

Chapitre 3. Estimation

Pierre Duchesne

February 2, 2017

Soit X un caractère étudié dans une population \mathcal{P} . On suppose que l'on connaît la loi de X mais à un (ou plusieurs) paramètres près.

Exemple: Soit X le nombre de pièces défectueuses produite par une machine dans un lot de n pièces. Dans un lot de n pièces sélectionnées au hasard et avec remise, nous avons ici que

$$X \sim \text{Bin}(n, \theta).$$

Ici, c'est θ qui est le paramètre inconnu.

Plus formellement, la loi de X appartient à une famille de loi:

$$\{F_\theta\}_{\theta \in \Theta},$$

où Θ est appelé **espace paramétrique**.

Dans l'exemple précédent, $X \sim \text{Bin}(n, \theta)$ et $\Theta = [0, 1]$.

La famille de lois correspondante est alors:

$$\{\text{Bin}(n, \theta)\}_{\theta \in [0,1]}.$$

Exemple: La taille d'un individu peut être modélisée par une loi normale. Soit X : la taille d'un individu adulte demeurant à Montréal. On aura donc que $X \sim \mathcal{N}(\mu, \sigma^2)$. La famille de lois est alors:

$$\{\mathcal{N}(\mu, \sigma^2)\}_{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+}$$

Dans ce cas-ci, $\Theta = \mathbb{R} \times \mathbb{R}^+$.

Dans cet exemple, le paramètre est le vecteur $\theta = (\mu, \sigma^2)^\top$, qui est de dimension deux, ou encore bivarié.

En clair, la loi de X sera entièrement connue si on connaît la valeur du paramètre inconnu θ (ou θ si le paramètre est de dimension supérieure à un).

La théorie de l'estimation a pour principal objectif d'approcher numériquement la valeur du paramètre inconnu.

On distingue deux formes d'estimation:

- (i) **Estimation ponctuelle**: attribuer une valeur unique à θ (ou encore à θ).
- (ii) **Estimation par intervalle de confiance**: attribuer un ensemble de valeurs à θ .

Règle générale, l'estimation se fait au moyen de l'information puisée dans un échantillon.

Soit X un caractère étudié et soit $\mathcal{E} : X_1, \dots, X_n$ un échantillon associé à X .

Définition: Un estimateur est une statistique

$$T = t(X_1, \dots, X_n)$$

telle que son support (c'est-à-dire l'ensemble des valeurs qu'elle est susceptible de prendre) soit situé dans Θ l'espace paramétrique.

Exemple: Considérons $X \sim N(\theta, 1)$. L'espace paramétrique du paramètre est ici $\Theta = \mathbb{R}$. Soit $\mathcal{E} : X_1, \dots, X_n$ un échantillon associé à X .

Les fonctions suivantes sont des exemples de statistiques:

- a) $T_1 = X_1$;
- b) $T_2 = (X_1 + X_2)/2$;
- c) $T_3 = (\min\{X_1, \dots, X_n\} + \max\{X_1, \dots, X_n\})/2$;
- d) $T_4 = \text{médiane}\{X_1, \dots, X_n\}$;
- e) $T_5 = \frac{1}{n-2} \sum_{i=2}^n X_i + X_1^{2010} - X_n^{2010}$;
- f) $T_6 = \frac{1}{n} \sum_{i=1}^n X_i$.

Il y a plusieurs estimateurs possibles dans un problème donné. C'est au moyen de critères que l'on va départager les estimateurs.

Critère 1: Absence de biais

Dans ce qui suit, $\hat{\theta}$ sera une notation pour un estimateur du paramètre θ .

Définition: Un estimateur $\hat{\theta}$ de θ est dit sans biais si:

$$E(\hat{\theta}) = \theta, \forall \theta \in \Theta.$$

Ainsi, cette condition d'absence de biais assure que, à la longue, les situations où l'estimateur surestime et sous-estime θ vont s'équilibrer, de sorte que la valeur estimée sera correcte *en moyenne*.

