Mathématiques Statistiques

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

16 octobre 2022

In this part, we focus on the step C of the ABC method. Using a sample of the output $\mathbf{Y}$, we want to estimate the quantity of interest e.g. the mean of the output.

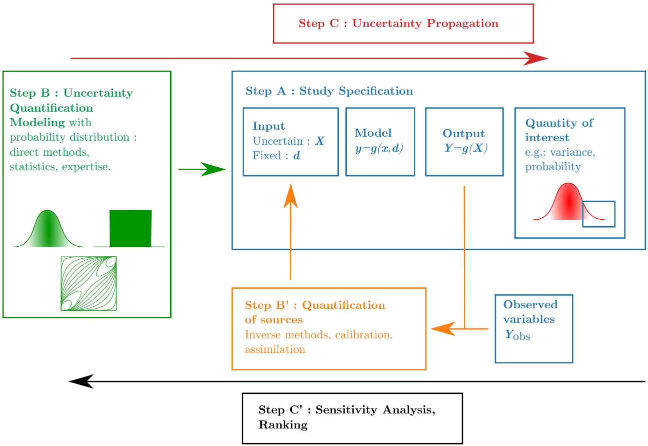


Figure 1 – Steps in the ABC method in the general case where $\mathbf{Y}$ is a random vector.

In some cases, some of these methods can be used in the step B as well, e.g. estimate the mean of the Gaussian distribution in some marginal of the random vector $\mathbf{X}$.

Notations

Définition 1

Vecteur aléatoire en entrée

Soit $p \in \mathbb{N}$ le nombre de composantes du vecteur d'entrée.
Soit $\mathcal{X} \subset \mathbb{R}^p$ le sous-ensemble d'échantillonnage.
Soit $\mathbf{X} = (X_1, ..., X_p)^T \in \mathcal{X}$ le vecteur aléatoire d'entrée.
On note $i \in \{1, ..., p\}$ l'indice d'une composante du vecteur d'entrée $\mathbf{X}$.

Pour tout $i = 1, ..., p$, on a donc $X_i \in \mathbb{R}$.

Définition 2

Densité de probabilité du vecteur d'entrée

On note f la densité de probabilité du vecteur aléatoire $\mathbf{X}$.
Dans le contexte de l'inférence (c'est à dire de l'étape B), lorsqu'on considère un modèle probabiliste paramétrique, on peut chercher à estimer les paramètres du modèle s'ajustant au mieux à un échantillon donné de $\mathbf{X}$.
Soit $m \in \mathbb{N}$ le nombre de paramètres de la densité de probabilité f.
Soit $\Theta \subseteq \mathbb{R}^m$ le sous-ensemble des paramètres.
Alors $f(\mathbf{x}, \boldsymbol{\theta})$ est la densité de probabilité dont le vecteur de paramètres est $\boldsymbol{\theta} \in \Theta$.

Notations : variable aléatoire en sortie

On simplifie le contexte et on considère une variable aléatoire Y en sortie.
En effet, dans les cas les plus simples, on peut traiter l'une après l'autre chaque marginale de sortie (c'est-à-dire indépendamment des autres marginales).
Soit g une fonction de $\mathcal{X}$ vers $\mathbb{R}$.
On considère la variable aléatoire :

$$Y = g(\mathbf{X}).$$

Nous allons estimer l'espérance de Y :

$$\mathbb{E}(Y) = \int_{\mathcal{X}} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}.$$

De plus, nous allons estimer la variance de Y :

$$\mathbb{V}(Y) = \mathbb{E}\left[(Y - \mathbb{E}(Y))^2\right].$$

Pour un seuil $s \in \mathbb{R}$ fixé, on peut souhaiter la probabilité de dépasser le seuil s :

$$\mathbb{P}(Y > s) = \int_{\mathcal{X}} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

où $\mathbb{1}_{g(\mathbf{X})>s}$ est la fonction indicatrice définie par :

$$\mathbb{1}_{g(\mathbf{x})>s}(\mathbf{x}) = \begin{cases} 1, & \text{si } g(\mathbf{x}) > s, \\ 0, & \text{sinon.} \end{cases}$$

Notations : échantillon Monte-Carlo simple

Soit $n \in \mathbb{N}$ un entier représentant la taille de l'échantillon.
Soit $\left\{\mathbf{x}^{(j)}\right\}_{j=1,\ldots,n}$ un échantillon i.i.d. du vecteur aléatoire $\mathbf{X}$. Par conséquent, la réalisation $x_i^{(j)}$ est la i-ème composante de la j-ème réalisation, pour $i = 1, ..., p$ et $j = 1, ..., n$. Le plan d'expériences A associé est donc :

$$A = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{pmatrix}$$

où chaque ligne représente une réalisation du vecteur aléatoire X et chaque colonne représente une composante.
Soit

$$y^{(j)} = g\left(\mathbf{x}^{(j)}\right)$$

pour $j = 1, ..., n$. L'estimateur Monte-Carlo de la moyenne empirique est :

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y^{(j)}.$$

Notations : la loi gaussienne

La définition suivante introduit la loi gaussienne [1] qui sera souvent utilisée par la suite.

Définition 3

(*Loi gaussienne univariée*) Soient $\mu \in \mathbb{R}$ et $\sigma > 0$ deux paramètres. La variable X suit la loi gaussienne de paramètres μ et σ^2 si sa densité de probabilité est :

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

pour tout $x \in \mathbb{R}$. Dans ce cas, on écrit $X \sim \mathcal{N}(\mu, \sigma^2)$.

Lorsqu'on écrit des mathématiques, le second paramètre est la variance σ^2, mais dans la plupart des logiciels le second paramètre est l'écart-type σ.

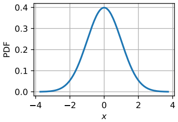


Figure 2 – La courbe en cloche pour $\mu = 0$ et $\sigma = 1$.

1. [Ross, 2004] page 168

Des probabilités aux statistiques

- Dans le cas général, lors de la propagation des incertitudes dans un modèle ou un code de calcul, la relation entrées-sortie(s) est complexe
- En fonction des lois de probabilité des entrées du modèle ou du code, déterminer la loi de la sortie Y n'est pas nécessairement facile.
- En général, on se limite à l'*estimation* d'une quantité d'intérêt : la moyenne, la variance, une probabilité de dépassement de seuil, un quantile, etc.
- On utilise la *statistique inférentielle* : à partir de n observations de la sortie, on approche la quantité d'intérêt.

De $\mathbf{X}$ à Y

On souhaite propager les incertitudes représentées par le vecteur aléatoire $\mathbf{X}$ au travers du modèle g :

$$Y = g(\mathbf{X}).$$

L'échantillon de $\mathbf{X}$ (de loi connue) :

$$\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\right\}$$

mène à un échantillon de Y :

$$\left\{y^{(1)} = g\left(\mathbf{x}^{(1)}\right), y^{(2)} = g\left(\mathbf{x}^{(2)}\right), \ldots, y^{(n)} = g\left(\mathbf{x}^{(n)}\right)\right\}.$$

C'est la méthode "*push-forward*".
Les étapes principales de la propagation des incertitudes des entrées vers la sortie sont :

- définir la loi du vecteur aléatoire $\mathbf{X}$;
- générer un échantillon de $\mathbf{X}$;
- générer un échantillon de Y ;
- estimer les quantités d'intérêt de Y : moyenne, variance, probabilité de dépasser un seuil.

De $\mathbf{X}$ à Y

- A partir de n réalisations $\left(y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right)$ d'une variable aléatoire Y, l'estimation consiste à approcher la *quantité d'intérêt* de Y comme par exemple la moyenne, la probabilité de dépassement d'un seuil donné ou un quantile
- Estimation paramétrique : dans certains cas, on pourra faire l'hypothèse que la loi de Y appartient à une certaine famille de lois $f_Y(y, \boldsymbol{\theta})$ paramétrées avec un vecteur $\boldsymbol{\theta}$ inconnu. C'est $\boldsymbol{\theta}$ qui sera estimé.
- Hypothèse : l'échantillon de taille n a été tiré de façon à assurer la représentativité de la loi :

  réalisations équiprobables et indépendantes

- Chaque valeur observée y_i est une réalisation aléatoire de la variable Y. l'échantillon de taille n est donc aléatoire :

  l'estimateur est une variable aléatoire

- Point de vue du statisticien : l'échantillon de taille n est une réalisation des variables $(Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)})$ de même loi que la loi parente Y et indépendantes, indépendantes et identiquement distribuées (i.i.d.)

Introduction à l'estimation

- Par exemple, l'espérance de Y estimée par la moyenne empirique à partir de l'échantillon $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

- La moyenne empirique $\bar{y}$ est une réalisation de la variable aléatoire $\bar{Y}$:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y^{(i)}$$

  où les $Y^{(i)}$ sont indépendants et de même loi que Y (i.i.d.)

Notation : simplification

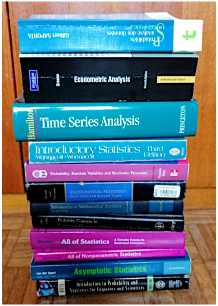
On procède à deux simplifications dans les notations

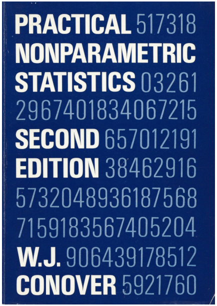
- Dans la suite, on concentre l'analyse sur la variable aléatoire de sortie Y : le rôle du vecteur aléatoire d'entrée $\mathbf{X}$ est ignoré.
  C'est pourquoi on considère une variable aléatoire, notée X.
- Dans ce contexte, il n'y a pas d'ambiguïté entre l'indice de la réalisation dans l'échantillon j et l'indice de la composante du vecteur i (puisque c'est un vecteur de dimension 1).
  C'est pourquoi, on note $\{x_1, x_2, ..., x_n\}$ l'échantillon et $\{X_1, X_2, ..., X_n\}$ les variables aléatoires associées.

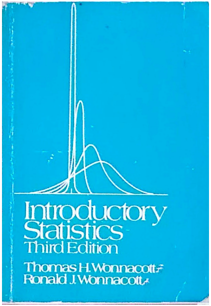
Bibliographie chronologique commentée des statistiques mathématiques : livres

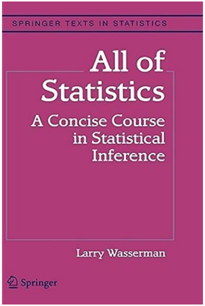
- [Hoel, 1971] : une introduction assez complète
- [Wonnacott and Wonnacott, 1977] : les résultats élémentaires très bien illustrés
- [Bickel and Doksum, 1977] : une approche mathématique
- [Conover, 1980] : les statistiques non paramétriques, chapitre 2 sur l'inférence, chapitre 3 sur l'utilisation de la loi binomiale (page 117 sur les limites de tolérance), chapitre 6 sur le test de Kolmogorov-Smirnov
- [Vaart, 2000] : les statistiques asymptotiques
- [Papoulis and Pillai, 2002] : chapitre 8 sur les statistiques. Des diapos sont disponibles sur Lien
- [Saporta, 2006] : un texte Français
- [Wasserman, 2006] : un résumé avec la plupart des résultats essentiels et des démonstrations (très) concises
- [Arnold et al., 2008] : sur les statistiques d'ordre, la loi du quantile
- [Greene, 2012] : beaucoup de résultats théoriques, une approche peu formelle
- [Guyader, 2016] : un cours universitaire

Bibliographie commentée des statistiques mathématiques









Statistique - Estimateur

La définition suivante introduit la notion de statistique [2].

