37 Estimation

ECG Maths Approfondies
Semestre 4

Il arrive fréquemment qu'un phénomène aléatoire soit régi par des paramètres inconnus. Il peut arriver que ces paramètres ne puissent être déterminés avec précision : par exemple, on peut savoir qu'une pièce n'est pas équilibrée et ne pas connaître avec précision la probabilité qu'elle donne « pile », on peut savoir que le nombre de clients se présentant à un guichet de la poste dans un intervalle de temps donné suit une loi de Poisson mais ne pas connaître son paramètre.

Il peut aussi arriver qu'il soit envisageable de les déterminer avec précision mais que le coût soit prohibitif : par exemple, un institut de sondage ne peut prévoir le résultat d'un référendum avec exactitude à moins d'interroger tous les individus de la population. Cet institut préférera donc interroger un échantillon de la population et extrapoler le résultat de ce sondage à la population entière.

Le phénomène aléatoire étudié conduit donc à définir une variable aléatoire X dont la loi μ_θ dépend d'un paramètre θ inconnu (réel ou vectoriel). On cherche alors à estimer la valeur de θ ou bien une valeur caractéristique $g(\theta)$ (g étant une fonction définie sur l'ensemble Θ des valeurs possibles de θ) de la loi μ_θ (par exemple son espérance, sa variance,...).

Le problème de l'estimation consiste alors à estimer la vraie valeur de $g(\theta)$ à partir d'un échantillon de données $x_1, \dots, x_n$ obtenues en observant n fois le phénomène.

Dans tout le cours, X est une variable aléatoire sur un espace probabilisable $(\Omega, \mathcal{A})$. On suppose que la loi de X n'est pas entièrement déterminée et appartient à une famille de lois dépendant d'un paramètre θ décrivant un sous-ensemble Θ de $\mathbb{R}$ (ou éventuellement de $\mathbb{R}^2$). $(\Omega, \mathcal{A})$ est muni d'une famille de probabilités $(\mathbb{P}_\theta)_{\theta\in\Theta}$.

Lorsqu'elles existent, l'espérance et la variance de X pour la probabilité $\mathbb{P}_\theta$ devraient être notées $\mathbb{E}_\theta(X)$ et $\mathbb{V}_\theta(X)$, mais, pour simplifier les notations, la probabilité sera plus simplement notée $\mathbb{P}$, l'espérance et la variance seront notées $\mathbb{E}(X)$ et $\mathbb{V}(X)$, mais on se souviendra qu'elles dépendent de la probabilité $\mathbb{P}_\theta$.

A. Estimation ponctuelle

A.1. Échantillonage

Dans ce paragraphe, n désigne un entier naturel n non nul.

Définition 37.1

On appelle ***n*-échantillon** de la loi μ_θ de X (ou plus simplement de X) toute famille $(X_i)_{1\leqslant i\leqslant n}$ de variables aléatoires définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ et de même loi que X.
On dit que $(X_i)_{1\leqslant i\leqslant n}$ est un n-échantillon indépendant et identiquement distribué (en abrégé *i.i.d.*) de X lorsque $(X_i)_{1\leqslant i\leqslant n}$ est un n-échantillon de X constitué de variables aléatoires mutuellement indépendantes.
Si $(X_i)_{1\leqslant i\leqslant n}$ est un n-échantillon de X, un échantillon observé est un n-uplet $(x_i)_{1\leqslant i\leqslant n} = (X_i(\omega))_{1\leqslant i\leqslant n}$ de valeurs prises par $X_1, \dots, X_n$.

Exemple 37.1 On dispose d'une pièce, non forcément équilibrée et l'on cherche à évaluer la probabilité p que cette pièce donne « pile ». On note X une variable aléatoire suivant la loi de Bernoulli de paramètre p. Si l'on effectue n ($n \in \mathbb{N}^*$) lancers successifs et indépendants de la pièce et si l'on note, pour tout entier $i \in [\![1, n]\!]$, X_i la variable aléatoire prenant la valeur 1 si le $i^{ème}$ lancer donne « pile » et 0 sinon, alors la famille $(X_i)_{1\leqslant i\leqslant n}$ est un n-échantillon *i.i.d.* de X.

