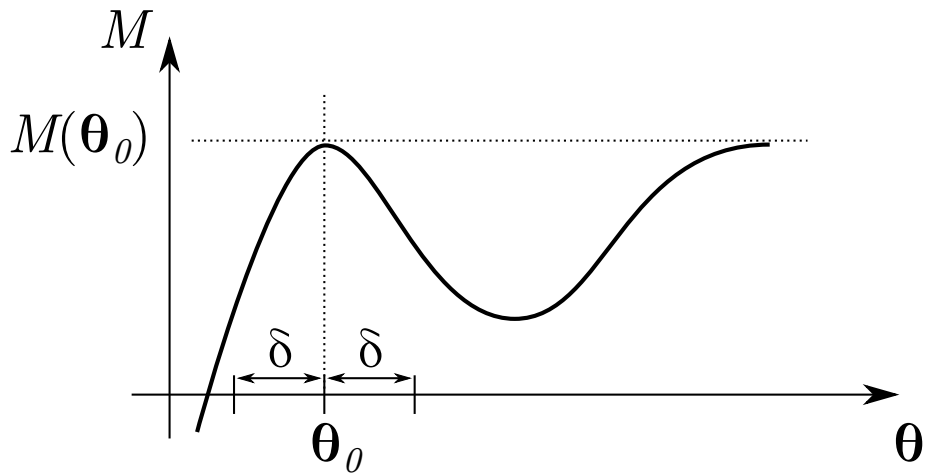

Figure 17 – Une fonction pour laquelle le maximum n'est pas bien séparé.

37. [Vaart, 2000] page 45

Démonstration (*)
The theorem 39 states the vector $\hat{\boldsymbol{\theta}}_{ML}$ maximises M_n, which implies that $M_n(\boldsymbol{\theta}) \leq M_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)$ for any $\boldsymbol{\theta} \in \Theta$. This must be true for $\boldsymbol{\theta}_0$, which implies

$$M_n(\boldsymbol{\theta}_0) \leq M_n\left(\hat{\boldsymbol{\theta}}_{ML}\right). \tag{61}$$

Furthermore, we have

$$M_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \log(1) = 0,$$

which implies $0 \leq M_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)$. We have already seen this property in the example 25.

Démonstration (suite)
In the proof, we denote $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{ML}$.
We have

$$M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) = M(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}_0) + M_n(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right).$$

But $M(\boldsymbol{\theta}_0) = 0$ and $M_n(\boldsymbol{\theta}_0) = 0$, which implies :

$$\begin{aligned} M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) &= M_n(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) \\ &\leq M_n\left(\hat{\boldsymbol{\theta}}\right) - M\left(\hat{\boldsymbol{\theta}}\right) \end{aligned}$$

by the inequality 61. This implies :

$$M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) \leq \sup_{\boldsymbol{\theta}\in\Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})|.$$

Démonstration (suite)

Let $\epsilon > 0$. If $M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) > \epsilon$, therefore

$$\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \geq M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) > \epsilon.$$

This implies :

$$\mathbb{P}\left(M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) > \epsilon\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| > \epsilon\right).$$

The equation 58 implies that the right hand size converges to zero when $n \to +\infty$. Hence,

$$\mathbb{P}\left(M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) > \epsilon\right) \to 0, \tag{62}$$

when $n \to +\infty$.

Démonstration (suite et fin)
Let $\delta > 0$. The inequality 59 implies that there exists $\epsilon > 0$, such that, if $|\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \delta$, then $M(\boldsymbol{\theta}_0) - M(\boldsymbol{\theta}) > \epsilon$. This must be true for $\hat{\boldsymbol{\theta}}$, hence, if $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| > \delta$, then $M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) > \epsilon$. This implies :

$$\mathbb{P}\left(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| > \delta\right) \leq \mathbb{P}\left(M(\boldsymbol{\theta}_0) - M\left(\hat{\boldsymbol{\theta}}\right) > \epsilon\right).$$

The convergence 62 implies

$$\mathbb{P}\left(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| > \delta\right) \to 0$$

when $n \to +\infty$, which concludes the proof.

The following theorem establishes the distribution of an affine transformation of a gaussian random vector [38].

Théorème 41 (Affine transformation of a normal vector)

Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^p$ *be a normal random vector where* $\boldsymbol{\mu} \in \mathbb{R}^p$ *is the mean and* $\Sigma \in \mathbb{R}^{p \times p}$ *is a positive definite matrix. Let*

$$\mathbf{Y} = c + B\mathbf{X}$$

be an affine transformation of $\mathbf{X}$*, where* $\mathbf{c} \in \mathbb{R}^p$ *and* $B \in \mathbb{R}^{p \times p}$*. Therefore, the vector* $\mathbf{Y}$ *is normal and*

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{c} + B\boldsymbol{\mu}, B\Sigma B^T\right). \tag{63}$$

38. [Greene, 2012] page 1083

The following theorem shows that the MLE is asymptotically Gaussian [39].