Une interprétation physique à l'absence de biais est que l'espérance mathématique est une mesure de position, de sorte que le centre de gravité de $\hat{\theta}$ est le paramètre que l'on veut estimer.

Définition: Le biais d'un estimateur $\hat{\theta}$ de θ est la quantité:

$$b_{\theta}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

On note que le signe a un sens ici. Un estimateur possèdera un biais positif si en moyenne il sur-estime la vraie valeur θ .

Il aura un biais négatif sinon.

Un estimateur $\hat{\theta}$ est sans biais pour θ si et seulement si $b_{\theta}(\hat{\theta}) = 0, \forall \theta \in \Theta$.

Théorème: Soit $\mathcal{E} : X_1, \dots, X_n$ un échantillon associé à X , tel que $E(X) = \mu$ et $\text{var}(X) = \sigma^2$.

Alors \bar{X} est un estimateur sans biais pour μ . En effet:

$$E(\bar{X}) = \frac{1}{n} E \left(\sum_{i=1}^n X_i \right) = \frac{1}{n} \left\{ \sum_{i=1}^n E(X_i) \right\} = \mu.$$

Si μ est connu, $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ est sans biais pour σ^2 .
En effet:

$$E(\hat{\sigma}^2) = \frac{1}{n} E \left\{ \sum_{i=1}^n (X_i - \mu)^2 \right\} = \frac{1}{n} \left\{ \sum_{i=1}^n \sigma^2 \right\} = \sigma^2.$$

Finalement, si μ est inconnu, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ est sans biais pour σ^2 .

Pour montrer le résultat facilement, on note que l'on peut réécrire $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$, avec $Y_i = X_i - \mu$. Donc, sans perte de généralité, on peut supposer que $E(X_i) = 0$.

On suppose $E(X_i) = 0$. Or $S^2 = \frac{1}{n-1} (\sum_{i=1}^n X_i^2 - n\bar{X}^2)$. Ainsi $E(\sum_{i=1}^n X_i^2) = n\sigma^2$. De même $E(\bar{X}^2) = \text{var}(\bar{X}) = \sigma^2/n$.

Donc:

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

Remarque: On note que le résultat de ce théorème est valide quelque soit la loi (en particulier nous n'avons pas présumé la normalité).

Définition: L'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ est:

$$EQM_{\theta}(\hat{\theta}) = E \left\{ (\hat{\theta} - \theta)^2 \right\}$$

On note que la littérature anglophone note MSE dont le sens est *mean squared error*.

Un résultat important exprime l'EQM en fonction de la variance et du biais carré:

$$EQM_{\theta}(\hat{\theta}) = \text{var}_{\theta}(\hat{\theta}) + b_{\theta}^2(\hat{\theta}).$$

Proposition: $EQM_{\theta}(\hat{\theta}) = E \left\{ (\hat{\theta} - \theta)^2 \right\} = \text{var}_{\theta}(\hat{\theta}) + b_{\theta}^2(\hat{\theta})$.

Pour montrer le résultat, on utilise l'artifice suivant:

$$\begin{aligned}(\hat{\theta} - \theta)^2 &= \left\{ \hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta \right\}^2, \\ &= \left\{ \hat{\theta} - E(\hat{\theta}) \right\}^2 + 2 \left\{ \hat{\theta} - E(\hat{\theta}) \right\} \left\{ E(\hat{\theta}) - \theta \right\} + \left\{ E(\hat{\theta}) - \theta \right\}^2 \\ &= \left\{ \hat{\theta} - E(\hat{\theta}) \right\}^2 + 2 \left\{ \hat{\theta} - E(\hat{\theta}) \right\} b_{\theta}(\hat{\theta}) + b_{\theta}^2(\hat{\theta}),\end{aligned}$$

On prend alors l'espérance mathématique de chaque côté.

On note que $E \left\{ (\hat{\theta} - E(\hat{\theta}))^2 \right\} = \text{var}_{\theta}(\hat{\theta})$; en effet, par définition, $\text{var}(X) = E \{ (X - E(X))^2 \}$.