A.2. Estimateur

Définition 37.2

On appelle **estimateur** de $g(\theta)$ toute variable aléatoire réelle de la forme $\varphi(X_1, \dots, X_n)$ où $(X_i)_{1\leqslant i\leqslant n}$ est un n-échantillon *i.i.d.* de X et φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$, au moins définie sur $X_1(\Omega)\times\cdots\times X_n(\Omega)$, éventuellement dépendante de n, mais indépendante de θ.
Si $\varphi(X_1, \dots, X_n)$ est un estimateur de $g(\theta)$, la réalisation de $\varphi(X_1(\omega), \dots, X_n(\omega))$ (où ω est le relevé effectué dans la population) est appelée **estimation** de $g(\theta)$.

Définition 37.3

On appelle **suite d'estimateurs** de $g(\theta)$ toute suite $(T_n)_{n\in\mathbb{N}^*}$ de variables aléatoires réelles telle que, pour tout $n \in \mathbb{N}^*$, T_n soit un estimateur de $g(\theta)$.

A.3. Exemple d'estimateur : la moyenne empirique

Si l'on dispose d'une pièce et que l'on souhaite estimer la probabilité qu'elle donne « pile », une première méthode consiste intuitivement à effectuer un certain nombre n de lancers puis à calculer le rapport du nombre de « piles » obtenus au nombre de lancers effectués. Ce rapport est appelé « moyenne empirique » et cette méthode est applicable dans la plupart des situations.

Définition 37.4

Soit X une variable aléatoire admettant une espérance m inconnue, n un entier naturel non nul et $(X_i)_{1\leqslant i\leqslant n}$ un n-échantillon *i.i.d.* de X. On note :

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

$\overline{X}_n$ est appelé **moyenne empirique** associée à $(X_i)_{1\leqslant i\leqslant n}$.

Proposition 37.5

Soit X une variable aléatoire admettant une espérance m inconnue, n un entier naturel non nul et $(X_i)_{1\leqslant i\leqslant n}$ un n-échantillon *i.i.d.* de X. On note :

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

$\overline{X}_n$ est un estimateur de m. De plus, $\overline{X}_n$ admet une espérance et :

$$\mathbb{E}(\overline{X}_n) = m$$

Si de plus X admet une variance σ^2, alors $\overline{X}_n$ admet une variance et :

$$\mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n}$$

Exercice 37.1 Démontrer la proposition 37.5.

B. Qualité d'un estimateur

B.1. Biais d'un estimateur

Définition 37.6

Si n est un entier naturel non nul, si T_n est un estimateur de $g(\theta)$ et si T_n admet une espérance, on appelle **biais** de T_n le réel

$$b_\theta(T_n) = \mathbb{E}(T_n) - g(\theta)$$

On dit que T_n est un **estimateur sans biais** de $g(\theta)$ si :

$$\mathbb{E}(T_n) = g(\theta)$$

Remarque En pratique, il est important de comprendre que, comme on ne connaît pas θ, on ne connaît pas non plus $g(\theta)$ (sinon, pourquoi chercherait-on à l'estimer ?). En revanche, on connaît un ensemble Θ contenant de manière certaine la valeur de θ (éventuellement $\mathbb{R}$ ou $\mathbb{R}$).

Pour être parfaitement rigoureux, il conviendrait donc d'écrire que T_n est un estimateur sans biais de $g(\theta)$ si :

$$\forall \tilde{\theta} \in \Theta,\ \mathbb{E}_{\tilde{\theta}}(T_n) = g(\tilde{\theta})$$

Cependant, comme ces notations n'apportent rien d'utile en général, nous ne les utilisons pas ici pour simplifier la compréhension.

Exemple 37.2 m est l'espérance d'une variable aléatoire X et si $(X_1, \dots, X_n)$ est un n-échantillon *i.i.d.* de X, alors la moyenne empirique $\overline{X}_n$ est un estimateur sans biais de m d'après 37.5.

