Chapitre III : Estimation ponctuelle

MATH-F-207

Davy Paindaveine

Bloc 2, Bachelier en sciences mathématiques
Université libre de Bruxelles

2023–2024

Contenu du chapitre

Contenu du chapitre

Introduction

Soit le modèle statistique paramétrique

$$\left(\mathbb{R}^n, \mathcal{B}^n, \mathcal{P}^{(n)} = \left\{ \mathrm{P}_\theta^{(n)} : \theta \in \Theta \subset \mathbb{R}^k \right\}\right),$$

où les mesures de probabilité $\mathrm{P}_\theta^{(n)}$ représentent les distributions possibles de $X^{(n)} = (X_1, \ldots, X_n)$.

Définition 1

Soit $g : \Theta \to \mathbb{R}^m$ une fonction mesurable.
Un estimateur $T(X^{(n)})$ de $g(\theta)$ est une statistique à valeurs dans $g(\Theta)$.

Dans le cas $g(\theta) = \theta$, on écrira parfois $\hat{\theta}$ plutôt que $T(X^{(n)})$.

Exemples d'estimateurs

Exemple électoral : soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathrm{Bern}(p)$, où $\theta = p \in \Theta = [0, 1] \subset \mathbb{R}$.

Un estimateur naturel de p est $T(X^{(n)}) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.
De même, un estimateur naturel de $g(p) = 3p^2 + 1$ est $T(X^{(n)}) = 3\bar{X}^2 + 1$.

Exemple du verre de bière : soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathcal{N}(\mu, \sigma^2)$, où $\theta = (\mu, \sigma^2)$ est un élement de $\Theta = \mathbb{R} \times \mathbb{R}_0^+ \subset \mathbb{R}^2$.

L'estimateur le plus classique de θ,

$$T(X^{(n)}) = \begin{pmatrix} \bar{X} \\ s^2 \end{pmatrix},$$

est celui qui apparaît dans le lemme de Fisher.

Paramètres d'intérêt et de nuisance

La fonction g permet aussi de restreindre le problème d'estimation à certaines composantes de θ seulement.

Dans l'exemple du verre de bière, la fonction $g(\theta) = g(\mu, \sigma^2) = \mu$ mène à l'estimation de la composante μ de θ. Dans ce cas, on dira que

- μ est le paramètre d'intérêt, et
- σ^2 est le paramètre de nuisance.

Estimateurs

Pour un problème d'estimation donné, il existe plusieurs estimateurs!

Exemple du bus :

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\text{Unif}([0, \theta])$, avec $\theta \in \Theta = \mathbb{R}_0^+$.
Les statistiques

$$X_{(n)}, \quad 2\bar{X}, \quad (n+1)X_{(1)}, \quad X_{(1)} + X_{(n)}, \quad \text{et} \quad 2X_1$$

constituent toutes des estimateurs de θ.

En fait, la classe de tous les estimateurs possibles est en général gigantesque (tous les coups sont permis!)

Objectifs du chapitre

$\rightsquigarrow$ Ce chapitre cherche à répondre aux questions suivantes :

- Comment comparer deux estimateurs?
- Quelles propriétés un “bon” estimateur doit-il avoir?
- Y a-t-il une limite à la précision d’un estimateur?
- Comment construire un estimateur atteignant la limite de précision? (si elle existe)
- Comment construire de façon “systématique” des estimateurs?

Contenu du chapitre

$T(X^{(n)})$ vs $T(x^{(n)})$

Soit $T(X^{(n)})$ un estimateur de $g(\theta)$.

Les données observées prennent la forme d'un n-uple $x^{(n)} = (x_1, \ldots, x_n)$, ce qui livre *l'estimation* $T(x^{(n)})$ *de* $g(\theta)$, laquelle est une valeur fixée dans $g(\Theta)$.

$\vee$
$\wedge$

L'estimateur $T(X^{(n)})$ est une variable (ou un vecteur) aléatoire, donc a une distribution échantillonnée. C'est sur celle-ci qu'on fondera les *critères d'estimation*.

Situation idéale

L'estimateur $T(X^{(n)})$ idéal estime $g(\theta)$ sans erreur, et ce, quel que soit θ:

$$\forall \theta \in \Theta, \quad \mathrm{P}_\theta^{(n)}[T(X^{(n)}) = g(\theta)] = 1$$

($\forall \theta \in \Theta$, sa distribution échantillonnée est donc dégénérée en $g(\theta)$ sous $\mathrm{P}_\theta^{(n)}$).