Définition 4

(*Statistique*) Soit X une variable aléatoire. Soit $\{x_1, \ldots, x_n\}$ un échantillon de taille n de la variable aléatoire X. Une statistique T est une fonction de l'échantillon :

$$T = T(x_1, \ldots, x_n).$$

2. [Bickel and Doksum, 1977] page 117

Statistique - Estimateur

Définition 5

(*Estimateur*) Soit X une variable aléatoire. Soit $\theta \in \mathbb{R}$ un paramètre associé à X. Soient $X_1, \ldots, X_n$ de même loi que X. Un estimateur $\hat{\theta}$ est l'application d'une statistique à $\{X_1, \ldots, X_n\}$ dans le but d'approcher θ :

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) \approx \theta.$$

Une *estimation* est une réalisation de la variable aléatoire $\hat{\theta}$

On distingue la valeur du paramètre θ et son estimateur $\hat{\theta}$ en ajoutant un "chapeau". L'estimateur $\hat{\theta}$ est donc une variable aléatoire, qui possède une distribution.

Exemple 6

(*Estimation de la moyenne d'une variable gaussienne*) Soit X une variable de loi gaussienne $\mathcal{N}(\mu, \sigma)$.
La moyenne empirique $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ est un estimateur de μ.
Il serait tout à fait cohérent de le noter $\hat{\mu}$ (mais ce n'est pas standard).

Le biais est l'écart entre l'espérance de l'estimateur et la vraie valeur [3].

Définition 7

(*Biais*) Supposons que $\hat{\theta}$ est un estimateur de θ. Le biais est :

$$Biais\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta}\right) - \theta$$

Définition 8

Un estimateur est sans biais si

$$\mathbb{E}\left(\hat{\theta}\right) = \theta.$$

Un estimateur est asymptotiquement sans biais si :

$$\lim_{n\to\infty} \mathbb{E}\left(\hat{\theta}\right) = \theta$$

3. [Bickel and Doksum, 1977] page 119, [Wonnacott and Wonnacott, 1977] page 180

Statistique - Estimateur

L'erreur quadratique moyenne est l'espérance du carré de la différence [4].

Définition 9

(*Erreur quadratique moyenne*) L'erreur quadratique moyenne MSE ("Mean Squared Error") est :

$$MSE\left(\hat{\theta}\right) = \mathbb{E}\left(\left(\hat{\theta} - \theta\right)^2\right). \tag{1}$$

Exemple 10

(*Estimation de la densité par une méthode non paramétrique*) Considérons par exemple l'histogramme c'est-à-dire l'estimateur non paramétrique de la densité de probabilité $f(x)$ en un point x donné.
Dans ce contexte, on recherche la largeur de classe h permettant de réaliser le meilleur compromis entre le biais et la variance dans le but de minimiser l'erreur quadratique moyenne.

4. [Wonnacott and Wonnacott, 1977] page 184

Statistique - Estimateur

En général on s'intéressera aux estimateurs sans biais pour lesquels on retiendra ceux de variance minimale et convergent *rapidement* en moyenne quadratique.

Remark 1

(Convergence en moyenne quadratique) Supposons que l'estimateur $\hat{\theta}$ est sans biais. Si il convergence en moyenne quadratique, alors sa variance tend vers zéro.
En effet,

$$\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right]=\mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left(\hat{\theta}\right)\right)^2\right]=\mathbb{V}\left(\hat{\theta}\right).$$

Supposons que l'estimateur convergence en moyenne quadratique :

$$\lim_{n\to\infty}\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right]=0.$$

Par conséquent, la variance de l'estimateur tend vers zéro :

$$\lim_{n\to\infty}\mathbb{V}\left(\hat{\theta}\right)=0.$$

Statistique - Estimateur

Un estimateur $\hat{\theta}$ peut être non biaisé ou biaisé [5].

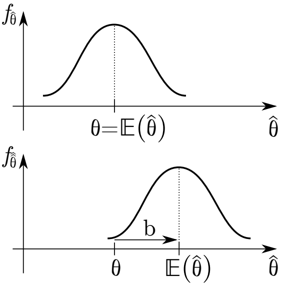


Figure 3 – En haut, un estimateur $\hat{\theta}$ non biaisé. En bas, un estimateur biaisé, de biais $b = \mathbb{E}\left(\hat{\theta}\right) - \theta$.

Exemple 11

(*Estimation de la variance*) L'estimateur de la variance de la population $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ est biaisé, alors que l'estimateur $\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ est non biaisé. Dans l'exercice 6, on propose d'expérimenter cela en Python.

5. [Wonnacott and Wonnacott, 1977] page 181

Statistique - Estimateur

Deux estimateurs $\hat{\theta}$ et $\hat{\hat{\theta}}$ peuvent avoir des variances différentes [6].

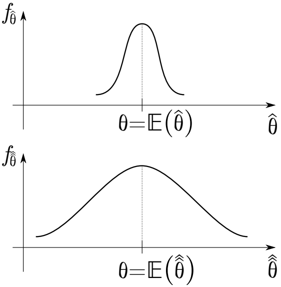


Figure 4 – Deux estimateurs non biaisés. L'estimateur du haut a une variance plus petite que celui du bas : l'estimateur du haut est plus efficace.

Exemple 12

(*Estimation de l'espérance*) L'estimateur de l'espérance fondé sur la médiane possède une variance supérieure à la moyenne empirique. Dans l'exercice 7, on propose d'expérimenter cela en Python.

6. [Wonnacott and Wonnacott, 1977] page 182

Coefficient de variation

Pour mesurer l'erreur relative d'estimation, on peut diviser l'écart-type par la moyenne, ce qui mène au coefficient de variation.

Définition 13

(*Coefficient de variation*) Le coefficient de variation de l'estimateur $\hat{\theta}$ est :

$$CV\left(\hat{\theta}\right) = \frac{\sqrt{\mathbb{V}\left(\hat{\theta}\right)}}{\mathbb{E}\left(\hat{\theta}\right)} \tag{2}$$

si $\mathbb{E}\left(\hat{\theta}\right) \neq 0$.

Coefficient de variation

Parfois, on exprime l'équation précédente en fonction de l'écart-type.
Pour toute variable aléatoire, l'écart-type est :

$$SD(X) = \sqrt{\mathbb{V}(X)}.$$

L'écart-type $SD(X)$ est de même "unité" physique que X. Par exemple, si X est en kg, alors $SD(X)$ est aussi en kg.
Par conséquent,

$$CV\left(\hat{\theta}\right) = \frac{SD\left(\hat{\theta}\right)}{\mathbb{E}\left(\hat{\theta}\right)}.$$

Coefficient de variation

Exemple 14

(*Coefficient de variation*) Supposons que nous disposons d'un estimateur non biaisé $\hat{\theta}$ du paramètre $\theta = 1$. Par conséquent, on a $\mathbb{E}\left(\hat{\theta}\right) = 1$.

Supposons que l'écart-type de l'estimateur est $\sqrt{\mathbb{V}\left(\hat{\theta}\right)} = 0.1$. Alors le coefficient de variation est

$$CV\left(\hat{\theta}\right) = \frac{0.1}{1} = 0.1$$

Le coefficient de variation est relativement petit, ce qui implique que l'estimation est relativement précise. Un coefficient de variation égal à 0.1=10% signifie que l'écart-type de l'estimateur représente un dixième de l'estimateur.

Supposons que l'écart-type de l'estimateur est $\sqrt{\mathbb{V}\left(\hat{\theta}\right)} = 2$. Alors le coefficient de variation est

$$CV\left(\hat{\theta}\right) = \frac{2}{1} = 2.$$

C'est un coefficient de variation relativement grand, qui implique que l'estimation est relativement peu précise.

Convergence of random variables

There are several modes of convergence that we commonly use when we consider estimators.
The first is the convergence in probability [7].

Définition 15

(*Convergence in probability*) Let $\{X_n\}_{n\geq 0}$ be a sequence of random variable and let X be a random variable. The sequence X_n converges to the random variable X in probability if, for any $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \rightarrow 0 \tag{3}$$

when $n \rightarrow +\infty$. In this case, we write :

$$X_n \xrightarrow{\mathbb{P}} X. \tag{4}$$

The previous definition is also called *convergence in measure* or *stochastic convergence*. The equation 3 can be written :

$$\lim_{n\rightarrow\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

7. [Bickel and Doksum, 1977] page 460, [Papoulis and Pillai, 2002] page 274

Convergence in distribution

Another type of convergence is the convergence in distribution [8]

Définition 16

(*Convergence in distribution*) Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables and let X be a random variable. Let F be the distribution function of X, i.e. $F : x \to \mathbb{P}(X \leq x)$ for any $x \in \mathbb{R}$. Let F_n be the distribution function of X_n. The sequence X_n converges to the random variable X in distribution if :

$$F_n(x) \to F(x) \tag{5}$$

when $n \to +\infty$ for any point $x \in \mathbb{R}$ where F is continuous. In this case, we write :

$$X_n \xrightarrow{d} X. \tag{6}$$

The equation is sometimes written :

$$X_n \xrightarrow{\mathcal{L}} X, \quad \text{or} \quad X_n \xrightarrow{\mathcal{D}} X.$$

8. [Bickel and Doksum, 1977] page 460, [Papoulis and Pillai, 2002] page 274

Convergence in distribution

The *Portmanteau lemma* presents several alternative definitions of the convergence in distribution [9]. The following theorem presents one of them.

Théorème 17

(Convergence in distribution) Let $\{X_n\}_{n\geq 0}$ be a sequence of random variable and let X be a random variable. The sequence X_n converges to the random variable X in distribution if and only if :

$$\mathbb{E}(F_n(X_n)) \to \mathbb{E}(F(X)), \tag{7}$$

when $n \to +\infty$ where F_n is the cumulative distribution function of X_n and F is the cumulative distribution function of X. In this case, we write

$$X_n \xrightarrow{d} X.$$

9. [Vaart, 2000] page 6

Définition 18

(*Mean squared error*) Let $\{X_n\}_{n\geq 0}$ be a sequence of random variable and let X be a random variable. The mean squared error is :

$$MSE(X_n) = \mathbb{E}\left((X_n - X)^2\right) \tag{8}$$

for any $n \geq 0$.

Sometimes, computing the expectation of the square of the difference $X_n - X$ is easier, in which case the following definition is used [10].

Définition 19

(*Convergence in the mean square sense*) Let $\{X_n\}_{n\geq 0}$ be a sequence of random variable and let X be a random variable. The sequence X_n converges to the random variable X in the mean square sense if :

$$MSE(X_n) = \mathbb{E}\left((X_n - X)^2\right) \to 0 \tag{9}$$

when $n \to +\infty$. In this case we write :

$$X_n \xrightarrow{MS} X.$$

10. [Papoulis and Pillai, 2002] page 274

Si l'estimateur $\hat{\theta}$ est non biaisé, alors $\mathbb{E}\left(\hat{\theta})\right) = \theta$, ce qui implique :

$$\mathbb{V}\left(\hat{\theta}\right) = \mathbb{E}\left(\left(\hat{\theta} - \mathbb{E}\left(\hat{\theta}\right)\right)^2\right) = \mathbb{E}\left(\left(\hat{\theta} - \theta\right)^2\right) = MSE\left(\hat{\theta}\right).$$

Par conséquent, si la variance d'un estimateur non biaisé tend vers zéro, alors l'estimateur est convergent en moyenne quadratique.

Inclusions of the previous modes of convergence [11].

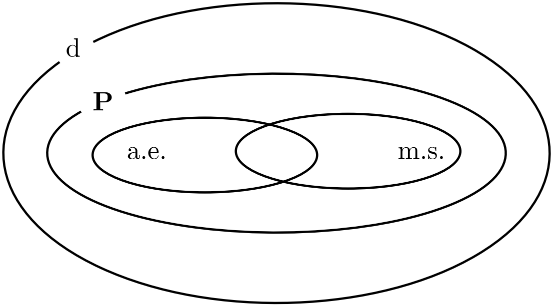


Figure 5 – Different modes of convergences of random variables

11. [Papoulis and Pillai, 2002] page 275

Markov's inequality [12] is the following.