Définition 37.7

Si $(T_n)_{n\in\mathbb{N}^*}$ est une suite d'estimateurs de $g(\theta)$ et si, pour tout entier naturel n non nul, T_n admet une espérance, on dit que $(T_n)_{n\in\mathbb{N}^*}$ est une suite d'estimateurs **asymptotiquement sans biais** de $g(\theta)$ si :

$$\lim_{n\to+\infty} \mathbb{E}(T_n) = g(\theta)$$

On dit parfois plus simplement que T_n est un **asymptotiquement sans biais** de $g(\theta)$.

Exercice 37.2 Soit X une variable aléatoire admettant une espérance m et une variance σ^2 inconnues, $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi que X. On note :

$$\forall n \in \mathbb{N}^*,\ V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

où $\overline{X}_n$ désigne la moyenne empirique associée à $(X_i)_{1\leqslant i\leqslant n}$.

Calculer l'espérance et la variance des variables aléatoires $X_i - \overline{X}_n$ pour $i \in [\![1, n]\!]$ et en déduire que la suite $(V_n)_{n\in\mathbb{N}^*}$ est une suite d'estimateurs de σ^2 asymptotiquement sans biais.

B.2. Convergence d'une suite d'estimateurs

Définition 37.8

Si $(T_n)_{n\in\mathbb{N}^*}$ est une suite d'estimateurs de $g(\theta)$, on dit que $(T_n)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs de $g(\theta)$ si :

$$T_n \xrightarrow{\mathbb{P}} g(\theta)$$

autrement dit si :

$$\forall \varepsilon \in \mathbb{R}_+^*,\ \lim_{n\to+\infty} \mathbb{P}(|T_n - g(\theta)| > \varepsilon) = 0$$

Par abus de langage, on dira aussi que T_n est un estimateur convergent de $g(\theta)$.

Exercice 37.3 Soit X une variable aléatoire admettant une espérance m inconnue et une variance σ^2 et $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi que X. On note :

$$\forall n \in \mathbb{N}^*,\ \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Montrer que la suite $(\overline{X}_n)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs sans biais de m.

Théorème 37.9

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $g(\theta)$ admettant tous une espérance et une variance. Si cette suite est telle que :

$$\lim_{n\to+\infty} \mathbb{E}(T_n) = g(\theta) \quad \text{et} \quad \lim_{n\to+\infty} \mathbb{V}(T_n) = 0$$

alors la suite $(T_n)_{n\in\mathbb{N}^*}$ est convergente.

Exercice 37.4 On se propose de démontrer le théorème 37.9.

1. Justifier que :

$$\forall \varepsilon \in \mathbb{R}_+^*,\ \forall n \in \mathbb{N}^*,\ \mathbb{P}(|T_n - g(\theta)| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}((T_n - g(\theta))^2}{\varepsilon^2}$$

2. Conclure.

Proposition 37.10

Si $(T_n)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs de $g(\theta)$ et si f est une fonction continue sur $\mathbb{R}$, à valeurs dans $\mathbb{R}$, alors $(f(T_n))_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs de $f(g(\theta))$.

C. Estimation par intervalle de confiance

On a vu que l'on pouvait estimer ponctuellement une grandeur $g(\theta)$ à l'aide d'estimateurs et même que l'on pouvait juger, sous certaines conditions, la qualité de cet estimateur. Cependant, aucune information n'était donnée sur la probabilité que la grandeur estimée soit effectivement proche de l'estimation fournie.
Le but de cette partie est de donner comme estimation un intervalle contenant $g(\theta)$ à estimer avec une certaine probabilité.
Dans tout ce paragraphe, $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ désigneront deux suites d'estimateurs de $g(\theta)$ telles que :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}(U_n \leqslant V_n) = 1$$

C.1. Définition

Définition 37.11

Soit $\alpha \in [0,1]$. $[U_n, V_n]$ est appelé **intervalle de confiance** de $g(\theta)$ au niveau de confiance $1-\alpha$ (ou au risque α) si :

$$\mathbb{P}(U_n \leqslant g(\theta) \leqslant V_n) \geqslant 1-\alpha$$

Sa réalisation est l'estimation de cet intervalle de confiance.