Théorème 42 (Convergence of the MLE (part III))

We consider the same hypotheses as the theorem 40. Assume that the log-likelihood function $\ell \in C^2(\Theta)$. Therefore,

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}_0, [\mathcal{I}(\boldsymbol{\theta}_0)]^{-1}\right). \tag{64}$$

39. [Wasserman, 2004] page 129, [Greene, 2012] page 559

Démonstration (*)
To simplify the expression, denote $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{ML}$. Since $\ell \in C^2(\Theta)$, we have $\frac{\partial \ell}{\partial \boldsymbol{\theta}} \in C^1(\Theta)$. Hence Taylor's theorem of the function $\frac{\partial \ell}{\partial \boldsymbol{\theta}}$ at point $\boldsymbol{\theta}_0$ implies :

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) = \frac{\partial \ell}{\partial \boldsymbol{\theta}}\left(\boldsymbol{\theta}_0\right) + \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2}\left(\boldsymbol{\xi}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

where $\boldsymbol{\xi}$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Using the definitions 5 and 4 of the score and Hessian matrix, we get :

$$\mathbf{s}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{s}\left(\boldsymbol{\theta}_0\right) + H\left(\boldsymbol{\xi}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

Démonstration (suite)
However, the MLE maximises the likelihood by hypothesis, which implies that the score is zero. In other words, the equation 12 implies :

$$\mathbf{s}\left(\boldsymbol{\theta}_0\right) + H\left(\boldsymbol{\xi}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \mathbf{0}.$$

We rearrange the previous equation and get :

$$\begin{aligned}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= \left[-H\left(\boldsymbol{\xi}\right)\right]^{-1}\mathbf{s}\left(\boldsymbol{\theta}_0\right) \\ &= \left[-H\left(\boldsymbol{\xi}\right)\right]^{-1}\frac{n}{\sqrt{n}}\frac{1}{\sqrt{n}}\mathbf{s}\left(\boldsymbol{\theta}_0\right)\end{aligned}$$

which implies :

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \left[-\frac{1}{n}H\left(\boldsymbol{\xi}\right)\right]^{-1}\frac{1}{\sqrt{n}}\mathbf{s}\left(\boldsymbol{\theta}_0\right). \tag{65}$$

Démonstration (suite)
We now work on the two parts of the right hand side of 65 and exhibit their asymptotic distributions.
Consider first the variable $\mathbf{Y}_i(\boldsymbol{\theta}_0)$ defined by the equation 11. Let $\overline{\mathbf{Y}}(\boldsymbol{\theta}_0)$ be the average :

$$\overline{\mathbf{Y}}\left(\boldsymbol{\theta}_0\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i(\boldsymbol{\theta}_0).$$

The equation 14 implies $\mathbf{s}(\boldsymbol{\theta}_0) = n\overline{\mathbf{Y}}(\boldsymbol{\theta}_0)$. Therefore,

$$\frac{1}{\sqrt{n}}\mathbf{s}(\boldsymbol{\theta}_0) = \sqrt{n}\,\overline{\mathbf{Y}}(\boldsymbol{\theta}_0).$$

The Lindeberg-Levy central limit theorem 51 implies :

$$\sqrt{n}\left[\overline{\mathbf{Y}}\left(\boldsymbol{\theta}_0\right) - \mathbb{E}(\mathbf{Y}_i(\boldsymbol{\theta}_0))\right] \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \mathrm{Cov}(\mathbb{E}(\mathbf{Y}_i(\boldsymbol{\theta}_0)))\right].$$

The theorem 16 implies $\mathbb{E}(\mathbf{Y}_i(\boldsymbol{\theta}_0)) = \mathbf{0}$ and the equation 23 implies $\mathrm{Cov}(\mathbf{Y}_i(\boldsymbol{\theta}_0)) = \mathcal{I}_1(\boldsymbol{\theta}_0)$. Therefore,

$$\sqrt{n}\,\overline{\mathbf{Y}}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_1(\boldsymbol{\theta}_0)\right).$$

This implies :

$$\frac{1}{\sqrt{n}}\mathbf{s}\left(\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_1(\boldsymbol{\theta}_0)\right). \tag{66}$$

Démonstration (suite)
Secondly, consider the random matrix variable $A_i(\boldsymbol{\theta}_0)$ defined by the equation :

$$A_i(\boldsymbol{\theta}_0) = -\frac{\partial^2 \log(f(X_i, \boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}^2} \in \mathbb{R}^{m \times m}.$$

Let $\overline{A}(\boldsymbol{\theta}_0)$ be its average defined by the equation :

$$\overline{A}(\boldsymbol{\theta}_0) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log(f(X_i, \boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}^2}. \tag{67}$$

The weak law of large numbers implies :

$$\overline{A}(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}} \mathbb{E}(A_i(\boldsymbol{\theta}_0)).$$

