Statistique inférentielle

Estimation

A. Godichon-Baggioni

Objectifs

On dispose de n données $x_1, \ldots, x_n$ qui sont des mesures d'une variable quantitative, et plus précisément des réalisations de variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$.

Objectifs :

- Estimer une caractéristique θ de X_1 (moyenne, variance,...)
- Estimer un paramètre θ de X_1 si cette variable est paramétrée (loi de Bernoulli, normale, exponentielle,...)

Exemples :

- Lancer de pièce : On s'intéresse à l'estimation de $\theta = \mathbb{E}[X_1]$.
- Loi exponentielle : On s'intéresse à l'estimation du paramètre $\theta = (\mathbb{E}[X_1])^{-1}$.

I. Définitions

Statistique et estimateur

Dans ce qui suit, on considère un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ dépendant d'un paramètre $\theta \in \Theta$, où $\Theta \subset \mathbb{R}$.

Définition

Une statistique $T(\mathbf{X})$ est une fonction mesurable de l'échantillon $\mathbf{X}$ ne dépendant pas de θ (mais dépendant éventuellement de paramètres connus). Un estimateur de θ est une statistique $\hat{\theta} = \theta(\mathbf{X})$ destinée à approcher θ.

Exemple : Lancer de pièce

Attention ! Ne pas confondre estimateur et estimation !

Notation : Dans ce qui suit, on considère $\hat{\theta}_n$ un estimateur de θ.

Erreur quadratique moyenne

Définition

On suppose que θ est à valeurs dans $\Theta \subset \mathbb{R}$. L'erreur quadratique moyenne (ou risque quadratique) de l'estimateur $\hat{\theta}_n$ est définie pour tout $\theta \in \Theta$ par

$$EQM\left(\hat{\theta}_n, \theta\right) = \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right].$$

Remarquons que grâce à l'inégalité de Markov, pour tout $c > 0$,

$$\mathbb{P}\left[\left|\hat{\theta}_n - \theta\right| \geq c\right] \leq \frac{\mathrm{EQM}\left(\hat{\theta}_n, \theta\right)}{c^2}.$$

Biais d'un estimateur

Définition

On appelle biais d'un estimateur $\hat{\theta}_n$ de θ la quantité

$$B\left(\hat{\theta}_n, \theta\right) = \mathbb{E}\left[\hat{\theta}_n\right] - \theta.$$

1. *S'il est nul, on dit que l'estimateur est sans biais ou non biaisé*
2. *Si*
$$B\left(\hat{\theta}_n, \theta\right) \xrightarrow[n \to +\infty]{} 0,$$
on dit que l'estimateur est asymptotiquement sans biais.

Exemple : lancer de pièce. *L'estimateur $\hat{\theta}_n$ est sans biais.*

Décomposition Bais-Variance

Proposition

Soit $\hat{\theta}_n$ un estimateur de θ, on a

$$EQM\left(\hat{\theta}_n, \theta\right) = B\left(\hat{\theta}_n, \theta\right)^2 + \mathbb{V}\left[\hat{\theta}_n\right].$$

Exemple : lancer de pièce Comme $\hat{\theta}_n$ est un estimateur sans biais de θ, on a

$$\mathrm{EQM}\left(\hat{\theta}_n, \theta\right) = \mathbb{V}\left[\hat{\theta}_n\right] = \frac{\theta(1-\theta)}{n}.$$

CONVERGENCE, CONSISTANCE ET NORMALITÉ ASYMPTOTIQUE

Définition

On dit que l'estimateur $\hat{\theta}_n$ est

1. *convergent ou consistant si*

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta,$$

2. *fortement consistant si*

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{p.s} \theta,$$

3. *asymptotiquement normal si il existe $\sigma^2 > 0$ tel que*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right).$$

II. Estimation de la moyenne et de la variance

ESTIMATION DE LA MOYENNE

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées. Un estimateur naturel de la moyenne $\mu = \mathbb{E}\left[X_1\right]$ est donc

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

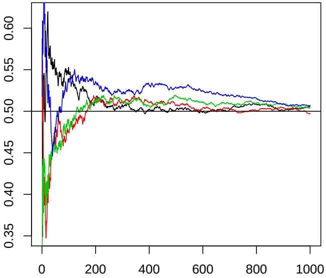
Proposition

1. *$\overline{X}_n$ est un estimateur sans biais et (fortement) consistant de μ.*
2. *Si $\sigma^2 = \mathbb{V}\left[X_1\right] < +\infty$, alors $\overline{X}_n$ est asymptotiquement normal et*

