Statistique et construction d'estimateur

Statistique :

Une statistique est la réduction de l'information apportée par un échantillon.

Définition : Soit $(\mathbb{R}^n, \mathcal{B}^n, \{P_\theta^{\otimes n}, \theta \in \Theta\})$ un modèle d'échantillonnage (des objets produits). On appelle statistique la variable aléatoire $T(X) = T(X_1, \dots, X_n)$ définie de $\mathbb{R}^n$ dans [illegible] par $x = (x_1, \dots, x_n) \longmapsto T(x) = T(x_1, \dots, x_n)$

Exemples :

$$T_1(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$T_2(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\}$$

$$T_3(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty, x]}$$

$T_3(X)$ est l'estimation de la fonction de répartition $F_\theta(x)$, appelée aussi fonction de répartition empirique $T_3(x) = \hat{F}_n(x)$.

Si θ est un paramètre l'estimateur $\hat{\theta}$ de θ est une statistique.

Exemple : Soit le modèle d'échantillonnage $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \{U_{[0,\theta]}, \theta \in \mathbb{R}_+^*\})$ où $U_{[0,\theta]}$ est la loi uniforme sur $[0, \theta]$, $\theta \in \mathbb{R}_+$.

On désire estimer θ, pour cela on a plusieurs estimateurs :

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{(moyenne empirique)}$$

$$\hat{\theta}_2 = X_{(n)} = \max\{X_1, \dots, X_n\}$$

$$\hat{\theta}_3 = X_{(n)} + \frac{X_{(n)}}{n} = \frac{(n+1)}{n} X_{(n)}$$

$$\hat{\theta}_4 = X_{(1)} + X_{(n)} \quad (X_{(1)} = \min\{X_1, \dots, X_n\})$$

Deux questions se posent :

Question ① : Comment construire les estimateurs.

Question ② : Comment comparer ces estimateurs pour choisir le meilleur.

Construction d'un estimateur

Estimation empirique :

Soit $(\mathbb{R}^n, \mathcal{B}^{n}, (P_\theta, \theta \in \mathbb{R}^k))$ un modèle d'échantillonnage, $X_1, \dots, X_n$ une suite de v.a. i.i.d. de même loi que X, de fonction de répartition $F_\theta(x)$ on note par la suite

$m_\theta = E_\theta(X)$ et $\sigma_\theta^2 = Var_\theta(X)$ (la moyenne et la variance de X).

Définition 4.1

On appelle moyenne empirique la statistique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

et est donc un estimateur de m_θ : $\hat{m}_\theta = \bar{X}_n$

Définition 4.2

Par analogie on définit le moment empirique d'ordre p par

$\hat{m}_p(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i^p$ qui est un estimateur du moment d'ordre p $m_p(\theta)$

Donc on a par conséquence et puisque $\sigma_\theta^2 = E_\theta(X^2) - (E_\theta(X))^2$

$\hat{\sigma}_\theta^2 = \hat{m}_2(\theta) - (\hat{m}_1(\theta))^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = S_n^2$

où $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ donc $\hat{\sigma}_\theta^2 = S_n^2$

S_n^2 est donc l'estimateur empirique de la variance.

Proposition 4.1.

$E_\theta(\bar{X}_n) = m_\theta$ ($\bar{X}_n$ est un estimateur sans biais de m_θ).

$Var_\theta(\bar{X}_n) = \frac{\sigma_\theta^2}{n}$

Cette méthode d'estimation des moments (empiriques) permet aussi d'estimer la fonction de répartition

$F_\theta(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{+\infty} 1_{]-\infty, x]}(t) f(t)\,dt = E(1_{[X \le x]})$

Posant $Y = 1_{[X \le x]}$ et $Y_i = 1_{[X_i \le x]}$ on a alors ; puisque $E(1_{[X \le x]}) = E(Y)$

$\widehat{E(Y)} = \frac{1}{n}\sum_{i} Y_i \Rightarrow \hat{F}(x) = \widehat{E(1_{[X \le x]})} = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \le x]}$

$\hat{F}_X(x) = F_n(x) = \frac{1}{n}\sum_{i} 1_{[X_i \le x]}$.