Remarque En pratique, si l'on connaît un estimateur T_n de $g(\theta)$, on cherchera le plus souvent un intervalle de confiance de la forme $[T_n - \varepsilon, T_n + \varepsilon]$ où ε est un réel strictement positif. Il s'agira alors de déterminer un réel ε strictement positif tel que :

$$\mathbb{P}(T_n - \varepsilon \leqslant g(\theta) \leqslant T_n + \varepsilon) \geqslant 1-\alpha$$

ou encore tel que :

$$\mathbb{P}(|T_n - g(\theta)| > \varepsilon) \leqslant \alpha$$

Dès lors, on voit que l'on pourra, dans le cas où T_n admet une espérance et/ou un moment d'ordre 2, utiliser l'inégalité de Markov et/ou de Bienaymé-Tchebychev pour déterminer un tel réel ε.

Définition 37.12

Soit $\alpha \in [0,1]$. On appelle **intervalle de confiance asymptotique** de $g(\theta)$ au niveau de confiance $1-\alpha$ (ou au risque α) toute suite $([U_n, V_n])_{n\in\mathbb{N}^*}$ telle qu'il existe une suite $(\alpha_n)_{n\in\mathbb{N}^*}$ telle que :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}(U_n \leqslant g(\theta) \leqslant V_n) \geqslant 1-\alpha_n \quad \text{et} \quad \lim_{n\to+\infty} \alpha_n = \alpha$$

Par abus de langage, on dira aussi que $[U_n, V_n]$ est un intervalle de confiance asymptotique de $g(\theta)$.

C.2. Estimation par intervalle de confiance d'une proportion

On suppose, dans cette partie, que X suit la loi de Bernoulli de paramètre p, inconnu, que l'on cherche à estimer. On considère également un réel α appartenant à $]0,1[$ et une suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires indépendantes et toutes de même loi que X.
Enfin, on note :

$$\forall n \in \mathbb{N}^*,\ \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Une première approche

Soit $n \in \mathbb{N}^*$. On a vu que la moyenne empirique $\overline{X}_n$ est un estimateur sans biais de p et que :

$$\mathbb{V}(\overline{X}_n) = \frac{p(1-p)}{n}$$

De l'inégalité de Bienaymé-Tchebychev, on déduit que :

$$\forall \varepsilon \in \mathbb{R}_+^*,\ \mathbb{P}(|\overline{X}_n - p| > \varepsilon) \leqslant \frac{p(1-p)}{n\varepsilon^2}$$

De plus, on peut remarquer que, comme p est réel :

$$\begin{aligned} p(1-p) &= p - p^2 \\ &= \frac{1}{4} - \left(\frac{1}{2} - p\right)^2 \\ &\leqslant \frac{1}{4} \end{aligned} \tag{37.1}$$

On en déduit donc que :

$$\forall \varepsilon \in \mathbb{R}_+^*,\ \mathbb{P}(|\overline{X}_n - p| > \varepsilon) \leqslant \frac{1}{4n\varepsilon^2}$$

Par conséquent, pour que $[\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon]$ soit un intervalle de confiance de p au niveau de confiance $1-\alpha$, il suffit que ε vérifie :

$$\frac{1}{4n\varepsilon^2} \leqslant \alpha$$

soit encore :

$$\varepsilon \geqslant \frac{1}{2\sqrt{n\alpha}}$$

On en déduit le résultat suivant :

Proposition 37.13

Soit $\alpha \in]0,1[$ et $n \in \mathbb{N}^*$. Si X suit la loi de Bernoulli de paramètre p, alors $\left[\overline{X}_n - \dfrac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \dfrac{1}{2\sqrt{n\alpha}}\right]$ est un intervalle de confiance de p au niveau de confiance $1-\alpha$.