Théorème 20

(Markov's inequality) Let X be a random variable that takes only nonnegative values. Therefore, for any $a > 0$,

$$\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}(X)}{a}. \tag{10}$$

Chebyshev's inequality [13] is the following.

Théorème 21

(Chebyshev's inequality) Let X be a random variable with finite variance. Therefore, for any $a > 0$,

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq a\right) \leq \frac{\mathbb{V}(X)}{a^2}. \tag{11}$$

12. [Papoulis and Pillai, 2002] page 151, [Ross, 2004] page 127
13. [Papoulis and Pillai, 2002] page 151, [Ross, 2004] page 128

Théorème 22

(MS convergence implies convergence in probability) Si $\hat{\theta}$ est un estimateur convergent en moyenne quadratique alors il est convergent en probabilité.

Preuve[14]

On applique l'inégalité de Markov à la variable $X = \left(\hat{\theta} - \theta\right)^2$ avec le réel $a = \epsilon^2$. Pour tout $\epsilon > 0$, l'inégalité de Markov implique :

$$\mathbb{P}\left(\left(\hat{\theta} - \theta\right)^2 \geq \epsilon^2\right) \leq \frac{\mathbb{E}\left(\left(\hat{\theta} - \theta\right)^2\right)}{\epsilon^2}.$$

Par conséquent,

$$\mathbb{P}\left(\left|\hat{\theta} - \theta\right| \geq \epsilon\right) \leq \frac{MSE\left(\hat{\theta}\right)}{\epsilon^2}.$$

Pour ϵ fixé, la convergence en moyenne quadratique implique $\lim_{n\to\infty} MSE\left(\hat{\theta}\right) = 0$. Par conséquent, le membre de droite de l'inégalité précédente tend vers zéro, ce qui implique que le membre de gauche également. □

14. [Papoulis and Pillai, 2002] page 275

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 [15].

Théorème 23

(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* $g(X_n) \xrightarrow{\mathbb{P}} g(\theta)$.

15. [Bickel and Doksum, 1977] page 460, [Vaart, 2000] page 7

Théorème central limite

The central limit theorem is extremely useful [16].

Théorème 24

(Central limit theorem)
Let $X_1, \ldots, X_n$ be n independent random variables with mean μ and finite variance σ^2. Then the random variable :

$$Z_n = \frac{X_1 + X_2 + ... + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard Gaussian distribution :

$$Z_n \xrightarrow{d} \mathcal{N}(0, 1).$$

On peut en déduire que :

$$\frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

où

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n).$$

16. [Wasserman, 2006] page 77

Exemple 25

(*Application du TCL à la loi log-normale*) Soit $Z \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu = 2$ et $\sigma = 1$.
On considère $X = \exp(Z) \sim \text{Lognormal}(\mu, \sigma^2)$.
Pour une taille d'échantillon n donnée, on considère la moyenne empirique $\bar{X}$.
On souhaite observer la distribution de la moyenne empirique $\bar{X}$ lorsque $n = 64$.
Pour cela, on répète $r = 1000$ fois l'expérience suivante : on génère un échantillon de taille n et on calcule la moyenne empirique [17].
Puis on estime la densité de probabilité de la distribution de $\bar{X}$ par une méthode de lissage par noyau.

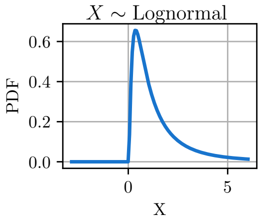


Figure 6 – Densité de probabilité d'une variable log-normale.

17. `convergence-mean-lognormal.py`

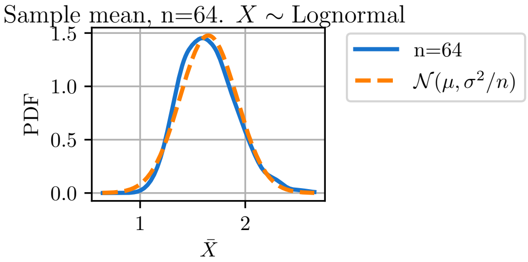


Figure 7 – Distribution de la moyenne empirique d'une variable log-normale et comparaison avec la distribution gaussienne.

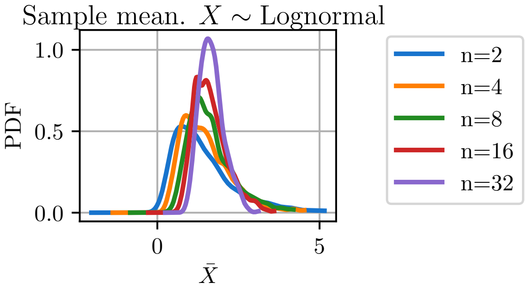


Figure 8 – Convergence de la distribution de la moyenne empirique d'une variable log-normale lorsque la taille de l'échantillon augmente.

Le graphique suivant [18] présente la distribution de $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$.

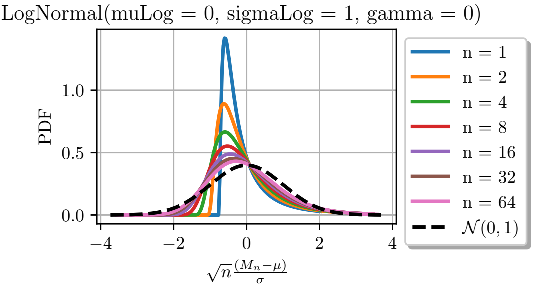


Figure 9 – Convergence de la distribution de $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ d'une variable log-normale lorsque la taille de l'échantillon augmente.

18. `convergence-TCL.py`, avec l'aide de Anne Dutfoy

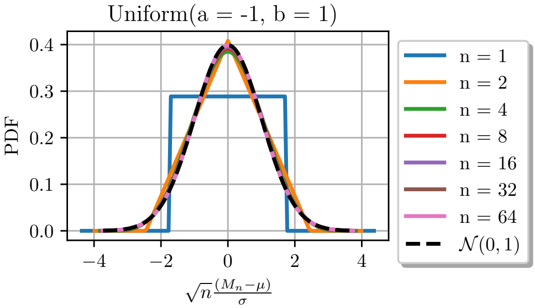


Figure 10 – Convergence de la distribution de $\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma}$ d'une variable uniforme lorsque la taille de l'échantillon augmente.

On peut généraliser le TCL à un vecteur aléatoire [19].

Théorème 26

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

$$\bar{\mathbf{X}} = \frac{\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_n}{n}$$

converges in distribution to the multivariate normal distribution :

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

19. [Wasserman, 2006] page 78, [Greene, 2012] page 1122

Delta méthode : cas général

Pour démontrer le théorème de la delta-méthode, nous aurons besoin du théorème de Slutsky [20].

Théorème 27

(Slutsky) Let X_n and Y_n be two sequences of random variables. Let X be a random variable and let a be a constant. If $X_n \xrightarrow{d} X$ and if $Y_n \xrightarrow{\mathbb{P}} a$ then :

$$X_n + Y_n \xrightarrow{d} X + a, \quad n \to \infty,$$

and

$$X_n Y_n \xrightarrow{d} aX, \quad n \to \infty.$$

20. [Bickel and Doksum, 1977] page 461

Delta method for a random variable
La delta-méthode est un outil très utile en pratique pour démontrer la convergence vers la loi gaussienne et évaluer la variance [21]

Théorème 28

(Méthode delta pour une variable aléatoire) Let X_n be a sequence of random variables. Let $\{u_n\}_{n\geq 0}$ be a sequence of constants tending to infinity :

$$\lim_{n\to+\infty} u_n = +\infty.$$

Let $\theta \in \mathbb{R}$ be a constant value. Assume that g is differentiable at θ and that g' is continuous at θ. If the following convergence in distribution holds :

$$u_n\left(X_n - \theta\right) \xrightarrow{d} X, \tag{12}$$

when $n \to \infty$, then :

$$u_n\left(g(X_n) - g(\theta)\right) \xrightarrow{d} g'(\theta)X, \tag{13}$$

when $n \to \infty$.

21. [Bickel and Doksum, 1977] page 461

Développement de Taylor

Remark 2

(Développement de Taylor de la fonction) Si g est dérivable en θ, alors[22]

$$g(X_n) \approx g(\theta) + g'(\theta)(X_n - \theta)$$

si $X_n \to \theta$.
Par conséquent[23],

$$\mathbb{V}(g(X_n)) \approx [g'(\theta)]^2 \mathbb{V}(X_n)$$

quand n est grand.
En termes d'écart-types :

$$SD(g(X_n)) \approx |g'(\theta)| SD(X_n)$$

quand n est grand.

22. [Greene, 2012] page 1124
23. [Papoulis and Pillai, 2002] page 150

Delta méthode : cas Gaussien

The theorem 28 is often applied to a Gaussian distribution with the sequence $u_n = \sqrt{n}$, which leads to the following theorem [24].

Théorème 29

(Delta method for a Gaussian random variable) Let X_n be a sequence of random variables. Let $\theta \in \mathbb{R}$ be a constant value. Let $\sigma > 0$ be a known standard deviation. Assume that g is differentiable at θ and that :

$$g'(\theta) \neq 0.$$

If the following convergence in distribution holds :

$$\sqrt{n}\,(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \tag{14}$$

when $n \to \infty$, then :

$$\sqrt{n}\,(g(X_n) - g(\theta)) \xrightarrow{d} \mathcal{N}\left(0, g'(\theta)^2\sigma^2\right), \tag{15}$$

when $n \to \infty$.

24. [Wasserman, 2006] page 79, [Greene, 2012] page 1123

Delta méthode et TCL

Remark 3

(Développement de Taylor de la fonction) Soit X une variable aléatoire de moyenne μ et de variance σ^2. Soient X_1, ..., X_n des variables aléatoires indépendantes et de même loi que X. Le T.C.L. implique :

$$\sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Soit g une fonction dérivable en μ telle que $g'(\mu) \neq 0$. Alors la delta méthode implique :

$$\sqrt{n}\left(g\left(\bar{X}\right) - g(\mu)\right) \xrightarrow{d} \mathcal{N}\left(0, g'(\mu)^2\sigma^2\right).$$

Delta méthode : cas vectoriel

The delta-method theorem has a vector version [25], that we will use in the theorem 45.

Théorème 30

(Delta method for a Gaussian random vector) Let $\mathbf{X}_n = (X_{n1}, ..., X_{np})^T \in \mathbb{R}^p$ *be a sequence of random vectors. Let* $\boldsymbol{\theta} \in \mathbb{R}^p$ *be a constant vector. Let* $\Sigma \in \mathbb{R}^{p \times p}$ *be a known covariance matrix. Assume that* $g : \mathbb{R}^p \to \mathbb{R}$ *is differentiable at* θ *and let :*

$$\nabla g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial g}{\partial x_p}(\mathbf{x}) \end{pmatrix}$$

for any $\mathbf{x} \in \mathbb{R}^p$*. Assume that* $\nabla g(\boldsymbol{\theta}) \neq \mathbf{0}$*. If the following convergence in distribution holds :*

$$\sqrt{n}\,(\mathbf{X}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{16}$$

when $n \to \infty$*, then :*

$$\sqrt{n}\,(g(\mathbf{X}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\boldsymbol{\theta})^T \Sigma \nabla g(\boldsymbol{\theta})\right), \tag{17}$$

when $n \to \infty$.

25. [Wasserman, 2006] page 79

Delta method and the TCL

Note that $\nabla g(\boldsymbol{\theta})^T \Sigma \nabla g(\boldsymbol{\theta}) \in \mathbb{R}$ in the previous definition.
We will use the previous theorem in the theorem 45.