C'est tout à fait irréaliste!

Mais on peut espérer qu'on se retrouve dans ce cas idéal quand $n \to \infty$.
Ceci mène au concept suivant...

Estimateurs convergents

Définition 2

Considérons un modèle statistique paramétrique et soit $g : \Theta \to \mathbb{R}^m$ *mesurable. Soit* $T(X^{(n)})$ *un estimateur de* $g(\theta)$*. Alors*

(i) $T(X^{(n)})$ *est faiblement convergent* (*pour* $g(\theta)$)
$\Leftrightarrow \forall \theta \in \Theta$, $T(X^{(n)}) \xrightarrow{\mathrm{P}} g(\theta)$ *sous* $\mathrm{P}_\theta^{(n)}$ *quand* $n \to \infty$

(ii) $T(X^{(n)})$ *est fortement convergent* (*pour* $g(\theta)$)
$\Leftrightarrow \forall \theta \in \Theta$, $T(X^{(n)}) \xrightarrow{\mathrm{p.s.}} g(\theta)$ *sous* $\mathrm{P}_\theta^{(n)}$ *quand* $n \to \infty$.

Ces convergences doivent tenir $\forall \theta \in \Theta$!

Typiquement, l'estimateur trivial $T(X^{(n)}) = g(\theta_0)$, avec $\theta_0 \in \Theta$ fixé, ne satisfera ces convergences qu'en la valeur θ_0.

Exemple du verre de bière : convergence de $\bar{X}$ et s^2

Dans le modèle d'échantillonnage paramétrique gaussien, $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est fortement convergent pour μ (par la loi forte des grands nombres).

L'estimateur $\hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ de σ^2 est également fortement convergent, puisque

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

Exemple du verre de bière : convergence de $\bar{X}$ et s^2

Dans le modèle d'échantillonnage paramétrique gaussien, $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est fortement convergent pour μ (par la loi forte des grands nombres).

L'estimateur $\hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ de σ^2 est également fortement convergent, puisque

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\Big\{(X_i - \mu) + (\mu - \bar{X})\Big\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + 2(\mu - \bar{X})\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + (\mu - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2 \overset{\text{p.s.}}{\to} \mathrm{E}[(X_1 - \mu)^2] - 0^2 = \sigma^2.
\end{aligned}
$$

Exemple du bus : convergence de $2\bar{X}$

Dans le modèle d'échantillonnage paramétrique uniforme, la loi forte des grands nombres implique que, sous $\mathrm{P}_\theta^{(n)}$,

$$2\bar{X} \overset{\text{p.s.}}{\to} 2\mathrm{E}_\theta[X_1] = 2 \times \frac{\theta}{2} = \theta \quad \text{quand } n \to \infty.$$

L'estimateur $2\bar{X}$ de θ est donc fortement convergent.

Dans ces deux exemples, les estimateurs sont aussi faiblement convergents (puisque $\overset{\text{p.s.}}{\to} \Rightarrow \overset{\text{P}}{\to}$).

Exemple du bus : convergence de $X_{(n)}$

Parfois, il faut recourir à la définition pour établir (ou écarter) la convergence.

Exemple : considérons l'estimateur $X_{(n)}$ de θ dans l'exemple du bus.
Pour tout $\theta > 0$ et tout $\varepsilon > 0$,

$$\begin{aligned}
\mathrm{P}_\theta^{(n)}[|X_{(n)} - \theta| > \varepsilon] &= \mathrm{P}_\theta^{(n)}[X_{(n)} < \theta - \varepsilon] + \mathrm{P}_\theta^{(n)}[X_{(n)} > \theta + \varepsilon] \\
&= F_\theta^{X_{(n)}}(\theta - \varepsilon) + 0 \\
&= \begin{cases} (\frac{\theta-\varepsilon}{\theta})^n & \text{si } 0 < \varepsilon < \theta \\ 0 & \text{si } \varepsilon \geq \theta \end{cases} \\
&\to 0
\end{aligned}$$

quand $n \to \infty$. Donc $X_{(n)}$ est faiblement convergent pour θ.

Exercice : montrer que $\frac{n+1}{n} X_{(n)}$ est aussi faiblement convergent pour θ.

