38 Estimation

ECG Maths Appliquées
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 38.1

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 38.1 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 38.2

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 $g(\theta)$.

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 38.3

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 38.4

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 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 38.1 Démontrer la proposition 38.4.

B. 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 :

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

B.1. Définition

Définition 38.5

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 38.6

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 $g(\theta)$.

B.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 :

$$\begin{aligned} p(1-p) &= p - p^2 \\ &= \frac{1}{4} - \left(\frac{1}{2} - p\right)^2 \\ &\leqslant \frac{1}{4} \end{aligned} \tag{38.1}$$

On en déduit donc que :

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

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 38.7

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 (38.1) :

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

et donc :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}(\left|\overline{X}_n - p\right| > \varepsilon) \leqslant \mathbb{P}\left(\left|\overline{X}_n^*\right| > 2\varepsilon\sqrt{n}\right) \tag{38.2}$$

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 $x \in \mathbb{R}_+^*$:

$$\lim_{n\to+\infty} \mathbb{P}\left(\left|\overline{X}_n^*\right| > x\right) = \mathbb{P}(|N| > x) \tag{38.3}$$

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

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

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 :

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

On a alors, d'après (38.2) :

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

d'où :

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

et d'après (38.3) :

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

ce qui prouve le résultat suivant :

Proposition 38.8

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. Correction des exercices

Correction de l'exercice 38-1

► $(X_1,\dots,X_n)$ 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 :

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

Sommaire