Une seconde approche

Soit $\varepsilon \in \mathbb{R}_+^*$. On peut aussi remarquer que, grâce à la majoration (37.1) :

$$\begin{aligned}\forall n \in \mathbb{N}^*,\ \left[\left|\overline{X}_n - p\right| > \varepsilon\right] &= \left[\left|\sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}\right| > \varepsilon\,\frac{\sqrt{n}}{p(1-p)}\right] \\ &\subset \left[\left|\sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}\right| > 2\varepsilon\sqrt{n}\right]\end{aligned}$$

et donc :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leqslant \mathbb{P}\left(\left|\overline{X}_n^*\right| > 2\varepsilon\sqrt{n}\right) \tag{37.2}$$

où l'on a posé :

$$\forall n \in \mathbb{N}^*,\ \overline{X}_n^* = \sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}$$

D'après le théorème de la limite centrée, comme la suite $(X_n)_{n\in\mathbb{N}^*}$ est une suite de variables aléatoires indépendantes, de même loi et admettant une variance non nulle, la suite $(\overline{X}_n^*)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire N suivant la loi normale centrée réduite, et donc que, pour $x \in \mathbb{R}_+^*$:

$$\lim_{n\to+\infty} \mathbb{P}\left(\left|\overline{X}_n^*\right| > x\right) = \mathbb{P}(|N| > x) \tag{37.3}$$

Par ailleurs, en notant Φ la fonction de répartition de la loi normale centrée réduite, on a :

$$\begin{aligned}\mathbb{P}(|N| > x) &= 1 - \mathbb{P}(-x \leqslant N \leqslant x) \\ &= 1 - \Phi(x) + \Phi(-x) \\ &= 2\,[1 - \Phi(x)]\end{aligned}$$

et donc :

$$\mathbb{P}(|N| > x) = \alpha \iff \Phi(x) = 1 - \frac{\alpha}{2}$$

Par ailleurs, comme Φ est strictement croissante et continue sur $\mathbb{R}$ avec :

$$\lim_{x\to-\infty} \Phi(x) = 0 \quad \text{et} \quad \lim_{x\to+\infty} \Phi(x) = 1$$

Φ réalise une bijection de $\mathbb{R}$ sur $]0,1[$, donc il existe un unique réel t_α tel que :

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$$

On peut alors considérer les suites $(\varepsilon)_{n\in\mathbb{N}^*}$ et $(\alpha_n)_{n\in\mathbb{N}^*}$ définies par :

$$\forall n \in \mathbb{N}^*,\ \varepsilon_n = \frac{t_\alpha}{2\sqrt{n}} \quad \text{et} \quad \alpha_n = \mathbb{P}\left(\left|\overline{X}_n^*\right| > t_\alpha\right)$$

On a alors, d'après (37.2) :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon_n\right) \leqslant \alpha_n$$

d'où :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\overline{X}_n - \varepsilon_n \leqslant p \leqslant \overline{X}_n + \varepsilon_n\right) \geqslant 1 - \alpha_n$$

et d'après (37.3) :

$$\lim_{n\to+\infty} \alpha_n = \mathbb{P}\left(|N| > t_\alpha\right) = \alpha$$

ce qui prouve le résultat suivant :

Proposition 37.14

Soit $\alpha \in]0,1[$ et t_α l'unique réel tel que :

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$$

Si X suit la loi de Bernoulli de paramètre p, alors $\left[\overline{X}_n - \dfrac{t_\alpha}{2\sqrt{n}}, \overline{X}_n + \dfrac{t_\alpha}{2\sqrt{n}}\right]$ est un intervalle de confiance asymptotique de p au niveau de confiance $1-\alpha$.

Remarques

a. Pour $\alpha = 0,05$, on a : $t_\alpha \simeq 1,96$ et on a alors :

$$\frac{1}{2\sqrt{n\alpha}} = \frac{2,24}{\sqrt{n}} \quad \text{et} \quad \frac{t_\alpha}{2\sqrt{n}} \simeq \frac{0,98}{\sqrt{n}}$$

b. Pour $\alpha = 0,01$, on a : $t_\alpha \simeq 2,58$ et on a alors :

$$\frac{1}{2\sqrt{n\alpha}} = \frac{5}{\sqrt{n}} \quad \text{et} \quad \frac{t_\alpha}{2\sqrt{n}} \simeq \frac{1,29}{\sqrt{n}}$$

c. On constate dans les deux exemples précédents que l'intervalle de confiance asymptotique obtenu par la seconde approche est plus intéressant que l'intervalle de confiance obtenu par la première approche. C'est le cas le plus souvent, mais il est important de bien comprendre que, la seconde approche étant obtenue par approximation, elle ne donnera de résultat vraiment fiable ou intéressant que pour des tailles d'échantillons suffisamment grandes.