Exemple du bus : convergence de $(n+1)X_{(1)}$

Dans l'exemple du bus, considérons l'estimateur $\hat{\theta}^{(n)} = (n+1)X_{(1)}$ de θ.
Sous $\mathrm{P}_\theta^{(n)}$, sa fonction de répartition est donnée par

$$F_\theta^{\hat{\theta}^{(n)}}(x) = \mathrm{P}_\theta^{(n)}[(n+1)X_{(1)} \le x] = 1 - \mathrm{P}_\theta^{(n)}[X_{(1)} > \tfrac{x}{n+1}]$$

$$= 1 - \textstyle\prod_{i=1}^n \mathrm{P}_\theta[X_i > \frac{x}{n+1}] = 1 - \big(1 - F_\theta^{X_1}(\frac{x}{n+1})\big)^n$$

$$= \begin{cases} 0 & \text{si } x < 0 \\ 1 - \big(1 - \frac{x}{(n+1)\theta}\big)^n & \text{si } 0 \le x < (n+1)\theta \\ 1 & \text{si } x \ge (n+1)\theta. \end{cases}$$

Par la règle de L'Hospital, on obtient que, pour tout $\theta > 0$,

$$\forall x \in \mathbb{R}, \quad F_\theta^{\hat{\theta}^{(n)}}(x) \to (1 - e^{-x/\theta})\mathbb{I}[x \ge 0] \quad \text{quand } n \to \infty.$$

Donc sous $\mathrm{P}_\theta^{(n)}$, $\hat{\theta}^{(n)} \xrightarrow{\mathcal{L}} \mathrm{Exp}(\theta)$, la loi exponentielle de moyenne θ.
Par suite, $\hat{\theta}^{(n)}$ n'est pas faiblement convergent pour θ (pourquoi?)

Remarque sur la convergence

La convergence est la qualité la plus fondamentale dont doit jouir un estimateur (un estimateur non convergent doit être écarté!)

Ceci permet d'écarter des estimateurs de la forme $T(X_1, \ldots, X_r)$ qui, indépendamment de n, n'utilisent qu'un nombre fixé r d'observations, comme

$$\hat{\theta} = X_1 + X_2$$

dans l'exemple du bus.

Des estimateurs utilisant toutes les observations et a priori raisonnables, comme (on le verra)

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

(toujours dans l'exemple du bus), sont également écartés grâce à ce critère.

Estimateurs exhaustifs : introduction

Soit le modèle statistique paramétrique

$$\left(\mathbb{R}^n, \mathcal{B}^n, \mathcal{P}^{(n)} = \left\{ \mathrm{P}_\theta^{(n)} : \theta \in \Theta \subset \mathbb{R}^k \right\}\right).$$

Un estimateur qui, indépendamment de n, n'utilise qu'un nombre fixé r d'observations ne fait pas usage de toute l'information pertinente sur θ qui réside dans $X^{(n)}$ et ne peut pas être convergent.

Comment préciser mathématiquement ce qu'est une statistique qui exploite toute l'information pertinente sur θ?

Estimateurs exhaustifs : définition

Dans un modèle statistique paramétrique, considérons une statistique $T(X^{(n)})$. Notons $\mathcal{T}^{(n)}$ l'ensemble de ses valeurs possibles.

Définition 3

$T(X^{(n)})$ est exhaustive $\Leftrightarrow$ pour tout $B \in \mathcal{B}^n$ et pour tout $t \in \mathcal{T}^{(n)}$,

$$\mathrm{P}_\theta^{(n)}[X^{(n)} \in B \,|\, T(X^{(n)}) = t]$$

ne dépend pas de θ.

Estimateurs exhaustifs : interprétation

Il existe toujours $B \in \mathcal{B}^n$ tel que $\mathrm{P}_\theta^{(n)}[X^{(n)} \in B]$ dépend de θ
$\rightsquigarrow$ Le fait de savoir si l'événement $[X^{(n)} \in B]$ s'est réalisé ou pas donne de l'information sur θ.

Mais une fois donnée l'information qu'une statistique exhaustive $T(X^{(n)})$ a pris la valeur t, les probabilités des événements $[X^{(n)} \in B]$ ne dépendent plus de θ.
$\rightsquigarrow$ Savoir si $[X^{(n)} \in B]$ s'est réalisé ou pas n'apporte plus d'information sur θ.