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right)$$

$$EQM\left(\overline{X}_n, \mu\right) = \frac{\sigma^2}{n}.$$

Estimation de la moyenne



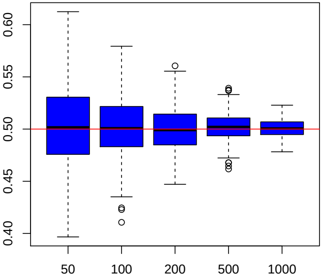


FIGURE – Evolution de $\overline{x}_n$ par rapport à n pour 4 échantillons (à gauche) et boxplots pour les $\overline{x}_n$ obtenus pour $n = 50, 100, 200, 500, 1000$ à l'aide de 4000 échantillons (à droite).

Estimation de la variance

Lorsque μ est connu, un estimateur naturel de la variance est

$$\hat{V}_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

$\hat{V}_n$ est sans biais et fortement consistant.
Si μ est inconnue, on a la variance empirique

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

ESTIMATEUR SANS BIAIS DE LA VARIANCE

Proposition

$\hat{\sigma}_n^2$ est un estimateur biaisié de σ^2 mais asymptotiquement sans biais. Plus précisément, on a

$$\mathbb{E}\left[\hat{\sigma}_n^2\right] = \frac{n-1}{n}\sigma^2.$$

Définition

L'estimateur sans biais de la variance S_n^2 est défini par

$$S_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right).$$

ESTIMATEUR SANS BIAIS DE LA VARIANCE

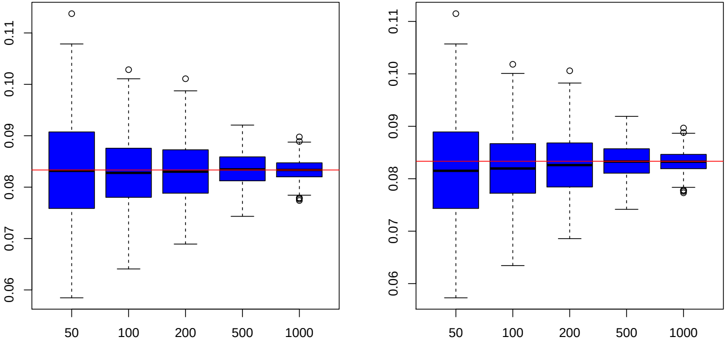


FIGURE – Boxplots pour l'estimateur non biaisé (à gauche) et biaisé (à droite)

Convergence

Proposition

Les estimateurs $\hat{\sigma}_n^2$ et S_n^2 sont consistants.

Proposition

Si X_1 admet un moment d'ordre 4, alors $\hat{\sigma}_n^2$ et S_n^2 sont asymptotiquement normaux, et on a

$$\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \tau^4 - \sigma^4\right)$$
$$\sqrt{n}\left(S_n^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \tau^4 - \sigma^4\right)$$

où

$$\tau^4 = \mathbb{E}\left[(X_1 - \mu)^4\right].$$

III. Méthode des moments

Méthode des moments

Proposition

Soit Θ un intervalle ouvert de $\mathbb{R}$ et $\theta \in \Theta$. Soit φ un C^1-difféomorphisme de Θ dans $\varphi(\Theta)$. Soit $\hat{\varphi}_n$ un estimateur consistant de $\varphi(\theta)$, alors $\hat{\theta}_n = \varphi^{-1}(\hat{\varphi}_n)$ est un estimateur consistant de θ, i.e

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta.$$

De plus, si $\hat{\varphi}_n$ est un estimateur asymptotiquement normal de $\varphi(\theta)$, i.e si il existe $\sigma^2 > 0$ tel que

$$\sqrt{n}\left(\hat{\varphi}_n - \varphi(\theta)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right),$$

et si $\varphi'(\theta) \neq 0$, alors $\hat{\theta}_n$ est un estimateur asymptotiquement normal de θ et

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{\left(\varphi'(\theta)\right)^2}\right).$$

Méthode des moments

La méthode des moments consiste à trouver un C^1-difféomorphisme φ et un moment k tel que $\mathbb{E}\left[X_1^k\right] = \varphi(\theta)$. Comme un estimateur de m_k est donné par

$$\hat{m}_{n,k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

on obtient l'estimateur

$$\hat{\theta}_n = \varphi^{-1}\left(\hat{m}_{n,k}\right)$$

EXEMPLES

Exemple : la loi uniforme. On considère des variables aléatoires i.i.d $X_1, \ldots, X_n$ suivant une loi uniforme sur $\left[0, \theta^2\right]$, avec $\theta > 0$, i.e de densité f_θ définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{1}{\theta^2} \mathbf{1}_{[0,\theta^2]}(x).$$