De même $E \left\{ \hat{\theta} - E(\hat{\theta}) \right\} = E(\hat{\theta}) - E(\hat{\theta}) = 0$.

Plus l'EQM d'un estimateur est petit, plus l'estimateur est considéré précis.

On note que si $\hat{\theta}$ est un estimateur sans biais de θ , alors:

$$EQM_{\theta}(\hat{\theta}) = \text{var}_{\theta}(\hat{\theta}).$$

Ceci fait ressortir tout l'intérêt que l'on peut porter à la variance d'un estimateur, comme critère de précision d'une classe composée de certains estimateurs.

L'application simultanée des critères 1 et 2 peut être conflictuelle.

Exemple: Supposons que $X \sim N(\mu, \sigma^2)$, où μ est inconnue et considérons deux estimateurs de σ^2 :

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

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

On se rappelle que $E(S^2) = \sigma^2$ et ayant présumé la normalité:

$$(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2;$$

Donc $\text{var}(S^2) = 2\sigma^4/(n-1)$. En clair:

$$EQM_{\sigma^2}(S^2) = \frac{2\sigma^4}{n-1} + 0^2 = \frac{2\sigma^4}{n-1}.$$

D'un autre côté, $S^{*2} = \frac{n-1}{n} S^2$. Ainsi

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

Le biais de S^{*2} est donc: $-\frac{1}{n} \sigma^2$.

De plus

$$\text{var}(S^{*2}) = \left(\frac{n-1}{n}\right)^2 \text{var}(S^2) = 2 \left(\frac{n-1}{n}\right)^2 \frac{\sigma^4}{n-1} = 2 \frac{n-1}{n^2} \sigma^4.$$

Donc:

$$EQM_{\sigma^2}(S^{*2}) = 2 \frac{n-1}{n^2} \sigma^4 + \frac{1}{n^2} \sigma^4 = \frac{2n-1}{n^2} \sigma^4.$$

En résumé, Pour le premier estimateur S^2 : $E(S^2) = \sigma^2$ et

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

Pour le second estimateur S^{*2} : $E(S^{*2}) = (1 - \frac{1}{n})\sigma^2$ et

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

Selon le premier critère:

$$E(\hat{\theta}) = \theta, \forall \theta \in \Theta.$$

Ainsi, selon ce critère, l'estimateur de choix de σ^2 devrait être S^2 .

Rappel: Pour le premier estimateur S^2 : $E(S^2) = \sigma^2$ et

$EQM_{\sigma^2}(S^2) = \frac{2\sigma^4}{n-1}$ Pour le second estimateur S^{*2} :

$E(S^{*2}) = (1 - \frac{1}{n})\sigma^2$ et $EQM_{\sigma^2}(S^{*2}) = \frac{2n-1}{n^2}\sigma^4$.

On note que $\frac{2n-1}{n^2} < \frac{2}{n-1}$ car

$$2n - 1 < 2n \Rightarrow (2n - 1)/n^2 < 2/n < 2/(n - 1).$$

Donc

$$EQM_{\sigma^2}(S^{*2}) < EQM_{\sigma^2}(S^2).$$

Le second critère est:

$$EQM_{\theta}(\hat{\theta}) = E \left\{ (\hat{\theta} - \theta)^2 \right\}$$

que l'on cherche à minimiser.

Ainsi l'estimateur de choix devrait être S^{*2} .

A priori, il n'y a aucune raison de privilégier un critère plutôt qu'un autre.

Cependant, une grande attention a été accordée à l'absence de biais des estimateurs.

Puisque l'estimateur S^{*2} n'est pas dans la classe des estimateurs sans biais, il devrait être mis de côté et on devrait privilégier l'estimateur S^2 .

Une approche consiste à rechercher le meilleur estimateur possible parmi ceux qui sont sans biais.