Remark 4

(Delta method and the TCL) Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. The Lindeberg-Levy multivariate TCL implies :

$$\sqrt{n}\left(\overline{\mathbf{X}} - \boldsymbol{\mu}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma\right)$$

when $n \rightarrow \infty$. Assume that $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is differentiable at $\boldsymbol{\mu}$ and that $\nabla g(\boldsymbol{\theta}) \neq \mathbf{0}$. Therefore, the multivariate delta method implies :

$$\sqrt{n}\left(g\left(\overline{\mathbf{X}}\right) - g(\boldsymbol{\mu})\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\boldsymbol{\mu})^T \Sigma \nabla g(\boldsymbol{\mu})\right),$$

when $n \rightarrow \infty$.

Propagation et estimation

- Soit $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ un vecteur aléatoire de densité de probabilité f.
- Soit $g : \mathcal{X} \rightarrow \mathbb{R}$ une fonction :
$$Y = g(\mathbf{X})$$
- L'espérance de Y est :
$$\mathbb{E}(Y) = \int_{\mathcal{X}} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$
- Soit s un seuil.
- Soit $\mathbb{1}_{g(\mathbf{x})>s}$ la fonction indicatrice définie par :

$$\mathbb{1}_{g(\mathbf{x})>s}(\mathbf{x}) = \begin{cases} 1, & \text{si } g(\mathbf{x}) > s, \\ 0, & \text{sinon.} \end{cases}$$

Propagation et estimation

- La probabilité $\mathbb{P}(Y > s)$ est l'espérance de la fonction indicatrice :

$$\mathbb{P}(Y > s) = \mathbb{E}\left(\mathbb{1}_{g(\mathbf{x})>s}(\mathbf{X})\right) = \int_{\mathcal{X}} \mathbb{1}_{g(\mathbf{x})>s}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}.$$

- Soit $\{y_1, \ldots, y_n\}$ un échantillon de n réalisations indépendantes de Y.
- Pour les obtenir, on génère n réalisations indépendantes de $\mathbf{X}$, notées $\mathbf{x}_1$, ..., $\mathbf{x}_n$ et on évalue les sorties :

$$y_i = g(\mathbf{x}_i)$$

pour $i = 1, \ldots, n$.

Propagation et estimation

- Tous les résultats que nous souhaitons sont sur la sortie Y : moyenne, variance, dépassement de seuil, etc...
- Or Y est une variable scalaire, réelle.
- C'est pourquoi, pour simplifier les notations, on ignore dans la suite la dépendance entre la sortie et les entrées via la fonction g.
- Attention : changement de notation !
- On suppose que X est une variable aléatoire réelle, scalaire.
- Etant donné n, on génère les réalisations indépendantes $x_1, \ldots, x_n$ de même loi que X.
- Puis, nous allons estimer la moyenne de X, sa variance et d'autres quantités d'intérêt associées à X.

Estimation de la moyenne

La définition suivante introduit la moyenne empirique[26].

Définition 31

(*Moyenne empirique*) Soit X une variable aléatoire. Soient $x_1, \ldots, x_n$ des réalisations indépendantes de X.
La moyenne empirique est :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Définition 32

(*Moyenne empirique*) Soit X une variable aléatoire. Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi que X. La variable aléatoire associée à la moyenne empirique est :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

26. [Wonnacott and Wonnacott, 1977] page 147,[Ross, 2004] page 17

Estimation de la moyenne

La variable aléatoire $\bar{X}$ est un estimateur de l'espérance μ sans biais [27].

Théorème 33

(Espérance de la moyenne empirique) Soit X une variable aléatoire de moyenne $\mu = \mathbb{E}(X)$. Soit $\{X_1, \ldots, X_n\}$ un échantillon i.i.d. de X. Sous les hypothèses précédentes, on a :

$$\mathbb{E}\left(\bar{X}\right) = \mu \tag{18}$$

Le théorème précédent montre que la moyenne empirique est un estimateur sans biais.

27. [Wonnacott and Wonnacott, 1977] page 147, [Ross, 2004] page 203

Démonstration.

$$\begin{aligned}
\mathbb{E}\left(\bar{X}\right) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu
\end{aligned}$$

ce qui mène à l'équation 18. □

Variance de la moyenne empirique

Le théorème suivant indique la variance de la moyenne empirique [28].

Théorème 34

(Variance de la moyenne empirique) Soit X une variable aléatoire de moyenne $\mu = \mathbb{E}(X)$. Soit $\{X_1, \ldots, X_n\}$ un échantillon i.i.d. de X. Supposons que $\mathbb{V}(X) = \sigma^2$ est finie. Alors :

$$\mathbb{V}\left(\bar{X}\right) = \frac{\sigma^2}{n}. \tag{19}$$

Le théorème précédent montre que l'estimateur empirique de la moyenne est convergent, puisque

$$\mathbb{V}\left(\bar{X}\right) = \frac{\sigma^2}{n} \to 0$$

lorsque $n \to \infty$. On va le démontrer grâce à la loi des grands nombres. Toutefois, remarquons que, si la variance de X est infinie, alors la convergence n'est pas assurée par le théorème précédent.

28. [Wonnacott and Wonnacott, 1977] page 147, [Ross, 2004] page 203

Démonstration.

$$
\begin{aligned}
\mathbb{V}\left(\bar{X}\right) &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i) \text{ (car les } X_i \text{ sont indépendants)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2} n\sigma^2
\end{aligned}
$$

ce qui mène à l'équation 19. □

- La vitesse de convergence est en $\frac{1}{\sqrt{n}}$:

$$\sqrt{\mathbb{V}\left(\bar{X}\right)} = \frac{\sigma}{\sqrt{n}}$$

- Le coefficient de variation de la moyenne empirique est :

$$CV\left(\bar{X}\right) = \frac{\sqrt{\mathbb{V}\left(\bar{X}\right)}}{\mathbb{E}\left(\bar{X}\right)} = \frac{\sigma}{\mu\sqrt{n}}.$$

- Pour diviser le CV par 2, il faut multiplier n par 4.

Estimate the mean of a sample

Exemple 35

(*Distribution of the sample mean with a gaussian distribution*[29]) We consider $X_1, X_2, X_3, X_4 \sim \mathcal{N}(0,1)$, independent, and a sample size equal to $n = 4$. The following plot is the distribution of

$$\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4) \sim \mathcal{N}\left(0, \sigma^2\right)$$

with $\sigma = \frac{1}{2}$. Indeed, we have $\sigma^2 = \left(\frac{1}{4}\right)^2 (1+1+1+1) = \frac{4}{4^2} = \frac{1}{4}$.

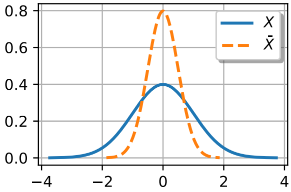


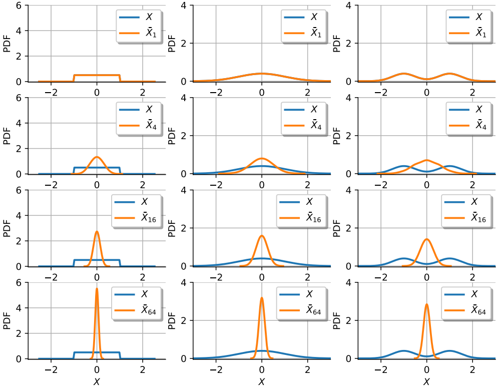
Figure 11 – Probability density of X and the sample mean $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$.

29. [Wonnacott and Wonnacott, 1977] page 148

Estimate the mean of a sample

Exemple 36

(*Distribution of the sample mean with increasing sample size*[30]) We consider several distributions of the random variable X and plot the distribution of $\bar{X}$ for an increasing value of $n = 1, 4, 32, 64$.



30. [Wonnacott and Wonnacott, 1977] page 152

We now analyse the convergence of the sample mean to the exact mean. The weak law of large numbers gives the convergence in probability [31].

Théorème 37

(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

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

31. [Ross, 2004] page 129, [Greene, 2012] page 1110

Démonstration.

We apply Chevyshev's inequality (theorem 21) to the random variable $\bar{X}$. The theorem 33 and 34 imply that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\bar{X} - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

where we have used ϵ instead of a in Chebyshev's inequality. Therefore,

$$\lim_{n \rightarrow +\infty} \mathbb{P}(|\bar{X} - \mu| > \epsilon) \leq \lim_{n \rightarrow +\infty} \frac{\sigma^2}{n\epsilon^2} = 0,$$

which concludes the proof. □

Estimation de la variance

L'estimateur de la variance est le suivant [32].

Définition 38

(*Estimateur biaisé de la variance*) L'estimateur empirique (biaisé) de la variance est :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{20}$$

On peut obtenir une expression plus simple pour le calcul mathématique [33] (mais à éviter pour le calcul numérique en arithmétique flottante).

Théorème 39

On a :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2. \tag{21}$$

32. [Wonnacott and Wonnacott, 1977] page 181
33. [Ross, 2004] page 23

Estimation de la variance

Le théorème suivant montre que $\hat{\sigma}^2$ est un estimateur biaisé de la variance [34].

Théorème 40

(Biais de l'estimateur de la variance) Supposons que X est une variable aléatoire de moyenne $\bar{X}$ et de variance $\sigma^2 = \mathbb{V}(X)$. Alors :

$$\mathbb{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2. \tag{22}$$

34. [Hoel, 1971] page 191, [Greene, 2012] page 1128

Estimation de la variance

Démonstration.

On évalue l'espérance de l'équation 21, ce qui implique :

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \mathbb{E}(\bar{X}^2) \\
&= \mathbb{E}(Y^2) - \mathbb{E}(\bar{X}^2) \\
&= \mathbb{V}(X) + [\bar{X}]^2 - [\mathbb{V}\left(\bar{X}\right) + [\mathbb{E}\left(\bar{X}\right)]^2] \\
&= \sigma^2 + [\bar{X}]^2 - [\mathbb{V}\left(\bar{X}\right) + [\bar{X}]^2] \text{ (par l'eq. 18)} \\
&= \sigma^2 - \mathbb{V}\left(\bar{X}\right) \\
&= \sigma^2 - \frac{1}{n}\sigma^2 \text{ (par l'eq. 19)}
\end{aligned}$$

ce qui mène à l'équation 22. □

Estimation de la variance

L'estimateur suivant est non biaisé [35].

Théorème 41

(Unbiased estimator of the variance) The estimator :

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{23}$$

is unbiased :

$$\mathbb{E}\left(\hat{S}^2\right) = \sigma^2. \tag{24}$$

35. [Ross, 2004] page 213

Variance de la moyenne

Exemple 42

(*Variance de la moyenne*) Soit X une variable aléatoire d'espérance $\mu_X = \mathbb{E}(X)$ et de variance $\sigma_X^2 = \mathbb{V}(X)$. On considère un échantillon $\{x_1, ..., x_n\}$ de réalisations de X, que l'on suppose indépendantes. La moyenne empirique $\bar{X}$ est une variable aléatoire d'espérance $\mu_{\bar{X}} = \mu$ et de variance $\sigma_{\bar{X}}^2$ où :

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Un estimateur non biaisé de la variance de $\bar{X}$ est :

$$\hat{\sigma}_{\bar{X}}^2 = \frac{\hat{s}^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Variance de la variance empirique

La variance de la variance empirique implique le moment d'ordre 4 [36].

Théorème 43

(Variance de la variance empirique) Soit X une variable aléatoire et $\{x_1, ..., x_n\}$ un échantillon i.i.d. de X. La variance de la variance empirique baisée est :

$$\mathbb{V}\left(\hat{\sigma}^2\right) = \frac{(n-1)^2}{n^3}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right). \tag{25}$$

où $\mu_4 = \mathbb{E}((X-\mu)^4)$. De plus,

$$\mathbb{V}\left(\hat{s}^2\right) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right). \tag{26}$$

On observe que $\frac{n-3}{n-1} \to 1, \quad n \to \infty$. Par conséquent, on a l'équivalence asymptotique suivante :

$$\mathbb{V}\left(\hat{s}^2\right) \sim \frac{1}{n}\left(\mu_4 - \sigma^4\right), \quad n \to \infty. \tag{27}$$

Un exercice sur ce thème est proposé dans l'exercice 5 page 120 pour $\hat{\sigma}^2$.

