Statistique Mathématique

Salim Lardjane

Université Bretagne Sud

Partie II
Estimation Statistique - Généralités

Estimateur

On appelle *estimateur* d'une quantité θ toute variable aléatoire dont les valeurs sont utilisées comme approximation de θ.

Ces dernières sont alors appelées *estimations* de θ.

Un estimateur T d'une quantité θ est dit *sans biais* si

$$\mathbb{E}(T) = \theta.$$

Moyenne arithmétique

Soient $X_1, \ldots, X_n$ des v.a.r. indépendantes et identiquement distribuées (ont note *i.i.d.*).

La *moyenne arithmétique* ou, plus simplement, *la moyenne*, des X_i est définie par :

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

On dit également que c'est la moyenne arithmétique de *l'échantillon* i.i.d. $X_1, \ldots, X_n$.

Variance empirique

La *variance empirique* ou, plus simplement, *la variance*, de l'échantillon est définie par :

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

La moyenne arithmétique comme estimateur de l'espérance

Si les X_i sont d'espérance μ et de variance théorique σ^2, alors :

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} \cdot n\mu = \mu$$

La moyenne arithmétique est donc un *estimateur sans biais* de l'espérance μ.

La moyenne arithmétique comme estimateur de l'espérance

D'autre part, on a :

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

Donc, $V(\bar{X}) \to 0$ lorsque $n \to \infty$.

Ainsi, la moyenne arithmétique est un *bon estimateur* de μ : pour n grand, la variance de $\bar{X}$ est petite et $\bar{X}$ est concentrée autour de son espérance, qui vaut μ.

La variance empirique comme estimateur de la variance

La variance empirique, quant à elle, est un *estimateur biaisé* de la variance σ^2.

En effet, on peut écrire :

$$\begin{aligned}
\mathbb{E}[(X_i - \bar{X})^2] &= \mathbb{E}[(X_i - \mu) - (\bar{X} - \mu)]^2 \\
&= V(X_i) + V(\bar{X}) \\
&\quad -2\mathbb{E}[(X_i - \mu)(\bar{X} - \mu)] \\
&= \sigma^2 + \frac{\sigma^2}{n} \\
&\quad -2\mathbb{E}\left[(X_i - \mu)\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)\right] \\
&= \sigma^2 + \frac{\sigma^2}{n} - \frac{2}{n}\mathbb{E}[(X_i - \mu)^2] \\
&= \sigma^2(1 + \frac{1}{n} - \frac{2}{n}) \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

La variance empirique comme estimateur de la variance

Par conséquent :

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i-\bar{X})^2] \\ &= \frac{n-1}{n}\,\sigma^1 \neq \sigma^2\end{aligned}$$

C'est pourquoi on définit parfois la variance de l'échantillon comme :

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

Cet estimateur de la variance σ^2 est *sans biais* :

$$\mathbb{E}\left(\frac{n}{n-1}\,S^2\right) = \sigma^2.$$

La variance empirique comme estimateur de la variance

Toutefois, *biaisé ne signifie pas nécessairement mauvais*.

Si on mesure la performance d'un estimateur T d'une quantité θ à l'aide de *l'erreur quadratique moyenne* :

$$\text{EQM}(T, \theta) = \mathbb{E}[(T - \theta)^2],$$

l'estimateur biaisé de la variance est *meilleur dans le cas gaussien (normal)*, comme on va le voir dans la suite.

La variance empirique comme estimateur de la variance : cas gaussien

A présent, supposons que les X_i, en plus d'être i.i.d., sont normalement distribuées.

On dit dans ce cas qu'on a un *échantillon gaussien*.

On peut obtenir dans ce cas la *loi de* S^2.

En effet, on a, pour un échantillon gaussien :

$$\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

La variance empirique comme estimateur de la variance : cas gaussien

Ceci peut être démontré via le résultat intermédiaire suivant.

Théorème : Pour un échantillon i.i.d. gaussien, $\bar{X}$ et S^2 sont *indépendants*.

Preuve. Celle-ci fait appel au résultat général sur la loi de la transformée d'une variable aléatoire (utilisation du Jacobien).

La variance empirique comme estimateur de la variance : cas gaussien

Remarque. La variable aléatoire

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n-1}}$$

est utile dans les situations où μ doit être estimée mais où la variance théorique σ^2 est inconnue.

D'après les résultats vus dans la première partie du cours, on voit que T suit une *loi de Student* à $n-1$ degrés de liberté.

La variance empirique comme estimateur de la variance : cas gaussien

Considérons à présent, dans le cas gaussien, l'ensemble des estimateurs de la forme cS^2 où $c \in \mathbb{R}$.