Autrement dit, quand on donne la valeur d'une statistique exhaustive, on donne toute l'information dans $X^{(n)}$ pertinente pour faire de l'inférence sur θ.

Exemple trivial

La statistique $T(X^{(n)}) = X^{(n)}$ est toujours exhaustive.

En effet: pour tout $B \in \mathcal{B}^n$ et pour toute valeur possible $x^{(n)}$ de $X^{(n)}$,

$$\mathrm{P}_\theta^{(n)}[X^{(n)} \in B \,|\, X^{(n)} = x^{(n)}] = \mathbb{I}[x^{(n)} \in B],$$

qui ne dépend pas de θ.

Mais, clairement, le concept d'exhaustivité n'aura d'intérêt que quand $T(X^{(n)})$ n'est pas en bijection avec $X^{(n)}$.

Statistique exhaustive : exemple électoral

Un exemple non trivial:

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathrm{Bern}(p)$, où $p \in [0,1] \subset \mathbb{R}$. Alors, la statistique $T(X^{(n)}) = \sum_{i=1}^{n} X_i$ est exhaustive.

Statistique exhaustive

Il n'est que très rarement possible d'utiliser la définition d'exhaustivité pour déterminer si une statistique est exhaustive ou non, ce qui rend utile la condition nécessaire et suffisante (CNS) d'exhaustivité que nous allons présenter.

Cette CNS utilise le concept de fonction de vraisemblance...

Fonction de vraisemblance

Soit un modèle statistique paramétrique. Soit $x^{(n)} \in \mathbb{R}^n$ fixé.

Définition 4

Si, pour tout $\theta \in \Theta$, la loi $\mathrm{P}_\theta^{(n)}$ de $X^{(n)}$ est discrète (respectivement, continue), alors la fonction de vraisemblance est

$$\begin{aligned} L_\cdot^{(n)}(x^{(n)}) : \quad \Theta &\to \mathbb{R} \\ \theta &\mapsto L_\theta^{(n)}(x^{(n)}) = \mathrm{P}_\theta^{(n)}[X^{(n)} = x^{(n)}] \\ (\textit{resp.},\ \theta &\mapsto L_\theta^{(n)}(x^{(n)}) = f_\theta^{X^{(n)}}(x^{(n)})), \end{aligned}$$

où $f_\theta^{X^{(n)}}(\cdot)$ est la densité de $X^{(n)}$ sous $\mathrm{P}_\theta^{(n)}$.

Fonction de vraisemblance : le cas d'échantillonnage

Pour un modèle d'échantillonnage paramétrique,

$$L_\theta^{(n)}(x^{(n)}) = \mathrm{P}_\theta^{(n)}[X^{(n)} = x^{(n)}] = \prod_{i=1}^{n} \mathrm{P}_\theta[X_1 = x_i] \quad \text{dans le cas discret}$$

et

$$L_\theta^{(n)}(x^{(n)}) = f_\theta^{X^{(n)}}(x^{(n)}) = \prod_{i=1}^{n} f_\theta^{X_1}(x_i) \quad \text{dans le cas continu.}$$

Intuitivement, plus la valeur de $L_\theta^{(n)}(x^{(n)})$ grande, plus la loi $P_\theta^{(n)}$ est en adéquation avec l'observation $x^{(n)}$ qui a été faite.

Fonction de vraisemblance : exemple électoral

Soit le vecteur aléatoire $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\text{Bern}(p)$, où $\theta = p \in \Theta = [0, 1] \subset \mathbb{R}$. La fonction de vraisemblance est

$$\begin{aligned} L_p^{(n)}(x^{(n)}) &= \prod_{i=1}^{n} \mathrm{P}_p[X_1 = x_i] \\ &= \prod_{i=1}^{n} \left\{ p^{x_i} (1-p)^{1-x_i} \mathbb{I}[x_i \in \{0,1\}] \right\} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} \mathbb{I}[x_1, \ldots, x_n \in \{0,1\}]. \end{aligned}$$

Fonction de vraisemblance : l'exemple du verre de bière

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathcal{N}(\mu, \sigma^2)$, où $\theta = (\mu, \sigma^2)$ appartient à $\Theta = \mathbb{R} \times \mathbb{R}_0^+ \subset \mathbb{R}^2$. La fonction de vraisemblance est

$$
\begin{aligned}
L_\theta^{(n)}(x^{(n)}) &= \prod_{i=1}^{n} f_\theta^{X_1}(x_i) \\
&= \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\tfrac{1}{2\sigma^2}(x_i - \mu)^2\right) \right\} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^{n} (x_i - \mu)^2\right).
\end{aligned}
$$