2) Méthode de substitution :

Principe de la méthode :

Si on dispose d'un estimateur $\hat{\theta}$ de θ et si $g(\cdot)$ est une fonction continue alors l'estimateur de $g(\theta)$ est $g(\hat{\theta})$:

$$\widehat{g(\theta)} = g(\hat{\theta}).$$

Exemple : L'estimateur de l'écart-type σ connaissant S_n^2 est :

$\hat{\theta} = S_n$: $\hat{\sigma} = \frac{1}{\sqrt{n}}\sqrt{\sum (x_i-\bar{x})^2}$. $(g(x)=\sqrt{x})$

3) Méthode des moments :

Principe de la méthode :

Supposons que l'application h définie de $\Theta \subset \mathbb{R}^p$ dans $h(\Theta) \subset \mathbb{R}^p$ est bijective et continue et soit φ une application de $\mathcal{X}$ dans $\mathbb{R}^p$ telle que $E(\varphi(X))$ existe et que :

$$h(\theta) = E_\theta(\varphi(X)) \quad \text{pour tout } \theta \in \Theta.$$

La méthode des moments consiste à estimer θ par $\hat{\theta}$:

$$\hat{\theta} = h^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\right)$$

En effet :

Si $h(\theta) = E_\theta(\varphi(X))$ alors $\widehat{h(\theta)} = \widehat{E_\theta(\varphi(X))}$

On pose $Y = \varphi(X)$ et $Y_i = \varphi(X_i)$ et comme on a :

$$\widehat{E_\theta(Y)} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{alors} \quad \widehat{h(\theta)} = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)$$

et par la méthode de substitution on sait que $\widehat{h(\theta)} = h(\hat{\theta})$

donc $h(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)$ ou $\hat{\theta} = h^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right)$

Exemple : Soit X de loi exponentielle de paramètre λ inconnu

$X \sim \mathcal{E}(\lambda)$.

On sait que $E_\lambda(X) = \frac{1}{\lambda}$ donc : $E_\lambda(\varphi(X)) = h(\lambda)$ permet d'écrire $\varphi(X) = x$ et $h(x) = \frac{1}{x}$, et que $h^{-1}(y) = \frac{1}{y}$.

D'où $\hat{\lambda} = h^{-1}(\varphi(X)) = \frac{1}{\frac{1}{n}\sum X_i} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}_n}$

Exemple :

Soit $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \{N(\mu,\sigma^2), \mu\in\mathbb{R}, \sigma^2>0\})$, et φ et h tels que
$\varphi(x) = (x, x^2)$; $h(\mu,\sigma^2) = (\mu, \mu^2+\sigma^2)$; $h(\theta) = E_\theta(\varphi(X)) = E_\theta(X, X^2)$

$h(\mu,\sigma^2) = (E_\theta X, E_\theta(X^2)) = (\mu, \mu^2+\sigma^2)$ $\quad (E(X^2) = \mu^2+\sigma^2)$

Pour déterminer $(\hat\mu, \hat\sigma^2) = h^{-1}(\hat E(X), \hat E(X^2))$ on a besoin de connaître $h^{-1}(u,v)$: $(u,v) = h(\mu,\sigma^2)$

donc $\mu = u$ et $\sigma^2 = v - u^2$

Comme $\hat E(X) = \frac{1}{n}\sum X_i$ et $\hat E(X^2) = \frac{1}{n}\sum_{i=1}^n X_i^2$ on obtient :

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

$$\hat\mu = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{et} \quad \hat\sigma^2 = \frac{1}{n}\sum X_i^2 - \frac{1}{n^2}\left(\sum X_i\right)^2$$

b) Méthode du maximum de vraisemblance.

Définition 4.1 :

Soit $(\mathcal{X}, \mathcal{E}, \mathcal{P} = \{P_\theta, \theta\in\Theta\})$ un modèle statistique et X une variable aléatoire de densité (vraisemblance) $f_\theta(x) = L(x,\theta)$. L'estimateur $\hat\theta(X)$ du maximum de vraisemblance (E.M.V, MLE) est tel que $L(x,\hat\theta) \geq L(x,\theta)$, $\forall\theta\in\Theta$.

Pour le modèle d'échantillonnage on a $L(x_1,\dots,x_n,\theta)$ la vraisemblance de l'échantillon.

Si la fonction est concave elle admet un unique maximum (L'estimateur du maximum de vraisemblance est obtenu en annulant la dérivée dans le cas unidimensionnel, et le gradient dans le cas multidimensionnel).

Remarque : Puisque la fonction logarithme est croissante le maximum de $\log(L(x,\theta))$ est le même que celui de $L(x,\theta)$.

$$\hat\theta = \arg\max_\theta(L(x,\theta)) = \arg\max_\theta(\log(L(x,\theta)))$$

Exemple 1:

X est de Bernoulli de paramètre "p".