Exemple : la loi exponentielle. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d suivant une loi exponentielle de paramètre $\theta > 0$, i.e de densité f_θ définie pour tout $x \in \mathbb{R}$,

$$f_\theta(x) = \theta \exp\left(-\theta x\right) \mathbf{1}_{\mathbb{R}_+}(x).$$

Remarques

Remarque : Pour éviter les erreurs, il est plus judicieux (et cela revient à peu près au même) d'écrire θ comme une fonction de $\mathbb{E}\left[X^k\right]$.

Remarque : Attention ! Il peut arriver qu'une variable aléatoire n'admette pas de moment d'ordre 1. Il faut alors essayer d'être malin !

Exemple : Soit $\theta > 0$, on considère une variable aléatoire X de densité définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{\theta}{x^2}\mathbf{1}_{x \geq \theta}.$$

Fonction génératrice

Définition (Fonction génératrice)

Soit X une variable aléatoire, on appelle fonction génératrice G_X de X la fonction définie par

$$G_X(t) = \mathbb{E}\left[e^{tX}\right].$$

Attention ! La fonction génératrice n'est pas nécessairement définie, et encore moins pour tout t.

FONCTION GÉNÉRATRICE

Théorème

On suppose que la variable aléatoire X admet des moments de tout ordre, i.e pour tout $k \in \mathbb{N}$ $\mathbb{E}\left[X^k\right] < +\infty$, et que la série $\sum_{k\geq 0} \frac{t^k \mathbb{E}\left[X^k\right]}{k!}$ admet un rayon de convergence R. Alors pour tout $|t| < R$ on a

$$G_X(t) = \sum_{k\geq 0} \frac{t^k \mathbb{E}\left[X^k\right]}{k!}.$$

En particulier, pour tout $k \in \mathbb{N}$, on a

$$G_X^{(k)}(0) = \mathbb{E}\left[X^k\right].$$

Exmple : la loi géométrique. Soit $X \sim \mathcal{G}(p)$ avec $p \in (0,1)$. Alors pour tout $t < -\log(1-p)$,

$$G_X(t) = \frac{pe^t}{1-(1-p)e^t}.$$

IV. Méthode du Maximum de Vraisemblance

Notations

Dans ce qui suit, on note $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d.
On note f_θ la densité de X_1.

Rappel : Dans le cas discret, on a pour tout $x \in \mathbb{R}$,

$$f_\theta(x) = \mathbb{P}\left[X_1 = x\right].$$

(LOG)-VRAISEMBLANCE

Définition (Vraisemblance et log-vraisemblance)

La vraisemblance de $\mathbf{X}$ *est définie pour tout* $\theta \in \Theta$ *par*

$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} f_\theta\left(X_i\right).$$

La log-vraisemblance de $\mathbf{X}$ *est définie pour tout* $\theta \in \Theta$ *par*

$$l_{\mathbf{X}}(\theta) = \log\left(L_{\mathbf{X}}(\theta)\right) = \log\left(\prod_{i=1}^{n} f_\theta\left(X_i\right)\right).$$

Attention ! La (log-)vraisemblance est aléatoire.

Exemples : cas discret

Exemple 1 : loi de Bernoulli. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et $X_1 \sim \mathcal{B}(\theta)$ avec $\theta \in (0, 1)$.

Exemple 2 : loi de Poisson. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et X_1 suit une loi de Poisson de paramètre $\theta > 0$.

EXEMPLES : CAS CONTINU

Exemple 1 : loi exponentielle. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et $X_1 \sim \mathcal{E}(\theta)$ avec $\theta > 0$.

Exemple 2 : loi normale. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$.

Exemple 3 : On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et X_1 admet pour densité f_θ définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{\theta}{x^2} \mathbf{1}_{x \geq \theta}$$

avec $\theta > 0$.

Remarques

Cas discret La réalisation de la vraisemblance est la probabilité de d'obtenir cette réalisation de l'échantillon.

Cas continue La réalisation de la vraisemblance est la densité de $\mathbf{X}$ en la réalisation de l'échantillon.

Objectifs : Maximiser cette probabilité (ou densité).