Critère d'exhaustivité

Théorème 5 (critère de factorisation de Neyman–Fisher)

Soit un modèle statistique paramétrique. La statistique $T(X^{(n)})$ *est exhaustive* $\Leftrightarrow$ *pour tout* $\theta \in \Theta$*, la vraisemblance* $L_\theta^{(n)}$ *se factorise sous la forme*

$$L_\theta^{(n)}(X^{(n)}) = g_\theta(T(X^{(n)}))\, h(X^{(n)}) \qquad \mathrm{P}_\theta^{(n)}\text{-}p.s.,$$

au sens où $\mathrm{P}_\theta^{(n)}[\{x^{(n)} : L_\theta^{(n)}(x^{(n)}) = g_\theta(T(x^{(n)}))\, h(x^{(n)})\}] = 1.$

Critère de factorisation : l'exemple électoral

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\text{Bern}(p)$, où $\theta = p \in \Theta = [0,1] \subset \mathbb{R}$. La fonction de vraisemblance se factorise en

$$L_p^{(n)}(x^{(n)}) = \underbrace{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}_{g_p(\sum_{i=1}^n x_i)} \times \underbrace{\mathbb{I}[x_1, \ldots, x_n \in \{0,1\}]}_{h(x^{(n)})},$$

ce qui montre que $T(X^{(n)}) = \sum_{i=1}^n X_i$ est exhaustive.

Remarque : Il découle du critère de factorisation que, si $T(X^{(n)})$ est exhaustive et H est une bijection, alors $H(T(X^{(n)}))$ est aussi exhaustive.

En particulier, $T(X^{(n)}) = \bar{X}^{(n)}$ est aussi exhaustive ci-dessus.

Critère de factorisation : l'exemple du verre de bière

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathcal{N}(\mu, \sigma^2)$, où $\theta = (\mu, \sigma^2)$ appartient à $\Theta = \mathbb{R} \times \mathbb{R}_0^+ \subset \mathbb{R}^2$. La fonction de vraisemblance se réécrit

$$
\begin{aligned}
L_\theta^{(n)}(x^{(n)}) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n (x_i-\mu)^2\right) \\
&= \underbrace{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n x_i^2 + \tfrac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \tfrac{n\mu^2}{2\sigma^2}\right)}_{g_{(\mu,\sigma^2)}\left(\begin{pmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{pmatrix}\right)} \times \underbrace{1}_{h(x^{(n)})}
\end{aligned}
$$

Donc $T(X^{(n)}) = \begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i^2 \end{pmatrix}$ et $T(X^{(n)}) = \begin{pmatrix} \bar{X} \\ s^2 \end{pmatrix}$ sont exhaustives.

Critère de factorisation : l'exemple du bus

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\text{Unif}([0, \theta])$, où $\theta \in \Theta = \mathbb{R}_0^+$.

Alors la fonction de vraisemblance est

$$
\begin{aligned}
L_\theta^{(n)}(x^{(n)}) &= \prod_{i=1}^{n} f_\theta^{X_1}(x_i) = \prod_{i=1}^{n} \left\{ \frac{1}{\theta} \mathbb{I}[0 \leq x_i \leq \theta] \right\} \\
&= \frac{1}{\theta^n} \mathbb{I}[0 \leq x_1, \ldots, x_n \leq \theta] \\
&= \frac{1}{\theta^n} \mathbb{I}[0 \leq x_{(1)}, \ldots, x_{(n)} \leq \theta] \\
&= \frac{1}{\theta^n} \mathbb{I}[x_{(1)} \geq 0, x_{(n)} \leq \theta] \\
&= \frac{1}{\theta^n} \mathbb{I}[x_{(1)} \geq 0] \mathbb{I}[x_{(n)} \leq \theta].
\end{aligned}
$$

Critère de factorisation : l'exemple du bus (2)