The equation 18 implies $\mathbb{E}(A_i(\boldsymbol{\theta}_0)) = \mathcal{I}_1(\boldsymbol{\theta}_0)$. Hence,

$$\overline{A}(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}} \mathcal{I}_1(\boldsymbol{\theta}_0).$$

Démonstration (suite)

Moreover, the theorem 40 implies $\hat{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$. Since $\boldsymbol{\xi}$ is between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$, this implies $\boldsymbol{\xi} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$. Finally, the function $\boldsymbol{\theta} \to A_i(\boldsymbol{\theta})$ is continuous since the function $\ell \in C^2(\Theta)$, by hypothesis. We can then apply the continuous mapping theorem 52 to the random variable $\overline{A}$, which implies $\overline{A}(\boldsymbol{\xi}) \xrightarrow{\mathbb{P}} \mathcal{I}_1(\boldsymbol{\theta}_0)$. However, the equation 67 implies :

$$\begin{aligned}\overline{A}(\boldsymbol{\theta}_0) &= -\frac{1}{n}\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\sum_{i=1}^{n}\log(f(X_i,\boldsymbol{\theta}_0)) \\ &= -\frac{1}{n}\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\ell(\boldsymbol{\theta}_0) \\ &= -\frac{1}{n}H(\boldsymbol{\theta}_0).\end{aligned}$$

by definition of H. This implies $-\frac{1}{n}H(\boldsymbol{\xi}) \xrightarrow{\mathbb{P}} \mathcal{I}_1(\boldsymbol{\theta}_0)$. Since the function $B \to B^{-1}$ for any non singular matrix B is a continuous matrix function, the continuous mapping theorem implies :

$$\left[-\frac{1}{n}H(\boldsymbol{\xi})\right]^{-1} \xrightarrow{\mathbb{P}} \mathcal{I}_1(\boldsymbol{\theta}_0)^{-1}. \tag{68}$$

We combine the equations 66 and 68 and Slutsky's theorem for the product to get :

$$\left[-\frac{1}{n}H(\boldsymbol{\xi})\right]^{-1}\frac{1}{\sqrt{n}}\mathbf{s}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{I}_1(\boldsymbol{\theta}_0)^{-1}\mathcal{N}(\mathbf{0},\mathcal{I}_1(\boldsymbol{\theta}_0)).$$

Démonstration (suite et fin)
The theorem 41, page 111, implies that the covariance matrix of the previous normal vector is :

$$\mathcal{I}_1(\boldsymbol{\theta}_0)^{-1}\mathcal{I}_1(\boldsymbol{\theta}_0)(\mathcal{I}_1(\boldsymbol{\theta}_0)^{-1})^T = \mathcal{I}_1(\boldsymbol{\theta}_0)^{-1}.$$

Indeed, we have $(\mathcal{I}_1(\boldsymbol{\theta}_0)^{-1})^T = \mathcal{I}_1(\boldsymbol{\theta}_0)^{-1}$ since Fisher's matrix is symmetric. The equation 65 implies :

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_1(\boldsymbol{\theta}_0)^{-1}\right).$$

We multiply the previous equation by $\frac{1}{\sqrt{n}}$, use the equation 25, and get the equation 64.

Synthèse sur les propriétés asymptotiques du maximum de vraisemblance

- Efficacité asymptotique : le MLE possède la variance asymptotique la plus petite car il atteint la borne de Cramér-Rao ;
- Asymptotiquement sans biais : le MLE converge vers la vraie valeur ;
- Asymptotiquement gaussien : cela permet d'évaluer un intervalle de confiance du paramètre.

Limitations du maximum de vraisemblance.

- Nécessite souvent de recourir à une méthode d'optimisation numérique. Cette méthode nécessite un point de départ. Le problème d'optimisation peut avoir plusieurs minimas locaux. La méthode peut diverger. Elle peut nécessiter la dérivée de la vraisemblance.
- Pour une taille d'échantillon petit, le MLE n'est pas nécessairement optimal.

Alternatives au maximum de vraisemblance.

- La méthode des moments : identification des moments par résolution d'un système d'équations non linéaires.
- Les méthodes bayésiennes : la loi a priori sur le vecteur de paramètres peut régulariser le problème, en particulier lorsque la taille de l'échantillon est petite. Lorsque la taille de l'échantillon tend vers l'infini, la loi a priori ne compte plus.