36. [Greene, 2012] page 1128, [Ang and Tang, 1975] page 248

Variance de la variance empirique

Dans le cas particulier où X suit la loi gaussienne, la variance de $\hat{s}^2$ est particulièrement simple[37].

Théorème 44

(Variance de la variance empirique non biaisée dans le cas gaussien) Supposons que $X \sim \mathcal{N}(\mu, \sigma^2)$. Soit $\{x_1, ..., x_n\}$ un échantillon i.i.d. de X. Alors

$$\mathbb{V}\left(\hat{s}^2\right) = \frac{2\sigma^4}{n-1}, \qquad \mathbb{V}\left(\hat{\sigma}^2\right) = \frac{2\sigma^4}{n}.$$

On remarque que l'on obtient le résultat de l'analyse asymptotique donnée par l'équation 27, puisque le moment d'ordre 4 centré de la loi gaussienne est $\mu_4(X) = 3\sigma^4$.

37. [Papoulis and Pillai, 2002] page 261, [Papoulis and Pillai, 2002] page 313

Variance de la variance empirique

Preuve
D'après le théorème 53 page 87 (théorème de Fisher), la variable $\frac{(n-1)\hat{s}^2}{\sigma^2}$ suit la loi du χ^2 à $n-1$ degrés de liberté. Or la variance de la loi du χ^2 à $n-1$ degrés de liberté est $2(n-1)$. Par conséquent,

$$\mathbb{V}\left(\frac{(n-1)\hat{s}^2}{\sigma^2}\right) = \frac{(n-1)^2}{\sigma^4}\mathbb{V}\left(\hat{s}^2\right) = 2(n-1),$$

ce qui mène au résultat pour $\hat{s}^2$. De plus

$$\hat{s}^2 = \frac{n}{n-1}\hat{\sigma}^2$$

ce qui mène au résultat pour $\hat{\sigma}^2$. $\square$

Variance de la variance empirique

Le théorème suivant [38] présente la loi asymptotique de l'estimateur biaisé et non biaisé de la variance.

Théorème 45

(Convergence de la variance empirique) Supposons que X est une variable aléatoire possédant des moments d'ordre 1 à 4 finis. En particulier, supposons que la variance $\sigma^2 = \mathbb{E}((X-\mu)^2)$ est finie et que le moment d'ordre 4 noté $\mu_4 = \mathbb{E}((X-\mu)^4)$ est fini. La variance empirique baisiée converge en distribution vers la loi gaussienne :

$$\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{d} \mathcal{N}\left(0, \mu_4 - \sigma^4\right) \tag{28}$$

où $\mu_4 = \mathbb{E}((X-\mu)^4)$. De même,

$$\sqrt{n}\left(\hat{S}^2 - \sigma^2\right) \xrightarrow{d} \mathcal{N}\left(0, \mu_4 - \sigma^4\right) \tag{29}$$

38. [Vaart, 2000] page 26, [Guyader, 2016] page 60

Variance de la variance empirique

Preuve
La preuve donnée dans [39] s'appuie sur la delta-méthode vectorielle présentée dans le théorème 30. En effet, on a :

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2 \\ &= \overline{X^2} - \left(\overline{X}\right)^2. \end{aligned}$$

Par conséquent,

$$\hat{\sigma}^2 = \phi\left(\overline{X}, \overline{X^2}\right)$$

où

$$\phi(x, y) = y - x^2.$$

En calculant les dérivées partielles de la fonction ϕ, on obtient le résultat : voir [Vaart, 2000] page 26. □

39. [Vaart, 2000] page 26

Introduction

Par rapport à l'estimation ponctuelle, l'estimation par intervalle de confiance consiste à définir un intervalle de valeurs contenant la *vraie* valeur avec un niveau de confiance donné [40].

Définition 46

(*Intervalle de confiance*) Soit X une variable aléatoire. Soit $\{X_1, ..., X_n\}$ des variables aléatoires indépendantes de loi X. Soit I un intervalle aléatoire dépendant de $\{X_1, ..., X_n\}$. L'intervalle I est un intervalle de confiance de niveau $1 - \alpha \in [0, 1]$ pour le paramètre θ si :

$$\mathbb{P}(\theta \in I) = 1 - \alpha. \tag{30}$$

Un intervalle aléatoire est défini par ses bornes A et B c'est-à-dire $I = [A, B]$. Par conséquent, l'intervalle $[A, B]$ est un intervalle de confiance de niveau $1 - \alpha$ si :

$$\mathbb{P}(A \leq \theta \leq B) = 1 - \alpha.$$

On considère souvent le niveau $1 - \alpha = 0.95$: dans ce cas, l'intervalle de confiance contient la *vraie* valeur du paramètre dans 95% des cas (puisque I est aléatoire).

40. [Papoulis and Pillai, 2002] page 307, [Bickel and Doksum, 1977] page 155

Introduction

La définition 46 peut être associée à des ambiguïtés [41].

- Changeons la définition et remplaçons l'équation 30 par la condition :

$$\mathbb{P}(\theta \in I) \geq 1 - \alpha. \tag{31}$$

  Pour un intervalle I donné, supposons que $1 - \alpha' \leq 1 - \alpha$ où α' est un réel. Si I est un intervalle de confiance de niveau $1 - \alpha$, alors c'est également un intervalle de confiance de niveau $1 - \alpha'$. Par exemple, avec cette définition si I est un intervalle de confiance de niveau 95%, alors c'est aussi un intervalle de confiance de niveau 50%. Pour résoudre l'ambiguïté, on peut définir le niveau de confiance de l'intervalle I comme étant la plus grande valeur possible de la probabilité $\mathbb{P}(\theta \in I)$. Un autre manière consiste à imposer l'égalité, comme nous l'avons fait dans l'équation 30.
- Considérons à nouveau la définition 46. Il peut arriver que deux intervalles I et I' soient associés au même niveau de confiance $1 - \alpha$. Dans ce cas, on peut sélectionner l'intervalle de plus petite longueur. Plus précisément, si l'intervalle $I = [a, b]$ est tel que $\mathbb{P}(a \leq \theta < b) = 1 - \alpha$, alors on souhaite minimiser la longueur $b - a$ [42]. C'est le cas par exemple pour la définition de l'intervalle de confiance de la variance d'une loi gaussienne dont la moyenne est connue [43].

41. [Bickel and Doksum, 1977] page 155
42. [Papoulis and Pillai, 2002] page 305
43. [Papoulis and Pillai, 2002] page 313

Introduction

Exemple 47

(*Intervalle de confiance de la moyenne*) On considère la variable $X \sim \mathcal{N}(\mu, 1)$ de moyenne μ inconnue.
Quel est l'intervalle $[a, b]$ contenant la valeur μ avec un seuil de $1 - \alpha = 95\%$ à partir d'une seul réalisation de X ?
Loi normale est symétrique, ce qui mène à un intervalle centré sur la moyenne μ :

$$\mathbb{P}(X < b) = 0.975, \quad P(X < a) = 0.025$$

Les bornes a et b correspondent aux quantiles à 0.025 et 0.975 de X :

$$a = \phi^{-1}(0.025), \quad b = \phi^{-1}(0.975).$$

Pour une loi centrée réduite ces quantiles sont égaux à -1.960 et 1.960 d'où $a = \mu - 1.960$ et $b = \mu + 1.960$. On a :

$$\mathbb{P}(\mu - 1.960 < X < \mu + 1.960) = 0.95.$$

Donc :

$$\mathbb{P}(X - 1.960 < \mu < X + 1.960) = 0.95.$$

Statistique - Estimateur

Principe d'un intervalle de confiance [44] de niveau $1 - \alpha$.

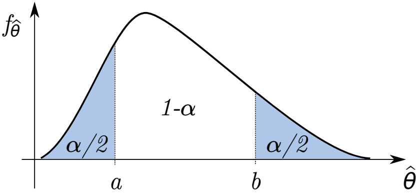


Figure 12 – Intervalle de confiance de niveau $1 - \alpha$ d'un estimateur $\hat{\theta}$.

44. [Papoulis and Pillai, 2002] p.305

Cas gaussien avec variance connue

Rappelons l'estimateur moyenne empirique

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.
Par conséquent,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Prenons comme précédemment un intervalle de confiance de seuil 5% et utilisons les quantiles de la loi normale centrée réduite. On a :

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.975, \quad \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -1.96\right) = 0.025$$

L'intervalle de confiance de la moyenne μ de niveau $1 - \alpha = 0.95$ est [45] :

$$\mathbb{P}\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

45. [Wonnacott and Wonnacott, 1977] page 201

Cas gaussien avec variance connue

On peut généraliser le résultat à un niveau de confiance $1 - \alpha$ arbitraire [46].

Théorème 48

(Intervalle de confiance de la moyenne avec variance connue) Soit $z_{1-\alpha/2} = \phi^{-1}(1 - \alpha/2)$ le quantile à $1 - \alpha/2$ de la loi normale centrée réduite. Soit $z_{\alpha/2} = -z_{1-\alpha/2}$ le quantile de niveau $\alpha/2$. Alors un intervalle de confiance de niveau $1 - \alpha$ de la moyenne $\mu = \mathbb{E}(X)$ est :

$$\mathbb{P}\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{32}$$

46. [Papoulis and Pillai, 2002] page 308, [Bickel and Doksum, 1977] page 157

Intervalle de confiance de la moyenne d'une gaussienne

Intervalle de confiance de la moyenne d'une variable gaussienne dont la variance est connue [47].

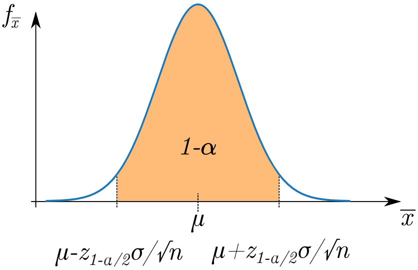


Figure 13 – Intervalle de confiance de niveau $1-\alpha$ de l'estimateur de μ par la moyenne empirique $\bar{x}$.

47. [Wonnacott and Wonnacott, 1977] p.200, [Papoulis and Pillai, 2002] p.305

Exemple

Exemple 49

(*Durée de vie d'une batterie de voiture*) On suppose que la durée de vie d'une batterie de voiture [48] est une variable aléatoire de loi gaussienne de moyenne μ inconnue et d'écart-type $\sigma = 0.5$ (ans) connu. On dispose d'un échantillon de 10 batteries dont la durée de vie a été mesurée :

$$x_1 = 4.336, \quad x_2 = 3.639, \quad x_3 = 4.494, \quad x_4 = 3.667, \quad x_5 = 4.568$$
$$x_6 = 4.051, \quad x_7 = 3.601, \quad x_8 = 4.626, \quad x_9 = 3.370, \quad x_{10} = 4.113.$$

Calculer un intervalle de confiance à 95% de la moyenne.

48. [Papoulis and Pillai, 2002] page 305

Exemple

Programme OpenTURNS/Python

```python
import openturns as ot
import numpy as np

sigma = 0.5
sample = ot.Sample([[4.336], ..., [4.113]])
x_bar = sample.computeMean()[0]
print("Sample mean = %.3f" % (x_bar))
Z = ot.Normal()
alpha = 0.05
z_alpha = Z.computeQuantile(1.0 - alpha / 2.0)[0]
print("Z(1 - alpha/2) = %.3f" % (z_alpha))
lower_bound = x_bar - z_alpha * sigma / np.sqrt(n)
upper_bound = x_bar + z_alpha * sigma / np.sqrt(n)
print(
    "Confidence interval at level %.3f = [%.3f, %.3f]"
    % (1.0 - alpha, lower_bound, upper_bound)
)
```

On observe que l'intervalle utilise la variable `sigma` puisque l'on suppose que la variance est connue.

