CHAPITRE 3: ERREURS D'ESTIMATION

- $x_1, \dots, x_n \in \mathcal{O} \subset \mathbb{I}$ observations indépendantes qui sont réalisations d'une loi $\mathcal{L}_{\theta^*}$, $\theta^* \in \Theta$
- θ^* inconnu $\rightarrow$ on introduit des n-échantillons, i.e. des suites $X_1, \dots, X_n$ de vaiid de même loi $\mathcal{L}_\theta$, $\theta \in \Theta$
- Le paramètre d'intérêt est $g(\theta)$, avec $g : \Theta \rightarrow \mathbb{R}$.
- Un estimateur est une fonction $\hat{g} = \hat{g}(X_1, \dots, X_n)$. Sa vocation est d'approcher $g(\theta)$, ceci $\forall \theta \in \Theta$
- Alors, comme $(x_1, \dots, x_n)$ est une réalisation de $\hat{g}(X_1, \dots, X_n)$ avec $X_1, \dots, X_n$ vaiid de loi $\mathcal{L}_{\theta^*}$, on en déduit que $\hat{g}(x_1, \dots, x_n) \approx g(\theta^*)$.

1 - Intervalles de confiance

L'idée est de construire un intervalle $I(x_1, \dots, x_n)$ tel que $g(\theta^*) \in I(x_1, \dots, x_n)$, avec un "niveau de confiance" suffisamment élevé.

Définition (Intervalle de confiance). Fixons $\alpha \in]0,1[$. Pour tout $\theta \in \Theta$ et $X_1, \dots, X_n$ un n-échantillon i.i.d. de la loi $\mathcal{L}_\theta$, un intervalle de confiance par excès, de niveau de confiance $1-\alpha$ pour $g(\theta)$, est un intervalle aléatoire $I(X_1, \dots, X_n)$ tel que

$$\boxed{\mathbb{P}(g(\theta) \in I(X_1, \dots, X_n)) \geq 1-\alpha.}$$

Comme les observations sont issues de la loi $\mathcal{L}_{\theta^*}$, on aura donc $g(\theta^*) \in I(x_1, \dots, x_n)$ avec un niveau de confiance $\geq 1-\alpha$. On a ainsi une approximation de $g(\theta^*)$, avec sa marge d'erreur.

Exemple de construction d'un IC (intervalle de confiance).

On lance 100 fois une pièce dans le but de donner une approximation de la proba $\theta^* \in]0,1[$ de tomber sur Pile.

On code par 0 = face et 1 = Pile et l'observation est $x_1, \dots, x_{100}$ telle que $\bar{x}_{100} = \frac{1}{100}\sum_{i=1}^{100} x_i = 0,7$. Notons $X_1, \dots, X_{100}$ v.a.i.d de loi $\mathcal{B}(\theta)$, $\theta \in]0,1[$.

Un estimateur naturel pour θ est la moyenne empirique $\bar{X}_{100} = \hat{\theta}$ car, par la LGN : $\bar{X}_{100} \approx \mathbb{E}(X_1) = \theta$ (vu à la séance n°3). Or, d'après l'inégalité de Bienaymé-Tchebitchev, on a $\forall a > 0$:

$$\mathbb{P}\left(|\bar{X}_{100} - \underbrace{\mathbb{E}(X_1)}_{=\theta}| \geq a\right) \leq \frac{\mathrm{var}(\bar{X}_{100})}{a^2}$$

Or, $\mathrm{var}(\bar{X}_{100}) = \mathrm{var}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) = \frac{1}{100^2}\mathrm{var}\left(\sum_{i=1}^{100} X_i\right)$

$= \frac{1}{100^2}\sum_{i=1}^{100}\mathrm{var}(X_i)$ car $X_1, \dots, X_{100}$ indép.

$= \frac{1}{100^2}\sum_{i=1}^{100}\mathrm{var}(X_1)$ car $X_1, \dots, X_{100}$ de même loi.