Method of moments

Consider the random variable X.
Let $\boldsymbol{\theta} \in \mathbb{R}^m$ be a vector of m parameters of the probability distribution function of X.
Let $x_1, ..., x_n$ be n independent observations of X.
Assume that the first m centered moments of the distribution $\mu_1(\boldsymbol{\theta}), \ldots, \mu_k(\boldsymbol{\theta})$ can be expressed [40] as a function of the vector $\boldsymbol{\theta}$:

$$\begin{aligned}
\mu_1(\boldsymbol{\theta}) &= \mathbb{E}[X], \\
\mu_2(\boldsymbol{\theta}) &= \mathbb{E}[(X - \mu_1)^2], \\
&\vdots \\
\mu_m(\boldsymbol{\theta}) &= \mathbb{E}[(X - \mu_1)^m].
\end{aligned}$$

40. [Papoulis and Pillai, 2002], p.147

Let $\hat{\mu}_1, \ldots, \hat{\mu}_m$ be the sample centered moments :

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu}_1)^2,$$

$$\vdots$$

$$\hat{\mu}_m = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu}_1)^m .$$

Définition 43 (Method of moments)

The method of moments is the solution $\hat{\boldsymbol{\theta}}_{MoM} \in \mathbb{R}^m$ of the non linear equations [41] :

$$\mu_1(\hat{\boldsymbol{\theta}}_{MoM}) = \hat{\mu}_1, \quad \mu_2(\hat{\boldsymbol{\theta}}_{MoM}) = \hat{\mu}_2, \quad \dots \quad \mu_m(\hat{\boldsymbol{\theta}}_{MoM}) = \hat{\mu}_m.$$

In OpenTURNS, we solve the optimization problem :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left[\frac{|\mu_1(\boldsymbol{\theta}) - \hat{\mu}_1|}{\sigma(\boldsymbol{\theta})}\right]^2 + \sum_{k=2}^{m} \left[\frac{|\mu_i(\boldsymbol{\theta})|^{\frac{1}{k}} - |\hat{\mu}_k|^{\frac{1}{k}}}{\sigma(\boldsymbol{\theta})}\right]^2.$$

where $\sigma(\boldsymbol{\theta})$ is the standard deviation of the parametric distribution.

41. [Wasserman, 2004], p.121

Exemple 44 (MoM for a Gaussian distribution)

Consider the Gaussian distribution [42] $\mathcal{N}(\mu, \sigma^2)$. We can prove that :

$$\bar{x} \xrightarrow{\mathbb{P}} \mathbb{E}[X] = \mu$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \xrightarrow{\mathbb{P}} \mathbb{V}[X] = \sigma^2.$$

This is why $\bar{x}$ and $\hat{\sigma}^2$ are the moment estimators of μ and σ^2.

42. [Greene, 2012], p.497

MoM for a gamma distribution

Exemple 45 (MoM for a gamma distribution)

Consider the gamma distribution [43] :

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{\alpha}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}$$

for any $x > 0$, where $\alpha > 0$ and $\beta > 0$ are parameters. The gamma function extends the factorial function to real numbers :

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

for any $\alpha > 0$. The mean and variance are :

$$\mathbb{E}[X] = \mu = \alpha\beta, \quad \mathbb{V}[X] = \sigma^2 = \alpha\beta^2.$$

43. [Hoel, 1971], p.86

MoM for a gamma distribution

Hence the method of moment can be based on the equations [44] :

$$\bar{x} = \hat{\alpha}\hat{\beta}, \quad \hat{\sigma}^2 = \hat{\alpha}\hat{\beta}^2.$$

This leads to :

$$\hat{\alpha}_{MoM} = \frac{\bar{x}^2}{\hat{\sigma}^2}, \quad \hat{\beta}_{MoM} = \frac{\hat{\sigma}^2}{\bar{x}}.$$

44. [Hoel, 1971], p.106

MoM for a exponential distribution

Exemple 46 (MoM for a Exponential distribution)

Consider the exponential distribution [45] :

$$f(x) = \lambda \exp\left(-\lambda\left(x-\gamma\right)\right)$$

for any $x > \gamma$, where $\lambda > 0$ is the rate parameter and $\gamma > 0$ is the location parameter. Its first moments are :

$$\begin{aligned} \mathbb{E}[X] &= \gamma + \frac{1}{\lambda} \\ \mathrm{Var}[X] &= \frac{1}{\lambda^2} \end{aligned}$$

45. [Hoel, 1971], p.87

MoM for a exponential distribution

Dans la méthode des moments, on estime la moyenne empirique $\bar{x}$ et la variance empirique $\hat{\sigma}^2$. Puis on cherche $\hat{\lambda}$ et $\hat{\gamma}$ solutions du système d'équations :

$$\begin{aligned} \bar{x} &= \hat{\gamma} + \frac{1}{\hat{\lambda}} \\ \hat{\sigma}^2 &= \frac{1}{\hat{\lambda}^2} \end{aligned}$$

Théorème 47

La méthode des moments est :

$$\hat{\lambda}_{MoM} = \frac{1}{\hat{\sigma}}, \qquad \hat{\gamma}_{MoM} = \bar{x} - \frac{1}{\hat{\lambda}} = \bar{x} - \hat{\sigma}.$$

Preuve La seconde équation du système implique :

$$\hat{\lambda}^2 = \frac{1}{\hat{\sigma}^2}$$

ce qui mène à l'équation pour $\hat{\lambda}$. La première équation du système mène à l'estimateur pour $\hat{\gamma}$.