On peut écrire :

$$\begin{aligned} \text{EQM}(cS^2, \sigma^2) &= \mathbb{E}[(cS^2 - \sigma^2)^2] \\ &= \mathbb{E}[(cS^2 - c\mathbb{E}(S^2) \\ &\quad + c\mathbb{E}(S^2) - \sigma^2)^2] \\ &= c^2 V(S^2) + (c\mathbb{E}(S^2) - \sigma^2)^2 \end{aligned}$$

La variance empirique comme estimateur de la variance : cas gaussien

Comme

$$\frac{nS^2}{\sigma^2} \sim \chi^2(n-1),$$

qui est de variance $2(n-1)$, on a :

$$\frac{n^2 V(S^2)}{\sigma^4} = 2(n-1)$$

d'où

$$V(S^2) = \frac{2\sigma^4(n-1)}{n^2}$$

et d'autre part :

$$\mathbb{E}\left(\frac{nS^2}{\sigma^2}\right) = \frac{n\mathbb{E}(S^2)}{\sigma^2} = n-1$$

La variance empirique comme estimateur de la variance : cas gaussien

On en déduit :

$$\begin{aligned} \text{EQM}(cS^2, \sigma^2) &= \frac{2c^2\sigma^4(n-1)}{n^2} \\ &\quad + \left(\frac{c(n-1)}{n} - 2\sigma^2\right)^2 \\ &= \frac{\sigma^4}{n^2} \cdot K(c) \end{aligned}$$

où

$$K(c) = cc^2(n-1) + c^2(n-1)^2 - 2c(n-1)n + n^2$$

Nous allons à présent chercher le minimum de $K(c)$, ce qui permettra de déterminer l'estimateur de σ^2 de la forme cS^2 d'erreur quadratique moyenne minimale.

La variance empirique comme estimateur de la variance : cas gaussien

Ecrivons la condition au premier ordre :

$$\begin{aligned}\frac{dK(c)}{dc} &= 4c(n-1) + 2c(n-1)^2 - 2(n-1)n \\ &= 0\end{aligned}$$

En divisant par $2(n-1)$, on obtient

$$2c + c(n-1) - n = 0$$

d'où

$$c = \frac{n}{n+1}.$$

De plus,

$$\frac{d^2K(c)}{dc^2} = 4(n-1) + 2(n-1)^2 > 0$$

Il s'agit donc bien d'un minimum.

La variance empirique comme estimateur de la variance : cas gaussien

Ainsi, le meilleur estimateur de σ^2 de la forme cS^2 pour un échantillon i.i.d. gaussien est :

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

Si on utilise S^2, alors $c = 1$ et si on utilise l'estimateur sans biais $\frac{n}{n-1}S^2$, alors $c = n/(n-1)$.

Comme on a :

$$\frac{n}{n+1} < 1 < \frac{n}{n-1}$$

et qu'une fonction quadratique décroît toujours quand on se déplace vers le minimum, on voit que l'estimateur biaisé S^2 est meilleur que l'estimateur non biaisé $\frac{n}{n-1}S^2$, mais qu'aucun des deux n'est optimal pour le critère de l'erreur quadratique moyenne.

La variance empirique comme estimateur de la variance : cas gaussien

De façon explicite :

$$\text{EQM}\left(\frac{nS^2}{n-1}, \sigma^2\right) = \frac{2\sigma^4}{n-1}$$

et

$$\text{EQM}(S^2, \sigma^2) = \frac{(2n-1)\sigma^4}{n^2}$$

et on a toujours

$$\text{EQM}(S^2, \sigma^2) < \text{EQM}\left(\frac{nS^2}{n-1}, \sigma^2\right)$$

dès lors que $n > 1$ (le vérifier à titre d'exercice).

La variance empirique comme estimateur de la variance : cas gaussien

Lorsque n *est grand*, les trois estimateurs précédents sont tous bons et la différence entre leurs performances sont négligeables.

Consistance d'un estimateur

Une suite d'estimateurs T_n d'une quantité θ est dite *consistante* ou *convergente en probabilité* si :

$$T_n \to_{\mathbb{P}} \theta \quad (n \to \infty)$$

Cela signifie que T_n est très probablement proche de θ lorsque la taille d'échantillon n est grande.

Moyenne arithmétique : consistance

A présent, soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires indépendantes identiquement distribuées, de mêmes espérance et variance μ et σ^2.

Notons $\bar{X}_n$ la moyenne arithmétique de $X_1, \ldots, X_n$.

Alors, d'après la Loi Faible des Grands Nombre, on a :

$$\bar{X}_n \to_{\mathbb{P}} \mu \quad (n \to \infty)$$

Ainsi, la moyenne arithmétique est un estimateur consistant de l'espérance.