Estimateur du maximum de vraisemblance

Définition (Estimateur du maximum de vraisemblance (EMV))

Le maximum de vraisemblance, si il existe, est un élément $\hat{\theta}_n$ *de* Θ *qui vérifie*

$$L_{\mathbf{X}}\left(\hat{\theta}_n\right) = \sup_{\theta \in \Theta} L_{\mathbf{X}}(\theta).$$

De manière équivalente, l'estimateur du maximum de vraisemblance, si il existe, vérifie

$$l_{\mathbf{X}}\left(\hat{\theta}_n\right) = \sup_{\theta \in \Theta} l_{\mathbf{X}}(\theta).$$

Remarques

Attention ! L'EMV est généralement aléatoire.

Attention ! Ni l'existence ni l'unicité de l'EMV ne sont assurées.

Remarque : A noter que si $\hat{\theta}_n$ est un EMV de θ, alors $\varphi\left(\hat{\theta}_n\right)$ est un EMV de $\varphi(\theta)$.

Remarque : On notera également

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} L_{\mathbf{X}}(\theta) \qquad \text{ou} \qquad \hat{\theta}_n = \arg\max_{\theta\in\Theta} l_{\mathbf{X}}(\theta)$$

EXEMPLES : CAS DISCRET

Exemple 1 : loi de Bernoulli. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et $X_1 \sim \mathcal{B}(\theta)$ avec $\theta \in (0, 1)$.

Exemple 2 : loi de Poisson. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et X_1 suit une loi de Poisson de paramètre $\theta > 0$.

EXEMPLES : CAS CONTINU

Exemple 1 : loi exponentielle. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et $X_1 \sim \mathcal{E}(\theta)$ avec $\theta > 0$.

Exemple 2 : loi normale. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$.

Exemple 3 : On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et X_1 admet pour densité f_θ définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{\theta}{x^2} \mathbf{1}_{x \geq \theta}$$

avec $\theta > 0$.

COMMENT TROUVER L'EMV

Soit l'EMV peut être donné "explicitement" par la vraisemblance, soit il est souvent plus facile de maximiser la log-vraisemblance. Pour cela, voici quelques options possibles :

- Dresser le tableau de variations de la log vraisemblance
- Chercher les zéros de la dérivée, i.e résoudre

$$l'_{\mathbf{x}}(\theta) = 0.$$

  - Vérifier qu'il(s) maximise(nt) la log-vraisemblance (étude de la concavité, ...)

UN CAS D'ÉTUDE

Exemple : Soit $X_1, \ldots, X_n$ des variables aléatoires i.i.d suivant une loi uniforme sur $[0, \theta]$, avec $\theta > 0$.

- Méthode des moments ?
- Méthode du maximum de vraisemblance ?

Objectifs

On s'intéresse à l'estimation d'un paramètre $\theta \in \Theta$ d'une variable aléatoire X, avec Θ un intervalle ouvert de $\mathbb{R}$. Les objectifs sont donc :

- Comparer différents estimateurs de θ.
- Savoir si on peut parler d'estimateur optimal.

V. Comparaison d’estimateurs

Comparaison des erreurs quadratiques moyennes

On rappelle qu'une façon de quantifier la qualité d'un estimateur $\hat{\theta}_n$ de θ est de considérer son risque quadratique

$$\mathrm{EQM}\left(\hat{\theta}_n, \theta\right) = \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right].$$

On considère que $\hat{\theta}_n$ est un meilleur estimateur que $\tilde{\theta}_n$ si

$$\forall \theta \in \Theta, \quad \mathrm{EQM}\left(\hat{\theta}_n, \theta\right) \leq \mathrm{EQM}\left(\tilde{\theta}_n, \theta\right).$$

Biais d'un estimateur

On donne trop souvent trop d'importance au biais d'un estimateur ! Débiaiser un estimateur ne donne pas forcément de meilleurs résultats !

Exemple : Soit $\theta > 0$ et $X \sim \mathcal{U}([0, \theta])$. Comparer les erreurs quadratiques moyennes des estimateurs suivants :

$$\hat{\theta}_n = X_{(n)}$$
$$\tilde{\theta}_n = \frac{n+1}{n} X_{(n)}$$
$$\theta_{\alpha,n} = \alpha X_{(n)}$$

avec $\alpha > 0$ choisi judicieusement.

Approche asymptotique

Le plus souvent, on dispose d'estimateurs $\hat{\theta}_n, \tilde{\theta}_n$ asymptotiquement normaux, i.e il existe σ_1^2 et σ_2^2 telles que

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma_1^2\right),$$
$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma_2^2\right).$$

Si $\sigma_1^2 \leq \sigma_2^2$, on choisira alors l'estimateur $\hat{\theta}_n$.