Distribution and asymptotic distribution of MoM

In general, the distribution of the MoM estimator is unknown.
There are particular cases where it is known.

Exemple 48 (Exact distribution of MoM for parameters of normal distribution)

Assume that X has the normal distribution with parameters μ and σ^2.
The sample mean has the distribution [46] :

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

The sample variance has the distribution :

$$\hat{\sigma}^2 \sim \mathcal{N}\left(\frac{n-1}{n}\sigma^2, \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1}\right).$$

46. See [Greene, 2012] p.501.

Distribution and asymptotic distribution of MoM

If X does not have the normal distribution, the distribution of the sample mean or sample variance is not known in general, but the asymptotic variance is known.

Exemple 49 (Asymptotic distribution of sample mean and sample variance)

Assume that X is a random variable with finite mean $\mu = \mathbb{E}(X)$ and finite variance $\sigma^2 = \mathbb{E}[(X-\mu)^2]$.
The asymptotic variance of the sample mean is given by the CLT :

$$\text{Asy. Var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

when $n \to \infty$. The asymptotic variance of the sample variance is [47] :

$$\text{Asy. Var}(\hat{\mu}) = \frac{\mu_4 - \sigma^2}{n}$$

when $n \to \infty$ where $\mu_4 = \mathbb{E}[(X-\mu)^4]$ is the fourth centered moment.

47. See [Greene, 2012] p.1128.

Distribution and asymptotic distribution of MoM

The asymptotic distribution of the MoM can be derivated if the moments are computed using the sample mean of some vector [48].
Let $\{x_1, ..., x_n\}$ be the independent observations of X.
Let $\lambda_1, ..., \lambda_m$ be m continuous and differentiable functions.
Assume that the moments are :

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} \lambda_k(x_i)$$

for $k = 1, ..., m$.
An estimator of the asymptotic covariance matrix of the random vector $(\hat{\mu}_1, ..., \hat{\mu}_m)$ is :

$$\hat{F}_{jk} = \frac{1}{n} \sum_{i=1}^{n} [\lambda_j(x_i) - \hat{\mu}_j] [\lambda_k(x_i) - \hat{\mu}_k]$$

for $j, k = 1, ..., m$.

48. See [Greene, 2012] p.501.

Distribution and asymptotic distribution of MoM

Let $J \in \mathbb{R}^{m \times m}$ be the Jacobian matrix of partial derivatives of $\hat{\mu}_k$:

$$J_{k,j}(\boldsymbol{\theta}) = \frac{\partial \hat{\mu}_k(\boldsymbol{\theta})}{\partial \theta_j}$$

for $k, j = 1, ..., m$. The first order Taylor expansion of the vector $\hat{\boldsymbol{\mu}}$ at point $\boldsymbol{\theta}_0$ implies :

$$\mathbf{0} \approx \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_0) + J(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

Therefore [49] :

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \approx -[J(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_0)$$

for $k = 1, ..., m$. The multivariate delta method implies that the covariance matrix of the method of moments can be estimated with :

$$\text{Est. Asy. Var}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{n}\left[J\left(\hat{\boldsymbol{\theta}}\right)\hat{F}^{-1}J\left(\hat{\boldsymbol{\theta}}\right)^T\right]^{-1}.$$

49. See [Greene, 2012] eq. 13-1 p.502.

- The MoM is sometimes tractable analytically.
- The MoM is sometimes asymptotically Gaussian, but the exact distribution may be intractable [50].
- The analytical solution of the MoM is often a good starting point for the MLE.

The method of generalized moments combines several moments to improve the estimation [51].

50. See [Greene, 2012] p.501.
51. [Hamilton, 2020], p.410

Maximum likelihood estimate of a probability

Exercice 1

(*Maximum likelihood estimate of a probability*) Assume that the random variable X has a Bernoulli distribution with parameter $p \in [0, 1]$. Hence, a realization x of the random variable X is defined by the equation :

$$x = \begin{cases} 1 \text{ with probability } p \text{ if trial is a success,} \\ 0 \text{ with probability } 1 - p \text{ otherwise.} \end{cases}.$$

The probability mass function of the Bernoulli distribution is :

$$f(x; p) = p^x (1 - p)^{1-x}, \tag{69}$$

for any $x \in \{0, 1\}$. This simply means that $f(1; p) = p$ and $f(0; p) = 1 - p$. Assume that $x_1, \ldots, x_n \in \{0, 1\}$ are n independent realizations of X, where n is the sample size. What is the maximum likelihood estimate of p ?

Application du maximum de vraisemblance à la taille d'un homme

Exercice 2

(*Python / OpenTURNS : Application du maximum de vraisemblance à la taille d'un homme*)
On cherche à reproduire la figure 2. Pour cela, on utilise les données de l'exemple de la taille d'un homme 12.
Question 1.