Rappels

- $\mathrm{var}(\alpha Z) = \alpha^2 \mathrm{var}(Z)$
- $\mathrm{var}\left(\sum_{i=1}^{n} Z_i\right) = \sum_{i=1}^{n}\mathrm{var}(Z_i)$ si $Z_1, \dots, Z_n$ indép.

$$= \frac{1}{100^2} \times 100 \operatorname{Var}(X_1) = \frac{\operatorname{Var}(X_1)}{100} = \frac{\theta(1-\theta)}{100} \quad \text{car } X_1 \sim \mathcal{B}(\theta).$$

On a donc, $\forall a > 0$:

$$\mathbb{P}\left(|\bar{X}_{100} - \theta| \geq a\right) \leq \frac{\theta(1-\theta)}{100 a^2}$$

On veut majorer $\theta(1-\theta)$. Soit $f(\theta) = \theta(1-\theta)$, $\theta \in]0,1[$. On a $f'(\theta) = 1 - 2\theta$

et $f'(\theta) \geq 0 \Leftrightarrow 1 - 2\theta \geq 0 \Leftrightarrow \theta \leq \frac{1}{2}$

θ	0		$1/2$		1
f'		+	0	−	
f		↗		↘	

Donc $f(\theta) \leq f\left(\frac{1}{2}\right) \quad \forall \theta \in]0,1[$

$\| \quad \frac{1}{2} \times \left(1 - \frac{1}{2}\right) = \frac{1}{4}$

$\Rightarrow \theta(1-\theta) = f(\theta) \leq \frac{1}{4} \quad \forall \theta \in]0,1[.$

Alors, $\mathbb{P}\left(|\bar{X}_{100} - \theta| \geq a\right) \leq \frac{1}{400 a^2} \Leftrightarrow 1 - \mathbb{P}\left(|\bar{X}_{100} - \theta| < a\right) \leq \frac{1}{400 a^2}$

$$\Leftrightarrow \mathbb{P}\left(|\bar{X}_{100} - \theta| < a\right) \geq 1 - \frac{1}{400 a^2}$$

Si on cherche un IC par excès de niveau 95%, on choisit $a \in \mathbb{R}$ de sorte que

$$1 - \frac{1}{400a^2} = 0,95 \iff \frac{1}{400a^2} = 0,05 \iff a^2 = \frac{1}{400 \times 0,05},$$

d'où $a = \frac{1}{\sqrt{20}}$. Alors, on a

$$\mathbb{P}\left(|\bar{X}_{100} - \theta| < \frac{1}{\sqrt{20}}\right) \geq 0,95$$

$$\iff \mathbb{P}\left(\bar{X}_{100} - \theta \in \left]-\frac{1}{\sqrt{20}}, \frac{1}{20}\right[\right) \geq 0,95$$

$$\iff \mathbb{P}\left(\theta \in \underbrace{\left]\bar{X}_{100} - \frac{1}{\sqrt{20}}, \bar{X}_{100} + \frac{1}{\sqrt{20}}\right[}_{= I(X_1, \dots, X_{100})}\right) \geq 0,95$$

$= I(X_1, \dots, X_{100})$: c'est l'IC au niveau de confiance par excès de 0,95

Rappel

$|x| < b$

$\iff -b < x < b$

$\iff x \in]-b, b[$

Décliné sur l'observation $x_1, \dots, x_{100}$ ($\bar{x}_{100} = 0,7$), on trouve

$$\theta^* \in \left]\underbrace{\bar{x}_{100}}_{0,7} - \frac{1}{\sqrt{20}}, \underbrace{\bar{x}_{100}}_{0,7} + \frac{1}{\sqrt{20}}\right[\text{ avec un niveau de confiance} \geq 0,95.$$

AN : $\theta^* \in]0{,}47 ; 0{,}92[$ avec un niveau de confiance $\geq 0{,}95$.