Définition: Un estimateur $\hat{\theta}$ qui est sans biais pour θ est dit à variance minimale si et seulement si pour tout autre estimateur $\hat{\theta}^*$ qui est sans biais pour θ la relation suivante est satisfaite:

$$\text{var}(\hat{\theta}) \leq \text{var}(\hat{\theta}^*).$$

Ainsi, parmi la classe des estimateurs sans biais de θ , l'estimateur sans biais à variance minimale est celui qui a la plus petite variance.

Exemple: Considérons X un caractère tel que $E(X) = \mu$ et $\text{var}(X) = \sigma^2$. Soit $\mathcal{E} : X_1, \dots, X_n$ un échantillon associé à X . Soient:

$$\begin{aligned}\hat{\mu}_1 &= X_1, \\ \hat{\mu}_2 &= (X_1 + X_2)/2, \\ \hat{\mu}_3 &= \frac{1}{n} \sum_{i=1}^n X_i.\end{aligned}$$

Ces trois estimateurs sont dans la classe des estimateurs sans biais:

$$\begin{aligned}E(\hat{\mu}_1) &= E(X_1) = \mu, \\ E(\hat{\mu}_2) &= E\{(X_1 + X_2)/2\} = \{E(X_1) + E(X_2)\}/2 = (\mu + \mu)/2 = \mu, \\ E(\hat{\mu}_3) &= \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu.\end{aligned}$$

Or $\text{var}(\hat{\mu}_1) = \sigma^2$; $\text{var}(\hat{\mu}_2) = \sigma^2/2$, $\text{var}(\hat{\mu}_3) = \sigma^2/n$; $\hat{\mu}_3$ est le gagnant!

Critère 3: Convergence

Tout estimateur $\hat{\theta}$ est une fonction de la taille de l'échantillon et une notation plus juste est en fait $\hat{\theta}_n$. Soit

$$\hat{\theta}_n = t(X_1, \dots, X_n).$$

On peut alors considérer une suite d'estimateurs $\{\hat{\theta}_n\}_{n=1}^{\infty}$.

Définition: Soit une suite d'estimateurs $\{\hat{\theta}_n\}_{n=1}^{\infty}$. On dira que $\hat{\theta}_n$ est convergent pour θ si:

$$\hat{\theta}_n \xrightarrow{P} \theta.$$

Rappel: Si $E(X_n) \rightarrow c$ et si $\text{var}(X_n) \rightarrow 0$ alors:

$$X_n \xrightarrow{P} c.$$

Nous avons alors que si $\hat{\theta}_n$ est sans biais pour θ , et si $\text{var}(\hat{\theta}_n) \rightarrow 0$, alors $\hat{\theta}_n$ est un estimateur convergent pour θ . C'est immédiat car $E(\hat{\theta}_n) = \theta$, et $\text{var}(\hat{\theta}_n) \rightarrow 0$ impliquent que

$$\hat{\theta}_n \xrightarrow{P} \theta.$$

Théorème: Soit $\mathcal{E} : X_1, \dots, X_n$ un échantillon associé à X , tel que $E(X) = \mu$, $\text{var}(X) = \sigma^2$ et $E\{(X - \mu)^4\} = \mu_4$. Alors

- (i) \bar{X} est un estimateur convergent pour μ ;
- (ii) Si μ est connu, $\hat{\sigma}^2$ est un estimateur convergent pour σ^2 .
- (iii) Si μ est inconnu, alors S^2 est un estimateur convergent pour σ^2 .

En effet, on a déjà vu que \bar{X} est sans biais pour μ . Or

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0, \quad n \rightarrow \infty.$$

De même $\hat{\sigma}^2$ est sans biais de σ^2 . Or

$$\text{var}(\hat{\sigma}^2) = \frac{\mu_4 - \sigma^4}{n} = \frac{\mu_4 - \mu_2^2}{n} \rightarrow 0, \quad n \rightarrow \infty$$

Finalement, S^2 est sans biais pour σ^2 . Cependant

$$\text{var}(S^2) = \left(\frac{n}{n-1}\right)^2 \left\{ \frac{\mu_4 - \mu_2^2}{n} - 2\frac{\mu_4 - 2\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} \right\} \rightarrow 0,$$

lorsque $n \rightarrow \infty$.