- Créer une variable aléatoire `Normal` de paramètres $\mu = 1.763$ (m) et $\sigma = 0.0680$ (m).
- Générer un échantillon de taille $n = 3$ réalisations indépendantes de la variable aléatoire précédente, notées x_1, x_2, x_3.
- Dessiner la densité de probabilité de la variable aléatoire.
- Dans le même graphique, dessiner les observations les observations x_1, x_2, x_3 sur l'axe des abscisses. Pour cela, créer un nuage de points de type `ot.Cloud`, puis utiliser la méthode `add` de l'objet `Graph` pour ajouter le nuage dans le graphique précédent.
- Calculer la vraisemblance de l'échantillon et configurer le titre pour afficher la valeur numérique.

Question 2. Faire de même pour la log-vraisemblance.

Etude de la probabilité d'avoir un infartus du myocarde

Exercice 3

(*Maximum likelihood estimate of a probability*) The Physician's Health Study aims at finding if low dose aspirin reduces the probability of cardiovascular mortality and whether beta caroten can help to prevent cancer
[Steering Committee of the Physicians' Health Study Research Group, 1989]. It is a study on 22071 participants, randomly split into two groups. Moreover, it is double-blind, meaning that neither the participants nor the experimenters know who is taking the treatment.
A summary of the results is presented in the table 2. We want to estimate the probability of a myocardial infarction with placebo. The goal of this exercise is to plot the likelihood and log-likelihood for this experiment.

	Myocardial infarctions	**No myocardial infarctions**	**Total**
Placebo	239	10795	11034
Aspirin	139	10898	11037

Table 2 – Results from the Physician's Health Study.

Etude de la probabilité d'avoir un infartus du myocarde

Exercice (suite)
Let X be the event of having a myocardial infarction with placebo "treatment". Assume that the random variable X has a Bernoulli distribution with unknown parameter $p \in [0, 1]$. Hence, an observation x of the random variable X is defined by the equation :

$$x = \begin{cases} 1 \text{ with probability } p \text{ if trial is a success,} \\ 0 \text{ with probability } 1 - p \text{ otherwise.} \end{cases} .$$

Here, "success" means that the participant has a myocardial infarction.

Etude de la probabilité d'avoir un infartus du myocarde

Exercice (suite)
Assume that $x_1, \ldots, x_n \in \{0, 1\}$ are n independent realizations of X, where $n = 11034$ is the sample size.
We want to plot the likelihood function depending on p.
Partie 1

- Define a `sample` with size n containing 239 ones and 10795 zeros. To do this, we suggest the following code.

```
import openturns as ot
import numpy as np

sample_size = 11034
placebo_infarctions = 239
sample = ot.Sample(sample_size, 1)
sample[0:placebo_infarctions, :] = \
    np.ones((placebo_infarctions, 1))
```

Etude de la probabilité d'avoir un infartus du myocarde

Exercice (suite)

- Define the function `likelihood_bernoulli(input_parameter)` which evaluates the likelihood function. A template for the function is :

```
def likelihood_bernoulli(input_parameter):
    p = input_parameter[0]
    TODO
    return [likelihood]
```

where the `TODO` must be replaced with valid code.
In order to do this, use the equation 69 in order to derive the likelihood function of p.

Etude de la probabilité d'avoir un infartus du myocarde

Exercice (suite et fin)
Partie 2

- Plot the likelihood function for $p \in [0, 0.1]$. To do this, create the function `likelihood_bernoulli_Py` using the `PythonFunction` class, and use the `draw` method. Plot the estimate $\hat{p}_{ML}$ on the graph. What is interesting in this plot ? Can you explain its content ?
- Plot the log-likelihood function for $p \in [0.01, 0.09]$. To do this, define the function `loglikelihood_bernoulli(theta)` which evaluates the log-likelihood function. A template for the function is :

```
def loglikelihood_bernoulli(input_parameter):
    p = input_parameter[0]
    TODO
    return [likelihood]
```

  Plot the estimate $\hat{p}_{ML}$ on the graph. What is interesting in this plot ? Can you explain its content ?

Fin de l'exercice 3.

Derivatives of the likelihood

Exercice 4

(*Derivative of the likelihood*) Assume that the hypotheses of the theorem 5 page 8 are true. Compute the partial derivatives of the log-likelihood :

$$\mathbf{s}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_m} \end{pmatrix}$$

for any $\boldsymbol{\theta} \in \Theta$ depending on the observations $\{x_1, ..., x_n\}$ and the partial derivatives of the log-density.