$f_p(x) = p^x (1-p)^{1-x}$

$L(x_1, \dots, x_n, p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$

$\log(L(x_1 \dots x_n p)) = \sum_{i=1}^{n} [x_i \log p + (1-x_i) \log(1-p)]$

$\varphi(p) = \log(L(x_1, \dots, x_n, p)) = \log(p) \left(\sum_{i=1}^{n} x_i\right) + \log(1-p) \left(\sum_{i=1}^{n} (1-x_i)\right)$

$\varphi'(p) = \dfrac{\sum x_i}{p} - \dfrac{\sum(1-x_i)}{1-p}$ et $\varphi''(p) = -\dfrac{\sum x_i}{p^2} - \dfrac{\sum(1-x_i)}{(1-p)^2} < 0$

$\varphi'(\hat{p}) = 0 \Leftrightarrow \dfrac{\sum x_i}{\hat{p}} - \dfrac{\sum(1-x_i)}{(1-\hat{p})} = 0$

$(1-\hat{p}) \sum_{i=1}^{n} x_i - \hat{p} \sum_{i=1}^{n} (1-x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i - \hat{p}\, n = 0$

$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Exemple 2:

X est de loi normale $\{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$, $\theta = (\mu, \sigma^2)$

$L(x_1, \dots, x_n, (\mu, \sigma^2)) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$\varphi(\mu, \sigma^2) = \log(L(x_1 \dots x_n, (\mu, \sigma^2))) = -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$

(1) $\dfrac{\partial \varphi(\mu, \sigma^2)}{\partial \mu} = \dfrac{1}{2\sigma^2} \sum_{i=1}^{n} (-2(x_i - \mu)) = 0$

(2) $\dfrac{\partial \varphi(\mu, \sigma^2)}{\partial \sigma} = -\dfrac{n}{\sigma} + \dfrac{1}{\sigma^3} \sum (x_i - \mu)^2 = 0$

Ce qui donne alors:

$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\hat{\mu} = \bar{X}_n$,

et $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X}_n)^2 = S_n$

Proposition: (Invariance du maximum de vraisemblance).

Si $\hat{\theta}$ est un estimateur du maximum de vraisemblance de θ dans un modèle $\{P_\theta, \theta \in \Theta\}$, et si g est une application bijective

...ensemble de Θ dans $g(\Theta)$, alors $g(\hat\theta)$ est l'estimateur du maximum de vraisemblance de $g(\theta)$ pour $\{P_{g(\theta)}, g(\theta) \in g(\Theta)\}$.

...alités d'un estimateur.

Estimateur convergent

Définition ...:

Un estimateur $\hat\theta_n = T(X_n)$ du paramètre θ est faiblement convergent consistant si $\hat\theta_n$ converge en probabilité sous $\{P_\theta, \theta \in \Theta\}$ vers θ

$$\hat\theta_n \xrightarrow{P} \theta, \text{ quand } n \to \infty$$

$\forall \varepsilon > 0$, $P(|\hat\theta_n - \theta| > \varepsilon) \longrightarrow 0$, $n \to \infty$ dans le cas unidimensionnel

ou $P(\|\hat\theta_n - \theta\| > \varepsilon) \longrightarrow 0$, $n \to \infty$ dans le cas multidimensionnel

Loi faible des grands nombres:

Si $X_1, \dots, X_n$ est une suite de v.a. i.i.d. de moyenne μ alors:

$\bar X_n = \frac{1}{n}\sum X_i$ (moyenne empirique) est estimateur de μ ($\hat\mu = \bar X_n$)

...: ① $\frac{1}{n}\sum X_i \xrightarrow{P} \mu$, $n \to \infty$ (inégalité de Bienaymé-Tchebychev)

...: $\widehat{E(X)} = \bar X_n$; $\widehat{E(X^2)} = \frac{1}{n}\sum X_i^2$ et on a:

② $\frac{1}{n}\sum X_i^2 \xrightarrow{P} E(X^2)$ (inégalité de Markov)

De ① et ② on a: $\hat\sigma^2 = S_n = \frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2 \xrightarrow{P} \sigma^2$

S_n est donc un estimateur consistant de σ^2.

On a la convergence forte de $\hat\theta_n$ vers θ si $\hat\theta_n \xrightarrow{p.s.} \theta$, $n \to \infty$

On a la convergence dans L^2 (en moyenne quadratique) si $\hat\theta_n \xrightarrow{L^2} \theta$, $n \to \infty$

Estimateur sans biais

Définition 5.2.1: Le biais de l'estimateur $\hat\theta_n$ de θ est la fonction définie sur Θ par: $B_{\hat\theta_n}(\theta) = E_\theta(\hat\theta) - \theta$, pour tout $\theta \in \Theta$ si bien sûr $E_\theta(\hat\theta)$ existe.