La factorisation

$$L_\theta^{(n)}(x^{(n)}) = \underbrace{\frac{1}{\theta^n}\mathbb{I}[0 \leq x_{(1)}, \ldots, x_{(n)} \leq \theta]}_{g_\theta(x_{(1)},\ldots,x_{(n)})} \times \underbrace{1}_{h(x^{(n)})}$$

montre que $T(X^{(n)}) = (X_{(1)}, \ldots, X_{(n)})$ est exhaustive. Par ailleurs,

$$L_\theta^{(n)}(x^{(n)}) = \underbrace{\frac{1}{\theta^n}\mathbb{I}[x_{(1)} \geq 0]\mathbb{I}[x_{(n)} \leq \theta]}_{g_\theta(x_{(1)},x_{(n)})} \times \underbrace{1}_{h(x^{(n)})}$$

indique que

$$T(X^{(n)}) = \begin{pmatrix} X_{(1)} \\ X_{(n)} \end{pmatrix}$$

est encore une statistique exhaustive.

Statistique exhaustive minimale

En fait, puisque

$$L_\theta^{(n)}(x^{(n)}) = \underbrace{\frac{1}{\theta^n}\mathbb{I}[x_{(n)} \leq \theta]}_{g_\theta(x_{(n)})} \times \underbrace{\mathbb{I}[x_{(1)} \geq 0]}_{h(x^{(n)})},$$

$T(X^{(n)}) = X_{(n)}$ est aussi une statistique exhaustive.

En fait, $T(X^{(n)}) = X_{(n)}$ est une statistique exhaustive *minimale*.

Définition 6

Une statistique exhaustive $T(X^{(n)})$ est minimale si, pour toute autre statistique exhaustive $S(X^{(n)})$, il existe une fonction ℓ telle que $T(X^{(n)}) = \ell(S(X^{(n)}))$.

Estimateur sans biais

Dans l'exemple du bus, l'estimateur $T(X^{(n)}) = X_{(n)}$ sous-estime toujours θ:

$$\mathrm{P}_\theta^{(n)}[T(X^{(n)}) < \theta] = 1 \text{ pour tout } \theta \in \Theta,$$

ce qui, bien sûr, implique une sous-estimation en moyenne: $\mathrm{E}_\theta^{(n)}[T(X^{(n)})] < \theta$.

Définition 7

Dans un modèle statistique paramétrique, (i) l'estimateur $T(X^{(n)})$ de $g(\theta)$ est sans biais (ou non biaisé) si et seulement si, pour tout $\theta \in \Theta$,

$$\mathrm{E}_\theta^{(n)}[T(X^{(n)})] = g(\theta).$$

(ii) L'estimateur $T(X^{(n)})$ de $g(\theta)$ est asymptotiquement sans biais (ou asymptotiquement non biaisé) si et seulement si, pour tout $\theta \in \Theta$,

$$\mathrm{E}_\theta^{(n)}[T(X^{(n)})] \to g(\theta) \quad \textit{quand } n \to \infty.$$

Biais : l'exemple du bus

Dans l'exemple du bus,

$$\mathrm{E}_\theta^{(n)}[X_{(n)}] = \int_{-\infty}^{+\infty} x f_\theta^{X_{(n)}}(x)\,dx = \int_0^\theta x \times \frac{n x^{n-1}}{\theta^n}\,dx$$
$$= \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n\theta}{n+1} < \theta,$$

de sorte que l'estimateur $X_{(n)}$ de θ est effectivement biaisé.

Par contre,

$$\mathrm{E}_\theta^{(n)}[X_{(n)}] \to \theta \quad \text{quand } n \to \infty,$$

ce qui montre que cet estimateur est asymptotiquement non biaisé.

Biais : l'exemple du bus

Il est possible de "débiaiser" l'estimateur en le remplaçant par

$$T(X^{(n)}) = \frac{n+1}{n} X_{(n)}.$$

Exercices : vérifier que $2\bar{X}$, $X_{(1)} + X_{(n)}$, $(n+1)X_{(i)}/i$, $i = 1, \ldots, n$, sont des estimateurs sans biais de θ.

En particulier, l'estimateur $(n+1)X_{(1)}$ de θ est sans biais...Bien qu'il ne soit pas faiblement convergent!

Biais d'un estimateur

Définition 8

Le biais d'un estimateur $T(X^{(n)})$ de $g(\theta)$, sous $\mathrm{P}_\theta^{(n)}$, est le réel

$$b_\theta^{(n)}(T(X^{(n)})) = \mathrm{E}_\theta^{(n)}[T(X^{(n)})] - g(\theta).$$

L'estimateur $T(X^{(n)})$ de $g(\theta)$ est sans biais $\Leftrightarrow b_\theta^{(n)} = 0$ pour tout $\theta \in \Theta$
Il est asymptotiquement sans biais $\Leftrightarrow b_\theta^{(n)} \to 0$ pour tout $\theta \in \Theta$.