Exemple

Le script affiche :

```
Sample mean = 4.047
alpha = 0.05
Z(1 - alpha/2) = 1.960
Confidence interval at level 0.950 = [3.737, 4.357]
```

Une option mathématiquement équivalente consiste à utiliser la méthode `computeBilateralConfidenceIntervalWithMarginalProbability` qui utilise les quantiles pour déterminer l'intervalle.

```
interval, marginalProb = Z.
    computeBilateralConfidenceIntervalWithMarginalProbability(
    1.0 - alpha)
z_min = interval.getLowerBound()[0]
z_max = interval.getUpperBound()[0]
print("z_min = %.4f" % (z_min))
print("z_max = %.4f" % (z_max))
```

Le script affiche :

```
z_min = -1.9600
z_max = 1.9600
```

Statistique - Estimateur

Exemple 50

(*Intervalle de confiance de la moyenne*) On considère la variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu = 4$ et $\sigma = 0.5$. On génère un échantillon de taille $n = 10$. On souhaite calculer un intervalle de confiance pour la moyenne μ et on fait l'hypothèse que la variance σ^2 est connue [49]. La courbe représente la densité de probabilité de la moyenne empirique $\mathcal{N}(\mu, \sigma^2/n)$. Les traits horizontaux représentent 10 intervalles à 95% de confiance.

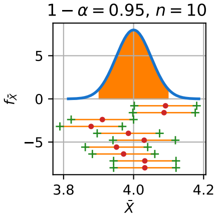


Figure 14 – Intervalle de confiance de niveau $1-\alpha$ de l'estimateur de μ par la moyenne empirique $\bar{x}$ lorsque la variance σ^2 est connue.

49. [Wonnacott and Wonnacott, 1977], p.204

L'intervalle de confiance de la moyenne a été défini en faisant deux hypothèses.

1. La loi de X est gaussienne.
2. La variance σ^2 est connue.

On peut lever l'hypothèse gaussienne en utilisant l'inégalité de Chebyshev [50].
Si X est de loi gaussienne, mais que la variance est inconnue, nous allons utiliser un théorème et deux autres lois de probabilité.

50. [Papoulis and Pillai, 2002] page 308

Loi du χ^2

La loi du χ^2 ("chi-deux") apparaît lorsqu'on considère une somme de carrés de variables gaussiennes [51].

Définition 51

(*Loi du* χ^2) Considérons n variables aléatoires $X_i \sim \mathcal{N}(0,1)$ indépendantes. Alors :

$$Q = \sum_{i=1}^{n} X_i^2$$

est une variable aléatoire du chi-deux à n degrés de liberté notée χ^2_n.

La moyenne et la variance de la variable Q sont [52] :

$$\mathbb{E}(Q) = n, \qquad \mathbb{V}(Q) = 2n.$$

51. [Ross, 2004] page 185
52. [Hoel, 1971] page 89

Loi du χ^2

La densité de probabilité de la loi du χ^2 à n degrés de liberté est [53] :

$$f_{\chi^2_n}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{n/2-1}$$

pour tout $x \geq 0$. La fonction Γ est une généralisation de la fonction factorielle pour des valeurs non entières :

$$\Gamma(x) = \int_0^x e^{-t} t^{x-1} dt$$

pour tout $x \geq 0$. On a $\Gamma(n+1) = n!$ pour tout entier n.

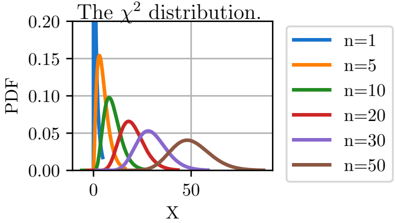


Figure 15 – Loi du χ^2_n en fonction du nombre de degrés de liberté n.

53. [Papoulis and Pillai, 2002] page 89

Loi de Student

La loi du T de Student [54] est dûe à William Gosset sous le pseudonyme de Student [55].

Définition 52

(*Loi du T de Student*) Soient $X \sim \mathcal{N}(0,1)$ et Y une variable du chi-deux à n degrés de liberté indépéndante de X. La variable

$$T = \frac{X}{\sqrt{Y/n}}$$

suit une loi de Student à n degrés de liberté, notée $\mathcal{T}_n$.

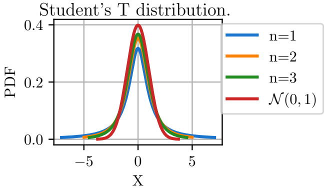


Figure 16 – Loi de Student à $n = 1, 2, 3$ degrés de liberté.

54. [Hoel, 1971] page 258, [Papoulis and Pillai, 2002] page 207
55. [Student, 1908]

Loi de Student

La loi de Student tend rapidement vers la loi normale centrée réduite.
Pour $n > 100$ elle peut-être approchée par la loi normale avec précision sur l'intervalle $[-3, 3]$.

Théorème de Fisher

Rappelons les estimateurs de la moyenne et de la variance empirique (biaisée) :

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

Le théorème suivant va permettre de définir un intervalle de confiance de la moyenne d'une loi normale[56].

Théorème 53

(Fisher) Soient $X_1, X_2, \ldots, X_n$ des variables indépendantes de loi normale $\mathcal{N}(\mu, \sigma^2)$. De plus

$$\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}.$$

Les variables aléatoires $\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sigma}$ *et* $\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$ *sont indépendantes.*
Enfin :

$$\frac{\sqrt{n-1}\left(\bar{X}-\mu\right)}{\hat{\sigma}} \sim \mathcal{T}_{n-1}.$$

56. [Bickel and Doksum, 1977] page 20, [Ross, 2004] page 217, [Saporta, 2006] page 282, [Bickel and Doksum, 1977] page 158

Cas gaussien - variance inconnue

Le théorème suivant donne l'intervalle de confiance de la moyenne d'une loi gaussienne de moyenne et de variance inconnues [57].

Théorème 54

(Intervalle de confiance de la moyenne d'une loi gaussienne de moyenne et de variance inconnues) Soient $X_1, X_2, \ldots, X_n$ *des variables indépendantes de loi normale* $\mathcal{N}(\mu, \sigma^2)$*. L'intervalle de confiance de niveau* $1-\alpha$ *est :*

$$\mathbb{P}\left(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} < \mu < \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}}\right) = 1 - \alpha$$

57. [Papoulis and Pillai, 2002] page 309, [Saporta, 2006] page 309

Cas gaussien - variance inconnue

Preuve
Par le théorème de Fisher, on a :

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n-1}} \sim \mathcal{T}_{n-1}.$$

Pour déterminer les bornes de l'intervalle de confiance, il suffit donc de remplacer les quantiles de la loi normale par ceux de la loi de Student à $n-1$ degrés de liberté. Soient $t_{\alpha/2}$ et $t_{1-\alpha/2}$ les quantiles de la loi de Student $\mathcal{T}_{n-1}$

$$t_{1-\alpha/2} = F^{-1}_{\mathcal{T}_{n-1}}\left(1 - \frac{\alpha}{2}\right), \quad t_{\alpha/2} = F^{-1}_{\mathcal{T}_{n-1}}\left(\frac{\alpha}{2}\right)$$

La loi de Student étant symétrique, les quantiles en $\alpha/2$ et $1-\alpha/2$ sont opposés :

$$t_{\alpha/2} = -t_{1-\alpha/2}.$$

ce qui mène au résultat. □

Exemple

Exemple 55

(*Intervalle de confiance de la moyenne de variance inconnue*) On suppose considère à nouveau les données de l'exemple 49, mais on fait l'hypothèse que la variance est inconnue.

Programme OpenTURNS/Python

```
import openturns as ot
import numpy as np

sample = ot.Sample([[4.336], ..., [4.113]])
x_bar = sample.computeMean()[0]
s_hat = sample.computeStandardDeviation()[0]
T = ot.Student(n - 1)
alpha = 0.05
t_alpha = T.computeQuantile(1 - alpha / 2.0)[0]
lower_bound = x_bar - t_alpha * s_hat / np.sqrt(n)
upper_bound = x_bar + t_alpha * s_hat / np.sqrt(n)
```

On observe que la variable `s_hat` contient une estimation de l'écart-type, puisqu'on suppose que la variance est inconnue.

Exemple

Le script affiche :

```
Sample mean = 4.047
Sample (unbiased) variance = 0.2068
T(1 - alpha/2) = 2.262
Confidence interval = [3.722, 4.372]
```

Lorsque l'on supposait la variance connue, nous avions obtenu l'intervalle $[3.737, 4.357]$.
On observe que l'estimation de la variance élargit un peu l'intervalle de confiance.

Cas général

Une limitation importante des intervalles de confiance précédents est que la loi de X n'est pas nécessairement connue ou pas nécessairement gaussienne.
Dans le cas général où la loi de X est inconnue, il n'y a pas de méthode pour calculer les bornes exactes de l'intervalle de confiance.
On peut toutefois avoir un intervalle de confiance asymptotique, c'est à dire lorsque n est grand [58].

Théorème 56

(Intervalle de confiance asymptotique de l'espérance) Soit X une variable aléatoire. Soit $\{x_1, ..., x_n\}$ un échantillon de réalisations indépendantes de même loi que X. Alors, l'intervalle suivant est un intervalle de confiance asymptotique de l'espérance :

$$\mathbb{P}\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}}\right) \rightarrow 1 - \alpha$$

quand $n \rightarrow \infty$.

58. [Delmas, 2013] page 151

Exemple

Preuve

Grâce au théorème central limite (TCL), et au théorème de Slutsky, on sait que l'estimateur $\bar{X}$ tend vers une loi normale :

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\hat{s}} \xrightarrow{d} \mathcal{N}(0, 1)$$

ce qui conclut la preuve. $\square$

Exemple

Exemple 57

(*Intervalle de confiance de la moyenne d'une loi log-normale*) On considère la variable aléatoire X de loi log-normale de paramètres $\mu = 2$ et $\sigma = 1$ où μ et σ sont les paramètres de la loi normale sous-jacente. En d'autres termes, on a $X = \exp(Z)$ où $Z \sim \mathcal{N}(\mu, \sigma^2)$. On génère un échantillon de taille $n = 100$ de réalisations indépendantes et de même loi que X. On souhaite estimer un intervalle de confiance de μ.

Exemple

Programme OpenTURNS/Python

```python
import openturns as ot
import numpy as np

mu = 2.0
sigma = 1.0
X = ot.LogNormal(mu, sigma)
n = 100
s = X.getSample(n)
x_bar = s.computeMean()[0]
s_hat = s.computeStandardDeviationPerComponent()[0]
Z = ot.Normal()
alpha = 0.05
z_alpha = Z.computeQuantile(1 - alpha / 2.0)[0]
lower_bound = x_bar - z_alpha * s_hat / np.sqrt(n)
upper_bound = x_bar + z_alpha * s_hat / np.sqrt(n)
```

Exemple

Résultat sur la console

```
Exact mean = 12.18
n = 100
Estimated mean = 12.55
alpha = 0.05
T(1 - alpha/2) = 1.984
Confidence interval = [9.905,15.19]
```

Exemple 58

(*Intervalle de confiance de la moyenne d'une variable log-normale*) On considère une v.a. X de loi Log-Normale, dont la gaussienne sous-jacente est $Z = \mathcal{N}(\mu, \sigma)$ avec $\mu = 2$ et $\sigma = 1$. L'espérance de X est $\mathbb{E}(X) = 12.18$. On considère un échantillon de taille $n = 100$ et on estime l'intervalle de confiance en utilisant la loi asymptotique gaussienne. La figure suivante présente 10 intervalles de confiance de niveau 95% de la moyenne de X.