C.3. Estimation par intervalle de confiance de l'espérance d'une variable aléatoire admettant un moment d'ordre 2

Le cas particulier d'une variable aléatoire gaussienne

Dans cette sous-partie, on suppose que la loi de X est la loi normale $\mathcal{N}(m,\sigma^2)$ où m est un réel inconnu et σ un réel strictement positif connu.
On considère alors un réel α appartenant à $]0,1[$, t_α l'unique réel tel que :

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$$

et une suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires indépendantes et toutes de même loi que X. Enfin, on note :

$$\forall n \in \mathbb{N}^*,\ \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \overline{X}_n^* = \sqrt{n}\,\frac{\overline{X}_n - m}{\sigma}$$

Comme la suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires indépendantes et toutes de même loi que X, la stabilité de la loi normale pour la somme permet d'affirmer que la suite $(\overline{X}_n^*)_{n\in\mathbb{N}^*}$ est formée de variables aléatoires suivant toutes la loi normale centrée réduite, et donc que :

$$\begin{aligned}\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\left|\overline{X}_n^*\right| > t_\alpha\right) &= 2\left[1 - \Phi(t_\alpha)\right] \\ &= \alpha\end{aligned}$$

Par ailleurs, on peut remarquer que :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\left|\overline{X}_n^*\right| > t_\alpha\right) = \mathbb{P}\left(\left|\overline{X}_n - m\right| > t_\alpha\,\frac{\sigma}{\sqrt{n}}\right)$$

En posant

$$\forall n \in \mathbb{N}^*,\ \varepsilon_n = t_\alpha\,\frac{\sigma}{\sqrt{n}}$$

on a donc :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\overline{X}_n - \varepsilon_n \leqslant m \leqslant \overline{X}_n + \varepsilon_n\right) = 1 - \alpha$$

Par conséquent, on en déduit le résultat suivant :

Proposition 37.15

Soit $\alpha \in]0,1[$ et t_α l'unique réel tel que :

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$$

Si X est une variable aléatoire suivant la loi normale d'espérance m inconnue et de variance σ^2 **connue**, alors $\left[\overline{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de m au niveau de confiance $1-\alpha$.

Le cas général

Dans cette sous-partie, on suppose que X est une variable aléatoire admettant une espérance m inconnue et une variance σ^2 ($\sigma > 0$).
On considère également un réel α appartenant à $]0,1[$, t_α l'unique réel tel que :

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$$

et une suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires indépendantes et toutes de même loi que X. Enfin, on note :

$$\forall n \in \mathbb{N}^*,\ \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Comme la suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires indépendantes et toutes de même loi que X, le théorème de la limite centrée assure encore que la suite $(\overline{X}_n^*)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire N suivant la loi normale centrée réduite et donc que :

$$\begin{aligned}\lim_{n\to+\infty} \mathbb{P}\left(\left|\overline{X}_n^*\right| > t_\alpha\right) &= \mathbb{P}(|N| > t_\alpha)\\ &= 2\left[1 - \Phi(t_\alpha)\right]\\ &= \alpha\end{aligned}$$

Par ailleurs, on peut remarquer que :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\left|\overline{X}_n^*\right| > t_\alpha\right) = \mathbb{P}\left(\left|\overline{X}_n - m\right| > t_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

En posant

$$\forall n \in \mathbb{N}^*,\ \varepsilon_n = t_\alpha \frac{\sigma}{\sqrt{n}} \quad \text{et} \quad \alpha_n = \mathbb{P}\left(\left|\overline{X}_n^*\right| > t_\alpha\right)$$

on a donc :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\overline{X}_n - \varepsilon_n \leqslant m \leqslant \overline{X}_n + \varepsilon_n\right) = 1 - \alpha_n \quad \text{et} \quad \lim_{n\to+\infty} \alpha_n = \alpha$$

Par conséquent, on en déduit le résultat suivant :

Proposition 37.16

Soit $\alpha \in]0,1[$ et t_α l'unique réel tel que :

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$$

Si X est une variable aléatoire admettant une espérance m inconnue et une variance σ^2 **connue**, alors $\left[\overline{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance asymptotique de m au niveau de confiance $1-\alpha$.

