<u>Rappels</u> Si $\hat{g}$ est un estimateur du paramètre d'intérêt $g(\theta)$, le biais en θ est $B_{\hat{g}}(\theta) = \mathbb{E}(\hat{g}) - g(\theta)$, avec $X_1, \dots, X_n$ iid de loi $\mathcal{L}_\theta$.

<u>Exemples</u> (Suite de la séance 5)

① $X_1, \dots, X_n$ iid de loi $\mathcal{B}(\theta)$. On s'intéresse au paramètre $g(\theta) = \theta(1-\theta)$ dont l'estimateur naturel est $\hat{g} = \bar{X}_n(1-\bar{X}_n)$. Pour calculer son biais, on a vu que

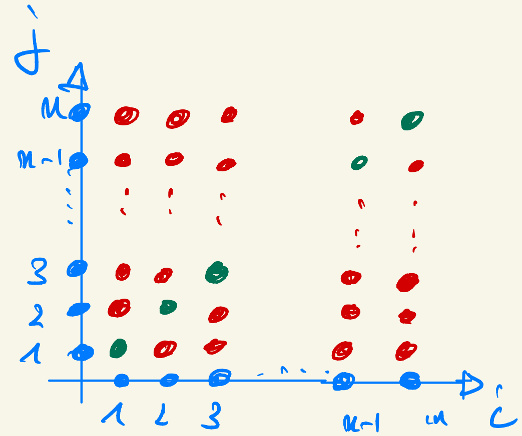
$$\mathbb{E}(\bar{X}_n^2) = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{E}(X_i X_j)$$

Or
$$\mathbb{E}(X_i X_j) = \begin{cases} \mathbb{E}(X_i)\mathbb{E}(X_j) \text{ si } i \neq j, \text{ car alors } X_i \text{ est indépendante de } X_j \\ \mathbb{E}(X_i^2) = \mathbb{E}(X_i) \text{ si } i = j, \text{ car } X_i^2 = X_i \end{cases}$$

$$= \begin{cases} \theta^2 \text{ si } i \neq j \\ \theta \text{ si } i = j \end{cases} \quad \text{car } X_i \sim \mathcal{B}(\theta), \text{ d'où } \mathbb{E}(X_i) = \theta$$

Donc $\mathbb{E}(\bar{X}_n^2) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbb{E}(X_i X_j) = \frac{1}{n^2} \sum_{i=1}^{n} \theta + \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \theta^2$

$$= \frac{1}{n^2} \times n\theta + \frac{1}{n^2}\,\theta^2 \underbrace{\sum_{\substack{i,j=1\\ i\neq j}}^{n} 1}_{\substack{\text{nb de couples}\\ (i,j) \text{ tq } i,j=1,\dots,n \text{ et } i\neq j}}$$



$$= \frac{\theta}{n} + \frac{\theta^2}{n^2}(n^2-n) = \frac{\theta}{n} + \frac{\theta^2}{n}(n-1)$$

Il y a n^2-n points rouges qui correspondent aux couples (i,j) t.q. $i,j=1,\dots,n$ et $i\neq j$

Alors,

$$\mathbb{E}(\hat{g}) = \mathbb{E}\left(\bar{X}_n(1-\bar{X}_n)\right) = \mathbb{E}(\bar{X}_n) - \mathbb{E}(\bar{X}_n^2) = \theta - \left(\frac{\theta+\theta^2(n-1)}{n}\right)$$

Par suite, le biais de $\hat{g}$ en θ vaut:

$$B_{\hat{g}}(\theta) = \mathbb{E}(\hat{g}) - g(\theta) = \mathbb{E}(\hat{g}) - \theta(1-\theta) = \theta - \frac{\theta+\theta^2(n-1)}{n} - \theta(1-\theta)$$

$$= \frac{\cancel{n\theta} - \theta - \cancel{\theta^2 n} + \theta^2 - \cancel{n\theta} + \cancel{n\theta^2}}{n} = \frac{\theta^2-\theta}{n} = \frac{\theta(\theta-1)}{n} \neq 0$$

et $\hat{g}$ est donc $\boxed{\text{biaisé}}$

(2) Dans le modèle statistique décrit par le n-échantillon $X_1, \dots, X_n$ iid de loi $\mathcal{G}(\theta)$ avec $\theta \in]0,1[$. L'estimateur par insertion du paramètre θ est

$$\hat{\theta} = \frac{1}{\overline{X}_n} \quad \text{(cf fin du chapitre précédent)}$$