Derivatives of the likelihood

Solution de l'exercice 4

The theorem 5 implies :

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} \log\left(f(x_i; \boldsymbol{\theta})\right)$$
$$= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(f(x_i; \boldsymbol{\theta})\right).$$

Comme dans le théorème 16, introduisons l'échantillon $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ les dérivées partielles de la log-densité :

$$\mathbf{y}_i = \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(f(x_i; \boldsymbol{\theta})\right) \in \mathbb{R}^m$$

pour $i = 1, ..., n$. Therefore,

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{y}_i.$$

Fisher matrix of a gaussian

Exercice 5

(*Fisher matrix of a gaussian*) In the theorem 22, we presented the Fisher information matrix of the gaussian distribution with the $\boldsymbol{\sigma} = (\mu, \sigma^2)^T$ parametrization. Another common parametrization of the gaussian distribution is using the standard deviation, so that the parameter is $\boldsymbol{\sigma} = (\mu, \sigma)^T$. Compute the Fisher matrix corresponding to $\boldsymbol{\sigma} = (\mu, \sigma)^T$.

Fisher matrix of a gaussian

Solution de l'exercice 5

The density probability function of X is :

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for any $x \in \mathbb{R}$. Notice that the density depends on σ, and not on σ^2 as in theorem 22. The log-density is [52] :

$$\log(f(x;\mu,\sigma)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

for any $\mu \in \mathbb{R}$ and any $\sigma > 0$. The partial derivative with respect to μ is :

$$\frac{\partial \log(f(x;\mu,\sigma))}{\partial \mu} = \frac{x-\mu}{\sigma^2}.$$

The partial derivative with respect to σ is :

$$\frac{\partial \log(f(x;\mu,\sigma))}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}.$$

The second partial derivative with respect to μ is :

$$\frac{\partial^2 \log(f(x;\mu,\sigma))}{\partial \mu^2} = -\frac{1}{\sigma^2}.$$

52. [Greene, 2012] page 553

Fisher matrix of a gaussian

The second partial derivative with respect to μ and σ is :

$$\frac{\partial^2 \log(f(x;\mu,\sigma))}{\partial\mu\partial\sigma} = -2\frac{x-\mu}{\sigma^3}.$$

The second partial derivative with respect to σ is :

$$\frac{\partial^2 \log(f(x;\mu,\sigma))}{\partial\mu^2} = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}.$$

The first entry of the Fisher matrix is :

$$\begin{aligned}\mathcal{I}(\mu,\sigma)_{11} &= -n\mathbb{E}\left[\frac{\partial^2 \log(f(x;\mu,\sigma))}{\partial\mu^2}\right] \\ &= -n\mathbb{E}\left[-\frac{1}{\sigma^2}\right] \\ &= \frac{n}{\sigma^2}.\end{aligned}$$

Fisher matrix of a gaussian

The off-diagonal entry of the Fisher matrix is :

$$\begin{aligned}\mathcal{I}(\mu,\sigma)_{12} &= -n\mathbb{E}\left[-2\frac{x-\mu}{\sigma^3}\right] \\ &= \frac{2n}{\sigma^3}\mathbb{E}\left[x-\mu\right] \\ &= 0\end{aligned}$$

since $\mathbb{E}\left[x-\mu\right]=0$. The last entry of the Fisher matrix is :

$$\begin{aligned}\mathcal{I}(\mu,\sigma)_{22} &= -n\mathbb{E}\left[\frac{1}{\sigma^2}-\frac{3(x-\mu)^2}{\sigma^4}\right] \\ &= n\left(-\frac{1}{\sigma^2}+\frac{3\sigma^2}{\sigma^4}\right) \\ &= \frac{2n}{\sigma^2}\end{aligned}$$

because $\mathbb{E}[(x-\mu)^2]=\sigma^2$.
Hence, the Fisher matrix is :

$$\mathcal{I}(\mu,\sigma)=\begin{pmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{2n}{\sigma^2}\end{pmatrix} \tag{70}$$

Fisher matrix of a gaussian

The comparison with the equation 26 reveals that the entry $\mathcal{I}(\mu, \sigma)_{11}$ is unchanged, as expected since the parameter μ is unchanged. The off-diagonal term is zero, as before, showing that the estimators of the two parameters μ and σ are asymptotically independent. The change is the last entry $\mathcal{I}$ which is multiplied by 4 and uses σ^2 instead of σ^4.

Fisher matrix of the exponential distribution

Exercice 6

(*Fisher matrix of the exponential*) Consider the exponential distribution :

$$f(x, \lambda) = \lambda \exp(-\lambda x),$$

for any $x \geq 0$, where $\lambda > 0$ is a parameter [53]. Compute the Fisher matrix.