Remarques

a. Le lecteur avisé aura remarqué que l'hypothèse « une variance σ^2 connue » est très contraignante. En effet, il est assez difficile d'imaginer que la variance puisse être connue, mais pas l'espérance (même si cela peut se produire).

b. En revanche, nous avons vu dans l'exercice 31.2 que la variance empirique V_n est un estimateur de la σ^2 (biaisé, mais asymptotiquement sans biais). Dans le cas où X admet un moment d'ordre 4, on peut calculer le risque quadratique de V_n et prouver qu'il converge vers 0. Il en découle que la suite $(V_n)_{n\in\mathbb{N}^*}$ converge en probabilité vers σ^2. Dès lors, comme la fonction $t \mapsto \sqrt{t}$ est continue sur $\mathbb{R}^+$, on en déduit que la suite $(\sqrt{V_n})_{n\in\mathbb{N}^*}$ converge en probabilité vers σ, on peut prouver qu'il est possible de remplacer σ par $\sigma_n = \sqrt{V_n}$ dans l'intervalle de confiance précédent.

c. On admettra que, plus généralement :

Proposition 37.17

Soit $\alpha \in]0,1[$ et t_α l'unique réel tel que :

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$$

Soit X est une variable aléatoire admettant une espérance m inconnue et une variance σ^2 **inconnue** et $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi que X. On note :

$$\forall n \in \mathbb{N},\ V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{et} \quad \sigma_n = \sqrt{V_n}$$

Alors $\left[\overline{X}_n - t_\alpha \frac{\sigma_n}{\sqrt{n}}, \overline{X}_n + t_\alpha \frac{\sigma_n}{\sqrt{n}}\right]$ est un intervalle de confiance asymptotique de m au niveau de confiance $1-\alpha$.

D. Correction des exercices

Correction de l'exercice 37-1

► $(X_1, \dots, X_n)$ est un n-échantillon *i.i.d.* de X et la fonction

$$\varphi : (x_1, \dots, x_n) \mapsto \frac{1}{n}\sum_{i=1}^{n} x_i$$

est définie sur $\mathbb{R}^n$ et indépendante de m, donc $\overline{X}_n = \varphi(X_1, \dots, X_n)$ est un estimateur de m.

► Comme les variables aléatoires de la suite $(X_n)_{n\in\mathbb{N}^*}$ admettent une même espérance m on a, par linéarité de l'espérance :

$$\begin{aligned}\forall n \in \mathbb{N}^*,\ \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k)\\ &= m\end{aligned}$$

► De plus, comme les variables aléatoires de la suite $(X_n)_{n\in\mathbb{N}^*}$ sont mutuellement indépendantes, si elles admettent une même variance σ^2, alors on a :

$$\begin{aligned}\forall n \in \mathbb{N}^*,\ \mathbb{V}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}(X_k)\\ &= \frac{\sigma^2}{n}\end{aligned}$$

Correction de l'exercice 37-2

◇ Soit $i \in [\![1, n]\!]$. Comme les variables aléatoires $X_1, \dots, X_n$ admettent une espérance, Y_i admet une espérance et on a, par linéarité de l'espérance :

$$\begin{aligned}\mathbb{E}(Y_i) &= \mathbb{E}(X_i) - \mathbb{E}(\overline{X}_n)\\ &= m - m\\ &= 0.\end{aligned}$$

◇ Soit $i \in [\![1,n]\!]$. Comme les variables aléatoires $X_1,\dots,X_n$ admettent une variance, Y_i admet une variance et on a, $X_1,\dots,X_n$ étant indépendantes :

$$\begin{aligned}\mathbb{V}(Y_i) &= \mathbb{V}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\\ &= \mathbb{V}\left(\frac{n-1}{n}X_i - \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_j\right)\\ &= \left(\frac{n-1}{n}\right)^2 \mathbb{V}(X_i) + \frac{1}{n^2}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathbb{V}(X_j)\\ &= \left(\frac{n-1}{n}\right)^2 \sigma^2 + \frac{1}{n^2}\sum_{\substack{j=1\\ j\neq i}}^{n}\sigma^2\\ &= \left(\frac{n-1}{n}\right)^2 \sigma^2 + \frac{n-1}{n^2}\sigma^2\\ &= \frac{n-1}{n}\,\sigma^2.\end{aligned}$$