Soient deux estimateurs $\hat{\theta}$ et $\tilde{\theta}$, d'un paramètre θ . Si les deux estimateurs $\hat{\theta}$ et $\tilde{\theta}$ sont sans biais, comparer leurs EQM revient à comparer les variances. On se rappelle de ce que l'on a vu selon le Critère 2:

$$EQM_{\theta}(\hat{\theta}) = \text{var}_{\theta}(\hat{\theta}) + b_{\theta}^2(\hat{\theta}).$$

Étant donné deux estimateurs sans biais $\hat{\theta}$ et $\tilde{\theta}$, d'un paramètre θ , l'**efficacité relative** de $\hat{\theta}$ par rapport à $\tilde{\theta}$ est:

$$eff(\hat{\theta}, \tilde{\theta}) = \frac{\text{var}(\tilde{\theta})}{\text{var}(\hat{\theta})}.$$

Si l'efficacité est inférieure à un, $\hat{\theta}$ a plus grande variance que $\tilde{\theta}$.

Interprétation de l'efficacité

Soient deux estimateurs sans biais $\hat{\theta}$ et $\tilde{\theta}$. Supposons que

$$\text{var}(\hat{\theta}) = \frac{a}{n}, \quad \text{var}(\tilde{\theta}) = \frac{b}{n}.$$

On calcule

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{var}(\tilde{\theta})}{\text{var}(\hat{\theta})} = \frac{b}{a}.$$

Supposons que l'efficacité est inférieure à un, donc que $b/a < 1$ et $\tilde{\theta}$ est préférable. Supposons $a = 2, b = 1$.

Si $n = 100$ lorsque l'on utilise $\tilde{\theta}$, il faudra 200 observations pour $\hat{\theta}$ pour obtenir la même variance que celle de $\tilde{\theta}$.

L'efficacité est le rapport des tailles échantillonnales nécessaires pour que les deux méthodes aient la même variance.

Soit X un caractère étudié tel que

- a) $p(x; \theta)$ est sa fonction de masse si X est discrète;
- b) $f(x; \theta)$ est sa fonction de densité si X est continue.

Définition: La quantité d'information de Fisher est notée $I_X(\theta)$ et est définie par:

$$I_X(\theta) = \begin{cases} E \left\{ \left(\frac{\partial}{\partial \theta} \log p(X; \theta) \right)^2 \right\}, & \text{Si } X \text{ est discrète,} \\ E \left\{ \left(\frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \right\}, & \text{Si } X \text{ est continue.} \end{cases}$$

La variance de tout estimateur sans biais ne peut jamais être inférieure à une certaine borne. Cette borne est appelée borne de Cramér-Rao.

Théorème: Inégalité de Cramér-Rao

Soit $\hat{\theta}_n = t(X_1, \dots, X_n)$ un estimateur sans biais de θ . Alors

$$\text{var}(\hat{\theta}_n) \geq \frac{1}{nI_X(\theta)},$$

où $I_X(\theta)$ est la quantité d'information de Fisher.

Définition: L'efficacité d'un estimateur sans biais $\hat{\theta}$ est définie par:

$$e_{\theta}(\hat{\theta}) = \frac{1}{n \text{var}(\hat{\theta}) I_X(\theta)}$$

On note que cette définition est le rapport de $\frac{1}{n I_X(\theta)}$ et de la variance $\text{var}(\hat{\theta})$.

On a toujours $e_{\theta}(\hat{\theta}) \leq 1$. On dira que $\hat{\theta}$ est efficace lorsque $e_{\theta}(\hat{\theta}) = 1$.

Ainsi un estimateur efficace atteint la borne de Cramér-Rao et est donc à variance minimale.

Remarque: La réciproque n'est pas vraie en général. Il existe des situations où aucun estimateur sans biais n'atteint la borne de Cramér-Rao.