Un estimateur $\hat\theta_n$ de θ est dit sans biais si $B_{\hat\theta_n}(\theta) = 0$ i.e. $E_\theta(\hat\theta) = \theta$.

Exemples :

$\hat{\theta}_n = \bar{X}_n$ est un estimateur sans biais de $\mu = E(X)$.

$E(\hat{\theta}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\mu = \mu$.

$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $E(S_n^2) = \frac{n-1}{n}\sigma^2$ (S_n^2 estimateur biaisé de σ^2).

(Voir la démonstration dans TD).

Définition 5.2

On dit qu'un estimateur $\hat{\theta}_n$ de θ est asymptotiquement sans biais si et seulement si :

$$\lim_{n\to\infty} B_\theta(\hat{\theta}_n) = 0 \quad \text{ou} \quad \lim_{n\to\infty} E(\hat{\theta}_n) = \theta.$$

Exemple :

$$E(S_n^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \; ; \quad B_\theta(S_n^2) = -\frac{\sigma^2}{n} \; ; \quad \lim_{n\to\infty} B_\theta(S_n^2) = 0$$

et $\lim_{n\to\infty} E(S_n^2) = \sigma^2$

Donc un estimateur biaisé peut être asymptotiquement sans biais.

Remarque :

Si on note par $S'^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

S'^2 est un estimateur sans biais de σ^2, appelé aussi estimateur de la variance modifiée.

$$E(S'^2) = E\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} E(S_n^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2.$$

6 Risque d'un estimateur

Un estimateur est une décision de prendre $\hat{\theta}_n$ au lieu de θ, le critère qui mesure l'erreur commise en remplaçant θ par $\hat{\theta}_n$ s'appelle fonction de coût (elle mesure l'écart entre $\hat{\theta}_n$ et θ).

Définition 6.1

On appelle fonction de coût (de perte) toute fonction $L(\cdot,\cdot)$ définie de $\Theta \times \Theta$ dans $\mathbb{R}_+$ mesurable par la première coordonnée $(\hat{\theta}, \theta) \longrightarrow L(\hat{\theta}, \theta)$, $L(\hat{\theta}, \theta)$ est une sorte de distance entre $\hat{\theta}$ et θ si $\Theta \subset \mathbb{R}$.

- Le coût absolu : $L(\hat{\theta},\theta) = |\hat{\theta}-\theta|$
- Le coût quadratique : $L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$
- Le coût pour tout h, $L(\hat{\theta},\theta) = h(\hat{\theta}-\theta)$

Si $\Theta \subset \mathbb{R}^d$ $(n \geq 1)$

- $L(\hat{\theta},\theta) = \|\hat{\theta}-\theta\|^2$ (la norme euclidienne).
- $L(\hat{\theta},\theta) = (\hat{\theta}-\theta)\cdot(\hat{\theta}-\theta)^t$ (vecteur transposé).

Remarque :

Tous les coûts sont symétriques et dépendent de n.

6.2 Risque

Noté $R(\hat{\theta}_n,\theta)$ le risque associé à la fonction de coût $L(\hat{\theta}_n,\theta)$ est défini par : $R(\hat{\theta}_n,\theta) = E(L(\hat{\theta}_n,\theta))$

Notant qu'en statistique le risque le plus utilisé est le risque lié au coût quadratique $L(\hat{\theta}_n,\theta) = (\hat{\theta}_n-\theta)^2$.

On obtient ainsi l'Erreur Quadratique Moyenne définie par $R(\hat{\theta}_n,\theta) = E[(\hat{\theta}_n-\theta)^2]$ notée MSE (Mean Square Error).

$$R(\hat{\theta}_n,\theta) = Var_\theta(\hat{\theta}_n) + b^2_{\hat{\theta}_n}(\theta).$$

Si l'estimateur est sans biais, le risque quadratique est égal à la variance : $R(\hat{\theta}_n,\theta) = Var_\theta(\hat{\theta}_n)$.

$R(\hat{\theta}_n,\theta)$ sert aussi à comparer la précision d'un estimateur.

Définition 6.2 :

Soit $\hat{\theta}_1$ et $\hat{\theta}_2$ deux estimateurs de θ, on dit que $\hat{\theta}_1$ est préférable à $\hat{\theta}_2$ si on a : $R(\hat{\theta}_1,\theta) \leq R(\hat{\theta}_2,\theta)$ pour tout $\theta \in \Theta$.

Définition 6.3 :

Un estimateur $\hat{\theta}_n$ est dit admissible s'il n'existe aucun estimateur qui lui soit préférable.