Résultat pas terrible ! Pour l'améliorer :

- augmenter le nombre d'expériences par ex. 1000 au lieu de 100.
- baisser le niveau de confiance. Alors l'IC sera de + faible longueur, mais résultat moins fiable
- utiliser l'inégalité de Hoeffding (cf feuille TD1) au lieu de l'inégalité de BT.

Remarques

① niveau de confiance vs longueur de l'IC. Il y a 2 critères de qualité pour un IC : sa longueur et son niveau de confiance. Or ces 2 critères s'opposent :

- si niveau de confiance ↗ ⇒ longueur de l'IC ↗
- longueur de l'IC ↘ ⇒ niveau de confiance ↘

Pour résoudre ce dilemme, on prend un niveau de confiance = 90% ou 95%.

② Un intervalle de confiance est construit avec un estimateur. La qualité de l'IC dépend donc de la qualité de l'estimateur.

2- Biais

Soit $\hat{g}$ un estimateur de $g(\theta)$. On s'attend à ce que les valeurs de la va $\hat{g}$ fluctuent autour de $g(\theta)$ et plus précisément qu'en moyenne elle soit égale à $g(\theta)$. C'est le concept de biais.

Définition (Biais) Soit $\theta \in \Theta$ et soit $X_1, \dots, X_n$ un n-échantillon iid de loi $\mathcal{L}_\theta$. Le biais en θ de l'estimateur $\hat{g} = \hat{g}(X_1, \dots, X_n)$ est la quantité

$$B_{\hat{g}}(\theta) = \mathbb{E}(\hat{g}) - g(\theta)$$

On dit que l'estimateur $\hat{g}$ est sans biais si $B_{\hat{g}}(\theta) = 0 \quad \forall \theta \in \Theta$.

Exemples

① Dans le jeu de pile ou face, on lance n fois une pièce. On peut modéliser cette expérience par un n-échantillon $X_1, \dots, X_n$ de variable de loi $\mathcal{B}(\theta)$, $\theta \in]0,1[$.

- Cas où le paramètre d'intérêt est θ. Un estimateur naturel est $\hat{\theta} = \bar{X}_n$ d'après la LGN. Alors, $\forall \theta \in]0,1[$:

$$B_{\hat{\theta}}(\theta) = \mathbb{E}(\hat{\theta}) - \theta = \mathbb{E}(\bar{X}_n) - \theta = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \theta$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) - \theta = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_1) - \theta \quad \text{car } X_1, \dots, X_n \text{ de même loi}$$

$$= \mathbb{E}(X_1) - \theta \overset{X_1 \sim \mathcal{B}(\theta)}{=} \theta - \theta = 0$$

Donc $\hat{\theta}$ est sans biais.

- Cas où le paramètre d'intérêt est $\theta(1-\theta)$ (i.e. la variance de $\mathcal{B}(\theta)$) Comme $\bar{X}_n$ est l'estimateur naturel de θ, $\bar{X}_n(1-\bar{X}_n)$ est l'estimateur par insertion de $\theta(1-\theta)$. On note $\hat{\sigma} = \bar{X}_n(1-\bar{X}_n)$.

Calculons son biais : $B_{\hat{g}}(\theta) = \mathbb{E}(\hat{g}) - \theta(1-\theta)$. Pour cela, on calcule

$$\mathbb{E}(\hat{g}) = \mathbb{E}\left(\bar{X}_n (1-\bar{X}_n)\right) = \mathbb{E}\left(\bar{X}_n - \bar{X}_n^2\right) = \underbrace{\mathbb{E}(\bar{X}_1)}_{=\theta} - \mathbb{E}\left(\bar{X}_n^2\right)$$

Alors,

$$\mathbb{E}\left(\bar{X}_n^2\right) = \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right) = \frac{1}{n^2}\,\mathbb{E}\left(\left(\sum_{i=1}^{n} X_i\right)^2\right)$$

$$= \frac{1}{n^2}\,\mathbb{E}\left(\sum_{i,j=1}^{n} X_i X_j\right)$$: faire ce calcul pour la prochaine séance de CM.