53. [Hoel, 1971] page 87, [Papoulis and Pillai, 2002] page 85

Fisher matrix of the exponential distribution

Solution de l'exercice 6

The log density is :

$$\log(f(x,\lambda)) = \log(\lambda) - \lambda x,$$

for any $x \geq 0$. The first partial derivative is :

$$\frac{\partial \log(f(x,\lambda))}{\partial \lambda} = \frac{1}{\lambda} - x,$$

for any $x \geq 0$. The second partial derivative is :

$$\frac{\partial \log(f(x,\lambda))}{\partial \lambda} = -\frac{1}{\lambda^2}$$

for any $x \geq 0$. Fisher information is :

$$\mathcal{I}(\lambda) = -\mathbb{E}\left[\frac{\partial \log(f(x,\lambda))}{\partial \lambda}\right] \tag{71}$$

$$= \frac{1}{\lambda^2}. \tag{72}$$

In this appendix, we provide the theorems that are used in the proof.
We analyse the convergence of the sample mean to the exact mean. The weak law of large numbers gives the convergence in probability.

Théorème 50

(Weak law of large numbers) Let X be a random variable with mean $\mathbb{E}(X) = \mu$ and finite variance $\mathbb{V}(X) = \sigma^2$. Let $X_1, \ldots, X_n$ be independent random variables with the same distribution as X. Then

$$\bar{X} \xrightarrow{\mathbb{P}} \mu.$$

On peut le généraliser à un vecteur aléatoire.

Théorème 51

(Lindeberg-Levy central limit theorem) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ be n independent random variables with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite covariance matrix $C \in \mathbb{R}^{p \times p}$. Then the random variable :

$$\bar{X} = \frac{X_1 + X_2 + ...X_n}{n}$$

converges in distribution to the multivariate normal distribution :

$$\sqrt{n}(\bar{X} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(0, C).$$

When a sequence of random variable converges in probability, we may be interested in the sequence made of the value of a continuous function on this sequence.

Théorème 52

(Continuous mapping) Let $\{X_n\}_{n\geq 0}$ be a sequence of random variable and let θ be a real constant. Let g be a function continuous at θ. If $X_n \xrightarrow{\mathbb{P}} \theta$ therefore $g(X_n) \xrightarrow{\mathbb{P}} g(\theta)$.

Références I


Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008).
A first course in order statistics.
SIAM.


Aryal, G. and Nadarajah, S. (2004).
Information matrix for beta distributions.
Serdica Mathematical Journal, 30(4) :513p–526p.


Bickel, P. J. and Doksum, K. A. (1977).
Mathematical statistics : basic ideas and selected topics.
Holden-Day, Inc.


Brémaud, P. (2012).
An introduction to probabilistic modeling.
Springer Science & Business Media.


Clopper, C. and Pearson, E. (1934).
The use of confidence or fiducial limits illustrated in the case of the binomial.
Biometrika, 26(4) :404–413.

Références II


Delattre, M. and Kuhn, E. (2019).
Estimating Fisher Information Matrix in Latent Variable Models based on the Score Function.
working paper or preprint.


Delmas, J.-F. (2010).
Introduction au calcul des probabilités et à la statistique.
Les Presses de l'ENSTA.


Greene, W. H. (2012).
Econometric analysis, Seventh Edition.
Pearson.


Guyader, A. (2010).
Introduction aux méthodes numériques.
Université Paris VI.


Hamilton, J. D. (2020).
Time series analysis.
Princeton university press.

Références III


Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The elements of statistical learning, Second Edition.
Springer.


Hoel, P. G. (1971).
Introduction to mathematical statistics.
John Wiley & sons, Inc.


Laplace, P. S. (1812).
Théorie analytique des probabilités.


Millar, R. B. (2011).
Maximum likelihood estimation and inference : with examples in R, SAS and ADMB, volume 111.
John Wiley & Sons.


Nagar, D. K., Zarrazola, E., and Sánchez, L. E. (2015).
Entropies and fisher information matrix for extended beta distribution.
Applied Mathematical Sciences, 9(80) :3983–3994.

Références IV


Papoulis, A. and Pillai, S. (2002).
Probability, random variables and stochastic processes.
Mc Graw Hill.


Ross, S. (2004).
Introduction to probability and statistics for engineers and scientists.
Elsevier. Academic Press.


Saporta, G. (2006).
Probabilités, analyse des données et statistiques.
Editions Technip.


Steering Committee of the Physicians' Health Study Research Group (1989).
Final report on the aspirin component of the ongoing physicians' health study.
N Engl J Med, 321 :129–135.


Tassi, P. (1989).
Méthodes statistiques, 2ème édition.
Economica.

Références V


U.S. Census Bureau, S. A. o. t. U. S. (2012).
Table 209. cumulative percent distribution of population by height and sex : 2007-2008.


Vaart, A. W. V. D. (2000).
Asymptotic Statistics.
Cambridge Series in Statistical and Probabilistic Mathematics.


Wasserman, L. (2004).
All of statistics : a concise course in statistical inference.
Springer.


Wasserman, L. (2006).
All of nonparametric statistics.
Springer.


Wilson, E. (1927).
Probable inference, the law of succession, and statistical inference.
Journal of the Americal Statistical Association, 22(158) :209–212.

Références VI


Wonnacott, T. H. and Wonnacott, R. J. (1977).
Introductory statistics, Third Edition.
Wiley.