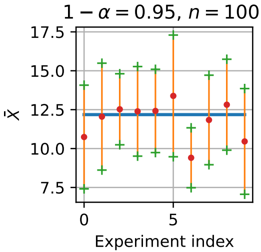


Figure 17 – Plusieurs intervalles de confiance de niveau 95% de la moyenne d'une variable log-normale sur un échantillon de taille $n = 100$. Le trait horizontal est la vraie moyenne.

Intervalle de confiance de l'estimateur de la variance d'une loi normale

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$.
Soit $x_1, \ldots, x_n$ des réalisations i.i.d. de X.
On souhaite estimer σ.
On considère deux cas :

1. la moyenne μ est connue ;
2. la moyenne μ est inconnue.

Moyenne connue

Cas 1) : supposons que la moyenne de la variable aléatoire gaussienne μ est connue.
On considère l'estimateur

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

L'estimateur $\hat{\kappa}^2$ est un estimateur de la variance (i.e. du moment d'ordre 2 centré).
L'estimateur $\hat{\kappa}^2$ n'est *pas* égal à l'estimateur de la variance empirique car on utilise μ au lieu de $\bar{X}$.

Intervalle de confiance de la variance d'une loi normale de moyenne connue

Le théorème suivant exploite la connaissance de la moyenne μ de la loi gaussienne pour déterminer un intervalle de confiance de la variance[59].

Théorème 59

(Intervalle de confiance de la variance d'une loi normale de moyenne connue) Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. Soit $\{x_1, ..., x_n\}$ un échantillon i.i.d. de X. On suppose que μ est connu et que σ^2 est inconnu. Soit $\alpha \in [0, 1]$ un paramètre réel. Soient $k_{\frac{\alpha}{2}}$ et $k_{1-\frac{\alpha}{2}}$ les quantiles de niveaux $\alpha/2$ et $1 - \alpha/2$ de la loi χ^2_n :

$$k_{\frac{\alpha}{2}} = F^{-1}_{\chi^2_n}\left(\frac{\alpha}{2}\right), \qquad k_{1-\frac{\alpha}{2}} = F^{-1}_{\chi^2_n}\left(1 - \frac{\alpha}{2}\right).$$

Alors, l'intervalle suivant est un intervalle de confiance de la variance de niveau $1 - \alpha$:

$$\mathbb{P}\left[\frac{n\hat{\kappa}^2}{k_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{n\hat{\kappa}^2}{k_{\frac{\alpha}{2}}}\right] = 1 - \alpha. \tag{33}$$

59. [Papoulis and Pillai, 2002] page 313, [Saporta, 2006] page 309

Preuve
La variable $n\hat{\kappa}^2/\sigma^2$ suit une loi χ^2_n car c'est une somme de n carrés de $\mathcal{N}(0,1)$
Soient $k_{\frac{\alpha}{2}}, k_{1-\frac{\alpha}{2}}$ les quantiles de niveaux $\alpha/2$ et $1-\alpha/2$ de la loi χ^2_n. L'intervalle de confiance de niveau $1-\alpha$ est :

$$\mathbb{P}\left[k_{\frac{\alpha}{2}} < \frac{n\hat{\kappa}^2}{\sigma^2} < k_{1-\frac{\alpha}{2}}\right] = 1-\alpha$$

ce qui mène au résultat. □

Remark 5

(Comportement asymptotique) Les bornes de l'intervalle défini dans l'équation 33 peuvent sembler diverger lorsque $n \to +\infty$, à cause du facteur n au numérateur. Il n'en est rien, ce que le l'on peut voir en observant que, si $X \sim \chi_n^2$, alors $X \to \mathcal{N}(n, 2n)$ lorsque $n \to +\infty$. Par conséquent, on a

$$\frac{n\hat{\kappa}^2}{\sigma^2} \to \mathcal{N}(n, 2n).$$

Cela implique :

$$\hat{\kappa}^2 \to \frac{\sigma^2}{n}\mathcal{N}(n, 2n) = \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right).$$

La convergence est donc une correction multiplicative de σ^2 par le terme $\sqrt{2/n}$. Ce résultat correspond au résultat du théorème 44 page 67.

Moyenne connue

Exemple 60

(*Intervalle de confiance de la variance d'une gaussienne de moyenne connue*) On considère la variable aléatoire X gaussienne de moyenne μ et on souhaite déterminer un intervalle de confiance de la variance.
On considère un échantillon de taille $n = 4$, ce qui mène à la loi du χ_4^2, c'est-à-dire à 4 degrés de liberté.
On fixe le niveau de confiance $1 - \alpha = 0.95$ associé à $\alpha = 0.05$.
Les quantiles et valeurs de densité de probabilité associées sont :

$$k_{\alpha/2} = 0.831, \quad f(k_{\alpha/2}) = 0.0665,$$
$$k_{1-\alpha/2} = 12.833, \quad f(k_{1-\alpha/2}) = 0.0100$$

La longueur de l'intervalle de confiance fondé sur les quantiles est :

$$\ell_{1-\alpha} = k_{1-\alpha/2} - k_{\alpha/2} = 12.00.$$

Moyenne connue

Recherchons un intervalle de longueur minimale, fondé sur les iso-valeurs de la densité de probabilité.
Pour tout réel $y_{1-\alpha} > 0$, soit $\mathcal{A}_{1-\alpha}$ l'ensemble des points dont la densité est supérieure à $y_{1-\alpha}$:

$$\mathcal{A}_{1-\alpha} = \left\{x \in \mathbb{R} \mid f_{\chi^2_n}(x) \geq y_{1-\alpha}.\right\}$$

On recherche $y_{1-\alpha}$ tel que :

$$\int_{\mathcal{A}_{1-\alpha}} f_{\chi^2_n}(x) dx = 1 - \alpha.$$

Pour $1 - \alpha = 0.95$, la solution de l'équation est $y_{1-\alpha} = 0.0185$ et on obtient :

$$a_{1-\alpha} = 0.296, \qquad b_{1-\alpha} = 11.191.$$

Ces deux observations sont toutes deux associées à la même valeur de densité de probabilité $y_{1-\alpha} = 0.0185$. La longueur de l'intervalle de confiance fondé sur les quantiles est :

$$\ell^\star_{1-\alpha} = b_{1-\alpha} - a_{1-\alpha} = 10.90.$$

On observe que le second intervalle possède une longueur strictement inférieure à l'intervalle précédent.

Moyenne connue

La figure 18 présente l'intervalle de confiance de la variance d'une variable gaussienne dont la moyenne est connue [60].
L'intervalle de confiance fondé sur les quantiles est de longueur strictement supérieure à l'intervalle fondé sur les iso-valeurs de densité de probabilité.

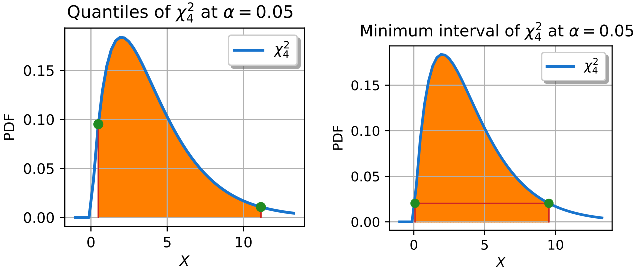


Figure 18 – A gauche, l'intervalle de confiance asymétrique de niveau $1-\alpha$ de la variance fondé sur les quantiles de la loi χ^2_n. A droite, l'intervalle de longueur minimale de niveau $1-\alpha$ de la variance.

60. [Papoulis and Pillai, 2002], p.313

Statistique - Estimateur

Lorsque n est grand, la loi du χ^2 est plus proche d'une loi gaussienne ce qui montre que les deux intervalles sont plus proches lorsque la taille de l'échantillon est suffisamment grande (pour $1 - \alpha = 0.95$).
Généraliser cette idée à une dimension supérieure mène à la notion de "minimum volume level set" [61].

61. [Nuñez Garcia et al., 2003]

Moyenne inconnue

Cas 2) : supposons que la moyenne de la variable aléatoire gaussienne μ est inconnue.
Puisque la moyenne μ est inconnue, on l'estime par $\bar{X}$.
On considère l'estimateur biaisé de la variance :

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

Moyenne inconnue

Le théorème suivant exploite estime la moyenne μ de la loi gaussienne pour déterminer un intervalle de confiance de la variance [62].

Théorème 61

(Intervalle de confiance de la variance d'une loi normale de moyenne inconnue) Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. Soit $\{x_1, ..., x_n\}$ un échantillon i.i.d. de X. On suppose que μ et σ sont inconnus. L'estimation de μ est réalisée par la moyenne empirique $\bar{x}$. Soit $\alpha \in [0, 1]$ un paramètre réel. Soient $k_{\frac{\alpha}{2}}$ et $k_{1-\frac{\alpha}{2}}$ les quantiles de niveaux $\alpha/2$ et $1 - \alpha/2$ de la loi χ^2_{n-1} :

$$k_{\frac{\alpha}{2}} = F^{-1}_{\chi^2_{n-1}}\left(\frac{\alpha}{2}\right), \qquad k_{1-\frac{\alpha}{2}} = F^{-1}_{\chi^2_{n-1}}\left(1 - \frac{\alpha}{2}\right).$$

Alors, l'intervalle suivant est un intervalle de confiance de la variance de niveau $1 - \alpha$:

$$\mathbb{P}\left[\frac{n\hat{\sigma}^2}{k_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{n\hat{\sigma}^2}{k_{\frac{\alpha}{2}}}\right] = 1 - \alpha \tag{34}$$

62. [Papoulis and Pillai, 2002] page 314, [Saporta, 2006] page 310, [Bickel and Doksum, 1977] page 159, [Ross, 2004] page 251

Moyenne inconnue

Preuve
D'après le théorème 53, la variable $n\hat{\sigma}^2/\sigma^2$ suit une loi χ^2_{n-1}. $\square$
Les calcul des bornes des intervalles via les quantiles ont été possibles parce que nous avons fait l'hypothèse que X suit la loi normale.
Le nombre de degrés de liberté est égal à $n-1$, au lieu de n précédemment.
Dans le cas général où X n'est pas nécessairement gaussien, on peut exploiter le théorème asymptotique 45 page 69.
Il est toutefois nécessaire de disposer d'un estimateur du moment d'ordre 4, et de disposer d'un échantillon de taille suffisamment grande pour que l'approximation asymptotique soit de précision acceptable.

Exemple

On considère un exemple adapté d'une application électronique [63].

Exemple 62

(*Variance d'une source de tension*) A voltage source V is measured six times :

$$x_1 = 109.937, \quad x_2 = 110.296, \quad x_3 = 110.760, \quad x_4 = 110.148,$$
$$x_5 = 109.135, \quad x_6 = 109.749.$$

The measurements are modeled by the random variable $X = v + \nu$ where v is a constant parameter and $\nu \sim \mathcal{N}(0, \sigma^2)$. Compute a 95% confidence interval of the variance σ^2.