Il n'est pas possible de calculer son biais, on se retranche vers une majoration du biais. Alors :

$$|B_{\hat{\theta}}(\theta)| = |\mathbb{E}(\hat{\theta}) - \theta| = \left|\mathbb{E}\left(\frac{1}{\overline{X}_n} - \theta\right)\right| \leq \mathbb{E}\left(\left|\frac{1}{\overline{X}_n} - \theta\right|\right)$$

$$\leq \mathbb{E}\left(\left|\frac{1}{\overline{X}_n} - \theta\right|\right) = \mathbb{E}\left(\frac{|1 - \theta \overline{X}_n|}{\overline{X}_n}\right)$$

Rappel si Z v.a. :

$|\mathbb{E}(Z)| \leq \mathbb{E}(|Z|)$

or, $X_1, \dots, X_n$ de loi $\mathcal{G}(\theta)$ donc $X_i \geq 1$ $\forall i = 1, \dots, n$ et $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \geq \frac{1}{n}\sum_{i=1}^{n} 1 = 1$

Alors

$$|B_{\hat{\theta}}(\theta)| \leq \mathbb{E}\left(\frac{|1 - \theta \overline{X}_n|}{\overline{X}_n}\right) \leq \mathbb{E}\left(\frac{|1 - \theta \overline{X}_n|}{1}\right) = \mathbb{E}(|1 - \theta \overline{X}_n|)$$

Rappel CS

si Z, Z' 2 v.a.

$\mathbb{E}(ZZ')^2 \leq \mathbb{E}(Z^2)\mathbb{E}(Z'^2)$

On applique CS avec $Z = |1-\theta \bar{X}_n|$ et $Z' = 1$:

$$|B_{\hat{\theta}}(\theta)| \leq \mathbb{E}(|1-\theta \bar{X}_n|) \leq \sqrt{\mathbb{E}((1-\theta\bar{X}_n)^2) \times \underbrace{\mathbb{E}(1^2)}_{=1}} = \sqrt{\mathbb{E}((1-\theta\bar{X}_n)^2)}$$

Or,

$$\mathbb{E}\left((1-\theta\bar{X}_n)^2\right) = \theta^2\, \mathbb{E}\left(\left(\frac{1}{\theta} - \bar{X}_n\right)^2\right)$$

On a $\mathbb{E}(X_1) = \frac{1}{\theta}$ car $X_1 \sim \mathcal{G}(\theta)$ (cf feuille TD1) d'où comme $X_i \sim \mathcal{G}(\theta)$:

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

Alors

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

$$= \frac{\theta^2}{n^2}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{\theta^2}{n^2}\sum_{i=1}^{n}\operatorname{var}(X_i), \text{ car } X_1,\dots,X_n \text{ indépendantes}$$

$$= \frac{\theta^2}{n^2}\sum_{i=1}^{n}\operatorname{var}(X_1) \text{ car } X_1,\dots,X_n \text{ de même loi}$$

$$= \frac{\theta^2}{n}\operatorname{var}(X_1) = \frac{\cancel{\theta^2}}{n}\,\frac{1-\theta}{\cancel{\theta^2}} \text{ car } X_1 \sim \mathcal{G}(\theta), \text{ cf feuille TD1}$$

Alors, $|B_{\hat{\theta}}(\theta)| \leq \sqrt{\mathbb{E}\left((1-\theta\bar{X}_n)^2\right)} \leq \sqrt{\frac{1-\theta}{n}} \leq \frac{1}{\sqrt{n}}$ car $\theta \in]0,1[$.

3 - Risque quadratique

Définition (Risque quadratique) Soient $\theta \in \Theta$ et $X_1, \dots, X_n$ un échantillon iid de la loi $\mathcal{L}_\theta$. Le risque quadratique en θ de l'estimateur $\hat{g} = \hat{g}(X_1, \dots, X_n)$ est la quantité

$$R_{\hat{g}}(\theta) = \mathbb{E}\left((\hat{g} - g(\theta))^2\right) \qquad g(\theta) : \text{paramètre d'intérêt.}$$

On veut des estimateurs qui ont un risque quadratique le + faible possible. Si c'est le cas, alors $\hat{g} \simeq g(\theta)$ en moyenne quadratique, c'est donc un "bon" estimateur. Cela nous permet de classer les estimateurs :

Définition Soient $\theta \in \Theta$ et $X_1, \dots, X_n$ un échantillon iid de la loi $\mathcal{L}_\theta$. On note $\hat{g}_1 = \hat{g}_1(X_1, \dots, X_n)$ et $\hat{g}_2 = \hat{g}_2(X_1, \dots, X_n)$ deux estimateurs de $g(\theta)$. Si

$$R_{\hat{g}_1}(\theta) \leq R_{\hat{g}_2}(\theta) \quad \forall \theta \in \Theta,$$

on dit que $\hat{g}_1$ est préférable à $\hat{g}_2$

Exemples

① Dans le modèle de Bernoulli décrit par le n-échantillon $X_1, \dots, X_n$ iid de loi $\mathcal{B}(\theta)$, $\theta \in]0,1[$, avec θ comme paramètre d'intérêt.