◇ Soit $n \in \mathbb{N}^*$. $(X_1,\dots,X_n)$ est un n-échantillon *i.i.d.* de X et la fonction

$$\varphi : (x_1,\dots,x_n) \mapsto \frac{1}{n}\sum_{i=1}^{n}\left[x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right]^2$$

est définie sur $\mathbb{R}^n$ et indépendante de σ^2, donc $V_n = \varphi(X_1,\dots,X_n)$ est un estimateur de σ^2. De plus, comme $X_1,\dots,X_n$ admettent toutes un moment d'ordre 2, V_n admet une espérance et, par linéarité de l'espérance :

$$\begin{aligned}\mathbb{E}(V_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i^2)\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbb{V}(Y_i) + (\mathbb{E}(Y_i))^2\Big]\\ &= \frac{n-1}{n}\,\sigma^2\end{aligned}$$

On a donc :

$$\lim_{n\to+\infty}\mathbb{E}(V_n) = \sigma^2$$

donc la suite $(V_n)_{n\in\mathbb{N}^*}$ est une suite d'estimateurs de σ^2 asymptotiquement sans biais.

Correction de l'exercice 37-3

On a déjà vu dans l'exercice 31.2 que, pour tout $n \in \mathbb{N}^*$, $\overline{X}_n$ est un estimateur sans biais de m. De plus, comme la suite $(X_n)_{n\in\mathbb{N}^*}$ est une suite de variables aléatoires indépendantes admettant une même espérance et une même variance, la loi faible des grands nombres nous permet d'affirmer que :

$$\overline{X}_n \xrightarrow{\mathbb{P}} m$$

donc que la suite $(\overline{X}_n)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs de m.

Correction de l'exercice 37-4

1. Soit $n \in \mathbb{N}^*$ et $\varepsilon \in \mathbb{R}_+^*$. T_n admet une variance, donc $(T_n - g(\theta))^2$ est une variable aléatoire positive admettant une espérance. On en déduit, avec l'inégalité de Markov :

$$\mathbb{P}\big((T_n - g(\theta)^2 \geqslant \varepsilon^2\big) \leqslant \frac{\mathbb{E}((T_n - g(\theta))^2}{\varepsilon^2}$$

d'où, comme la fonction $t \mapsto \sqrt{t}$ est strictement croissante sur $\mathbb{R}^+$:

$$\forall \varepsilon \in \mathbb{R}_+^*,\ \forall n \in \mathbb{N}^*,\ \mathbb{P}(|T_n - g(\theta)| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}((T_n - g(\theta))^2}{\varepsilon^2}$$

2. Une probabilité étant toujours positive, on a donc, d'après la formule de Koenig-Huygens :

$$\begin{aligned}\forall \varepsilon \in \mathbb{R}_+^*,\ \forall n \in \mathbb{N}^*,\ 0 \leqslant \mathbb{P}(|T_n - g(\theta)| \geqslant \varepsilon) &\leqslant \frac{\mathbb{V}(T_n - g(\theta)) + \left[\mathbb{E}(T_n - g(\theta))\right]^2}{\varepsilon^2} \\ &\leqslant \frac{\mathbb{V}(T_n) + \left[\mathbb{E}(T_n) - g(\theta)\right]^2}{\varepsilon^2}\end{aligned}$$

Or on a :

$$\lim_{n\to+\infty} \frac{\mathbb{V}(T_n) + \left[\mathbb{E}(T_n) - g(\theta)\right]^2}{\varepsilon^2} = 0$$

donc, d'après le théorème de l'encadrement :

$$\forall \varepsilon \in \mathbb{R}_+^*,\ \lim_{n\to+\infty} \mathbb{P}(|T_n - g(\theta)| \geqslant \varepsilon) = 0$$

ce qui prouve le résultat attendu.

Sommaire