1. Suppose first that the source is a known standard with $v = 110$V.
2. Suppose now that v is unknown.

63. [Papoulis and Pillai, 2002] page 315

Exemple

Programme OpenTURNS/Python.
Partie 1 : la moyenne μ est connue.

```python
import openturns as ot
import numpy as np
mu = 110.0
sample = ot.Sample.BuildFromPoint(
    [109.937, 110.296, 110.76, 110.148, 109.135, 109.749 ]
)
n = sample.getSize()
# Part 1 : the mean is known
n = sample.getSize()
delta = sample - mu
delta_point = delta.asPoint()
k_squared = delta_point.normSquare() / n
print("K^2 = %.3f" % (k_squared))
distribution = ot.ChiSquare(n)
alpha = 0.05
k_min = distribution.computeQuantile(alpha / 2.0)[0]
k_max = distribution.computeQuantile(1.0 - alpha / 2.0)[0]
print("Chi(alpha/2) = %.3f" % (k_min))
print("Chi(1 - alpha/2) = %.3f" % (k_max))
lower_bound = n * k_squared / k_max
upper_bound = n * k_squared / k_min
```

Exemple

```python
print("Confidence interval of sigma^2 at level %.3f = [%.3f, %.3f]"
    % (1.0 - alpha, lower_bound, upper_bound)
)
print("Corresponding interval of sigma = [%.3f, %.3f]"
    % (np.sqrt(lower_bound), np.sqrt(upper_bound))
)
```

Le script précédent produit la sortie :

```
K^2 = 0.250
alpha = 0.05
Chi(alpha/2) = 1.237
Chi(1 - alpha/2) = 14.449
Confidence interval of sigma^2 at level 0.950 = [0.104, 1.214]
Corresponding interval of sigma = [0.322, 1.102]
```

Exemple

Programme OpenTURNS/Python.
Partie 2 : la moyenne μ est inconnue.

```
print("+ Part 2 : the mean is unknown")
s_squared_hat = sample.computeCovariance()[0, 0]
print("Sample (unbiased) variance = %.4f" % (s_squared_hat))
distribution = ot.ChiSquare(n - 1)
alpha = 0.05
print("alpha =", alpha)
k_min = distribution.computeQuantile(alpha / 2.0)[0]
k_max = distribution.computeQuantile(1.0 - alpha / 2.0)[0]
print("Chi(alpha/2) = %.3f" % (k_min))
print("Chi(1 - alpha/2) = %.3f" % (k_max))
lower_bound = (n - 1) * s_squared_hat / k_max
upper_bound = (n - 1) * s_squared_hat / k_min
print("Confidence interval = [%.3f, %.3f]" % (lower_bound,
    upper_bound))
print(
    "Corresponding interval of sigma = [%.3f, %.3f]"
    % (np.sqrt(lower_bound), np.sqrt(upper_bound))
)
```

Exemple

Partie 2 : la moyenne μ est inconnue.
On obtient :

```
Sample (unbiased) variance = 0.3004
alpha = 0.05
Chi(alpha/2) = 0.831
Chi(1 - alpha/2) = 12.833
Confidence interval = [0.117, 1.807]
Corresponding interval of sigma = [0.342, 1.344]
```

Cas général

Le cas général où la variable aléatoire X n'est pas nécessairement gaussienne peut être traité par le théorème asymptotique 45 page 69, à condition de disposer d'un estimateur convergent du moment centré d'ordre 4 et un échantillon de taille suffisamment grande.

Confidence intervals from Chebyshev

Exercice 1

(*Intervalle de confiance de la moyenne avec l'inégalité de Chebyshev*)
Assume that X is a random variable with finite expectation $\mu = \mathbb{E}(X)$ and finite variance $\sigma^2 = \mathbb{V}(X)$. We assume that μ is unknown and that σ is known. We want to derive a confidence interval of the sample mean $\bar{X} = X_1 + ... + X_n$, where n is the sample size and $X_1, ..., X_n$ are independent random variables with the same distribution as X.
Apply Chebyshev inequality (theorem 21 page 31) to the sample mean and derive a confidence interval with level $1 - \alpha$.
Compare the numerical values of the factors involved in this confidence interval with the one obtained in equation 32, p. 76, for $\alpha = 0.05$.
In which circumstances is Chebyshev's bound interesting ?

Théorème central limite

Exercice 2

(*Python / OpenTURNS : Théorème central limite*)
Rédiger un script Python/OpenTURNS qui montre la convergence de la variable

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

où X est une variable aléatoire de loi exponentielle de paramètre $\lambda = 3.215$.
Pour cela, on considère une taille d'échantillon égale à $n = 32$. On répètera l'expérience $r = 1000$ fois, ce qui mène à un échantillon de taille r de réalisations de la moyenne empirique $\bar{X}$. On approchera la densité de probabilité de la variable aléatoire $\bar{X}$ par l'histogramme empirique.

Convergence de l'estimateur de la variance empirique non biaisé

Exercice 3

Démontrer le théorème suivant.

Théorème 63

(Convergence of the sample (unbiased) variance) Let X be a random variable with mean $\mathbb{E}(X) = \mu$, finite variance $\mathbb{V}(X) = \sigma^2$ and finite order 4 moment $\mu_4 = \mathbb{E}((X-\mu)^4)$. Let $X_1, \ldots, X_n$ be independent realizations of the random variable X. Then

$$\hat{s}^2 \xrightarrow{\mathbb{P}} \sigma^2.$$

Indications. Utiliser l'inégalité de Chebyshev.

Convergence de l'estimateur de la variance empirique non biaisé

Exercice 4

Démontrer le théorème suivant.

Théorème 64

(Asymptotic distribution of the mean of a random variable) Let X be a random variable with mean $\mathbb{E}(X) = \mu$, finite variance $\mathbb{V}(X) = \sigma^2$ and finite order 4 moment. Let $X_1, \ldots, X_n$ be independent random variables with the same distribution as X. Therefore,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\hat{s}} \xrightarrow{d} \mathcal{N}(0, 1). \tag{35}$$

Indications. Démontrer que :

$$\frac{\sigma}{\hat{s}} \xrightarrow{\mathbb{P}} 1.$$

en utilisant le théorème de continuité puis appliquer le théorème de Slutsky.

Convergence de l'estimateur de la variance empirique biaisé

Exercice 5

(*Convergence of the sample (biased) variance*)
Démontrer le théorème suivant.

Théorème 65

(Convergence of the sample (biased) variance) Let X be a random variable with mean $\mathbb{E}(X) = \mu$, finite variance $\mathbb{V}(X) = \sigma^2$ and finite order 4 moment $\mu_4 = \mathbb{E}((X-\mu)^4)$. Let $X_1, \ldots, X_n$ be independent realizations of the random variable X. Then

$$\hat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2.$$

Indications. Utiliser l'inégalité de Chebyshev.

Distribution de l'estimateur de la variance empirique biaisé

Exercice 6

(*Python : Distribution de l'estimateur biaisé de la variance*)
L'estimateur de la variance de la population

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

est biaisé. L'estimateur

$$\hat{s}^2 = \frac{1}{n-1} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

est non biaisé.
L'objectif de cet exercice est de montrer la distribution empirique de ces deux estimateurs grâce à une simulation en Python avec OpenTURNS.
On considère la distribution gaussienne de moyenne nulle et de variance $\sigma^2 = 1$. On considère la taille de l'échantillon égale à $n = 10$. Générer un échantillon de taille n et utiliser l'estimateur biaisé de la variance. Répéter l'opération $r = 1000$ fois et stocker les estimations dans un tableau. Estimer la densité de probabilité de l'estimateur, par exemple grâce à un histogramme (ou une méthode de lissage à noyau). Qu'observez-vous ?
Répéter l'opération pour l'estimateur non biaisé.

Distribution de l'estimateur de la médiane empirique

Exercice 7

(*Python : Distribution de l'estimateur de la médiane empirique*)
Soient $x_1, \ldots, x_n$ des réalisations indépendantes de la variable X. L'estimateur empirique de la moyenne est :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Sa variance est

$$\mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}.$$

La médiane d'une variable aléatoire est son quantile de niveau 0.5, i.e. $m = F^{-1}(0.5)$ où F est la fonction de répartition de X. Pour évaluer la médiane empirique, on ordonne les réalisations par ordre croissant.

Distribution de l'estimateur de la médiane empirique - suite

Pour $i = 1, ..., n$, soit $x_{(i)}$ la i-ème réalisation dans l'échantillon trié par ordre croissant :

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

La médiane empirique est

$$\tilde{X} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} \text{ si } n \text{ est impair,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \text{ si } n \text{ est pair.} \end{cases}$$

On peut démontrer que la médiane empirique est un estimateur sans biais de μ. Sa variance asymptotique est ([Greene, 2012] pages 656 et 1126)

$$\mathbb{V}\left(\tilde{X}\right) \rightarrow \frac{\pi}{2}\frac{\sigma^2}{n}$$

quand $n \rightarrow +\infty$.
L'objectif de cet exercice est de montrer la distribution empirique de ces deux estimateurs grâce à une simulation en Python avec OpenTURNS.
Utiliser la même méthode que celle présentée dans l'exercice 6.

Estimation de la moyenne de la loi de Cauchy

Exercice 8

(*Python : Loi de Cauchy*)
La loi de Cauchy a la densité de probabilité f définie par [64] :

$$f(x) = \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2}$$

où x_0 est un paramètre de position et γ est un paramètre d'échelle.
La moyenne et la variance de la loi de Cauchy sont indéfinies.
La loi de Cauchy n'est pas disponible dans le catalogue des distributions de PERSALYS. En revanche, on peut obtenir une loi de Cauchy en utilisant une loi de Student associée à df $= 1$ degré de liberté, de moyenne x_0 et d'écart-type γ.
Malheureusement, on ne peut pas utiliser la loi de Student avec 1 degré de liberté dans PERSALYS (le vérifier). Dans les deux exercices qui suivent, on utilise un contournement pour utiliser tout de même cette distribution.

64. [Papoulis and Pillai, 2002] page 92, [Wasserman, 2006] page 30

Estimation de la moyenne de la loi de Cauchy - suite

Exercice 8, A
Considérer le modèle g défini par :

$$Y = \frac{X_1}{X_2}$$

où

$$X_i \sim \mathcal{N}(0, 1)$$

pour $i = 1, 2$.
Utiliser la méthode de Monte-Carlo simple et essayer d'estimer la moyenne et la variance. Qu'observez-vous ?

Estimation de la moyenne de la loi de Cauchy - suite

Exercice 8, B
Considérer le modèle g défini par :

$$Y = \tan\left(\pi\left(X_1 - \frac{1}{2}\right)\right)$$

où

$$X_1 \sim \mathcal{U}(0, 1).$$

Utiliser la méthode de Monte-Carlo simple et essayer d'estimer la moyenne et la variance. Qu'observez-vous ?

Estimation de la moyenne de la loi de Cauchy - suite

Exercice 8, C
On considère deux variables aléatoires X_1 et X_2 définies par :

$$X_i \sim \mathcal{N}(0,1)$$

pour $i = 1, 2$.
Considérer le modèle g défini par :

$$Y = \frac{M}{S/\sqrt{2}}$$

où

$$M = \frac{X_1 + X_2}{2}, \quad S = (X_1^2 - M) + (X_2^2 - M).$$

Utiliser la méthode de Monte-Carlo simple et essayer d'estimer la moyenne et la variance. Qu'observez-vous ?

Références I

Ang, Alfredo, H.-S. and Tang, W. H. (1975).
Probability concepts in engineering planning and design. Volume 1 - Basic principles.
John Wiley and Sons.

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

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

Conover, W. J. (1980).
Practical nonparametric statistics. 2d Edition.
John Wiley & Sons.

Delmas, J.-F. (2013).
Introduction au calcul des probabilités et à la statistique : exercices, problèmes et corrections (2e édition).
Les Presses de l'ENSTA.

Références II

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

Guyader, A. (2016).
Statistique mathématique. année 2015-2016, second semestre.
Université Pierre et Marie Curie. Master Mathématiques et Applications.

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

Nuñez Garcia, J., Kutalik, Z., Cho, K.-H., and Wolkenhauer, O. (2003).
Level sets and minimum volume sets of probability density functions.
International Journal of Approximate Reasoning, 34(1) :25–47.

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

Références III

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.

Student (1908).
The probable error of a mean.
Biometrika, 6(1) :1–25.

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

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

Références IV

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