Variance empirique : consistance

Soit $X_1, X_2, \ldots, X_n, \ldots$ une suite de v.a.r. **gaussiennes** indépendantes et identiquement distribuées et soit S_n^2 la variance empirique de l'échantillon $X_1, \ldots, X_n$.

On va montrer que S_n^2 est un estimateur consistant de la variance théorique σ^2, c'est-à-dire que :

$$S_n^2 \to_{\mathbb{P}} \sigma^2 \quad (n \to \infty)$$

Variance empirique : consistance

A cet effet, notons que, comme l'échantillon est gaussien, on a :

$$\frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1)$$

d'où

$$\mathbb{E}\left(\frac{nS_n^2}{\sigma^2}\right) = n-1$$

et

$$V\left(\frac{nS_n^2}{\sigma^2}\right) = 2(n-1).$$

Variance empirique : consistance

On en déduit que :

$$\mathbb{E}(S_n^2) = \frac{(n-1)\sigma^2}{n} \to \sigma^2 \quad (n \to \infty)$$

et

$$V(S_n^2) = \frac{2(n-1)\sigma^4}{n^2} \to 0 \quad (n \to \infty),$$

d'où la convergence en probabilité de S_n^2 vers σ^2 d'après un résultat vu dans la première partie du cours (p. 29).

Moyenne arithmétique : loi asymptotique

Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires indépendantes identiquement distribuées, de mêmes espérance et variance μ et σ^2.

Notons $\bar{X}_n$ la moyenne arithmétique de $X_1, \ldots, X_n$.

Alors, d'après le Théorème Central Limite (p. 30), on a :

$$\sqrt{n}(\bar{X}_n - \mu) \to_L \mathcal{N}(0, \sigma^2) \quad (n \to \infty)$$

Estimation : cas général

On peut considérer dans le cas général qu'on joue une partie contre la Nature.

Celle-ci choisit un état θ (usuellement un réel ou un vecteur réel) et effectue une expérience aléatoire.

On ne connaît pas θ mais on observe la valeur d'une variable (ou d'un vecteur) aléatoire X, appelée *l'observation*, dont la loi dépend de θ.

Cette loi est généralement caractérisée par sa densité $f_\theta(x)$.

Estimation : cas général

Après avoir observé une *réalisation* x de X, on estime θ par $\delta(x)$, où δ est une fonction bien choisie.

La quantité obtenue $\delta(x)$ est appelée *estimation ponctuelle* de θ car elle consiste en un nombre (ou vecteur) unique qu'on espère proche de θ.

La principale alternative à cette approche est *l'estimation par intervalle de confiance*, qu'on verra dans la suite.

Estimation : cas général

Notons que, pour qu'une estimation ponctuelle $\delta(x)$ ait un sens, elle ne doit dépendre que de x et pas du paramètre inconnu θ.

La **variable aléatoire** $\delta(X)$ est un *estimateur* de θ.

Il existe plusieurs estimateurs possibles d'un paramètre inconnu θ et *il n'existe pas de règle systématique* pour en chosir un.

Estimation : cas général

On se base souvent sur des considérations pratiques du type :

(a) Quel est le coût de la collecte des données ?

(b) La performance de l'estimateur est-elle facile à quantifier ? Par exemple peut-on calculer $\mathbb{P}(|\delta(X) - \theta| \leq \varepsilon)$ à ε fixé ?

(c) Les avantages de l'estimateur sont-ils adaptés au problème traité ?

Estimation : cas général

On va considérer dans la suite plusieurs méthodes d'estimation :

1. *Les estimateurs du maximum de vraisemblance* : ces estimateurs ont de très bonnes propriétés théoriques (consistance, normalité asymptotique) et sont usuellement faciles à calculer.

(ii) *Les intervalles de confiance* : ces estimateurs ont une caractéristique *pratique* très utile – on construit un intervalle à partir des données et on connaît la probabilité que notre intervalle (aléatoire) contienne la valeur (inconnue) du paramètre θ.

Estimation : cas général

(iii) *Les estimateurs Uniformément Sans Biais de Variance Minimale (ang. UMVUE estimators)* : des résultats théoriques permettent d'en obtenir un grand nombre mais, comme on l'a vu, un estimateur biaisé peut parfois être meilleur.

(iv) *Les estimateurs bayésiens* : ces estimateurs sont appropriés lorsqu'il est raisonnable de supposer que l'état de la nature θ est une *variable aléatoire* de densité *connue*.

Estimation : cas général

De façon générale, la Théorie Statistique fournit des candidats et l'expérience pratique permet d'en retenir un.