On s'intéresse aux 2 estimateurs $\hat{\theta}_1 = \bar{X}_n$ et $\hat{\theta}_2 = X_1$. On a

$$R_{\hat{\theta}_1}(\theta) = \mathbb{E}\left((\hat{\theta}_1 - \theta)^2\right) = \mathbb{E}\left((\bar{X}_n - \theta)^2\right). \text{ Or, } \mathbb{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \theta$$

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

$$\underset{\text{indép.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}(X_i) \underset{\text{m loi}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}(X_1) = \frac{1}{n^2}\sum_{i=1}^{n}\theta(1-\theta) = \frac{\theta(1-\theta)}{n}$$

On $\lim_{n \to +\infty} R_{\hat{\theta}_1}(\theta) = 0$: la qualité de l'estimateur $\hat{\theta}_1$ augmente à mesure que l'on reproduit les expériences aléatoires

$$R_{\hat{\theta}_2}(\theta) = \theta(1-\theta) \text{ car } \hat{\theta}_2 = \hat{\theta}_1 \text{ si } n = 1$$

Alors, si $n \geq 2$, $R_{\hat{\theta}_1}(\theta) \leq R_{\hat{\theta}_2}(\theta)$ $\forall \theta \in]0,1[$ donc $\hat{\theta}_1$ est préférable à $\hat{\theta}_2$.

(2) Dans le modèle statistique décrit par le n-échantillon $X_1, \dots, X_n$ iid de loi $\mathcal{P}(\theta)$ avec $\theta > 0$ et θ comme paramètre d'intérêt.

Calculer le risque quadratique de $\bar{X}_n$.

$$R_{\bar{X}_n}(\theta) = \mathbb{E}\left((\bar{X}_n - \theta)^2\right)$$

Rappel si $Z \sim \mathcal{P}(\theta)$,

$\mathbb{E}(Z) = \text{var}(Z) = \theta$

$\text{var}(Z) = \mathbb{E}\left((Z - \mathbb{E}(Z))^2\right)$

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \underset{X_i \sim \mathcal{P}(\theta)}{=} \frac{1}{n}\sum_{i=1}^{n} \theta = \theta$$

Donc
$$R_{\bar{X}_n}(\theta) = \mathbb{E}\left((\bar{X}_n - \mathbb{E}(\bar{X}_n))^2\right) = \text{var}(\bar{X}_n) = \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) \underset{\text{indep.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i)$$

$$\underset{\text{m loi}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_1) = \frac{n}{n^2}\text{var}(X_1) = \frac{1}{n}\text{var}(X_1) = \frac{\theta}{n} \text{ car } X_1 \sim \mathcal{P}(\theta).$$

(3) Dans le modèle stat décrit par le n-échantillon $X_1, \dots, X_n$ iid de loi $\mathcal{G}(\theta)$, $\theta \in]0,1[$, et $g(\theta) = \frac{1}{\theta}$ comme paramètre d'intérêt.

$\bar{X}_n$ est un estimateur naturel de $\frac{1}{\theta}$ $\left(\text{car } \bar{X}_n \overset{\text{LGN}}{\approx} \mathbb{E}(X_1) = \frac{1}{\theta}\right)$

$$R_{\bar{X}_n}(\theta) = \mathbb{E}\left(\left(\bar{X}_n - \frac{1}{\theta}\right)^2\right) = \mathbb{E}\left(\left(\bar{X}_n - \mathbb{E}(\bar{X}_n)\right)^2\right)$$

$$= \mathrm{var}(\bar{X}_n) = \dots = \frac{\mathrm{var}(X_1)}{n} = \frac{1-\theta}{n\theta^2} \quad \text{car } X_1 \sim \mathcal{G}(\theta).$$

<u>Rappel</u> si $Z \sim \mathcal{G}(\theta)$,

$\mathbb{E}(Z) = \frac{1}{\theta}$ et

$\mathrm{var}(Z) = \frac{1-\theta}{\theta^2}$

Si maintenant le paramètre d'intérêt est θ. Alors $\hat{\theta} = \frac{1}{\bar{X}_n}$ est un estimateur naturel. Evaluons le risque de $\hat{\theta}$:

$$R_{\hat{\theta}}(\theta) = \mathbb{E}\left(\left(\hat{\theta} - \theta\right)^2\right) = \mathbb{E}\left(\left(\frac{1}{\bar{X}_n} - \theta\right)^2\right)$$

: majorer ce risque pour la prochaine séance de CM.