Dans l'exemple du bus, le biais de $X_{(n)}$ en θ vaut

$$b_\theta^{(n)}(X^{(n)}) = \mathrm{E}_\theta[X_{(n)}] - \theta = \frac{n\theta}{n+1} - \theta = -\frac{\theta}{n+1}.$$

Biais: exemple du verre de bière

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathcal{N}(\mu, \sigma^2)$, où $\theta = (\mu, \sigma^2)$ appartient à $\Theta = \mathbb{R} \times \mathbb{R}_0^+ \subset \mathbb{R}^2$.

Pour $g(\theta) = \mu$, l'estimateur $\bar{X}$ est sans biais. Pour $g(\theta) = \sigma^2$, l'estimateur s^2 vérifie

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - 2\bar{X}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \bar{X}^2$$
$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \bar{X}^2.$$

Ceci livre

$$\mathrm{E}_\theta^{(n)}[s^2] = \mathrm{E}_\theta^{(n)}[X_1^2] - \mathrm{E}_\theta^{(n)}[\bar{X}^2]$$
$$= \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{(n-1)\sigma^2}{n}.$$

Estimateur non biaisé de σ^2

Le biais de s^2 en $\theta = (\mu, \sigma^2)$ est donc

$$b_\theta^{(n)} = \mathrm{E}_\theta^{(n)}[s^2] - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n} \; (< 0) \cdot$$

Donc s^2 est un estimateur biaisé, mais asymptotiquement sans biais.

Clairement, l'estimateur

$$S^2 \stackrel{\text{def}}{=} \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

est sans biais pour σ^2.

Transformations non affines

Le non-biais ne résiste pas de façon générale aux transformations non-affines.

Considérons l'estimation de $\sigma = \sqrt{\sigma^2}$ dans l'exemple du verre de bière.
Un estimateur naturel est $T(X^{(n)}) = S = \sqrt{S^2}$. Cet estimateur vérifie

$$0 < \mathrm{Var}_\theta^{(n)}[S] = \mathrm{E}_\theta^{(n)}[S^2] - \left(\mathrm{E}_\theta^{(n)}[S]\right)^2 = \sigma^2 - \left(\mathrm{E}_\theta^{(n)}[S]\right)^2,$$

ce qui livre $\left(\mathrm{E}_\theta^{(n)}[S]\right)^2 < \sigma^2$, donc aussi $\mathrm{E}_\theta^{(n)}[S] < \sigma$.

$\rightsquigarrow$ S est un estimateur biaisé de σ

Remarque : par contre, le non-biais résiste aux transformations affines.

Existence d'un estimateur sans biais

Il n'existe pas toujours d'estimateur sans biais.

Soit le problème de l'estimation de $g(p) = p^2$ dans l'exemple électoral fondé sur une unique observation X_1 de loi $\mathrm{Bern}(p)$, où $\theta = p \in \Theta = [0, 1] \subset \mathbb{R}$.

Un estimateur $T(X_1)$ de p^2 est non biaisé si et seulement si

$$p^2 = \mathrm{E}_p[T(X_1)] = T(1)p + T(0)(1 - p)$$

pour tout $p \in [0, 1]$. Quels que soient $T(0)$ et $T(1)$, cette équation a au plus deux solutions. Il n'existe donc pas d'estimateur sans biais!

Erreur quadratique moyenne

La propriété de non-biais, seule, est loin d'être suffisante pour qu'un estimateur soit satisfaisant (un estimateur sans biais peut ne pas être convergent!)

Un estimateur sera d'autant plus satisfaisant qu'il fournit, en moyenne, une erreur d'estimation petite...

Définition 9

Soient un modèle statistique paramétrique et une fonction mesurable $g : \Theta \to \mathbb{R}$. *L'erreur quadratique moyenne de l'estimateur* $T(X^{(n)})$ *de* $g(\theta)$, *sous* $\mathrm{P}_\theta^{(n)}$, *est*

$$\mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] = \mathrm{E}_\theta^{(n)}[(T(X^{(n)}) - g(\theta))^2].$$

En anglais, on parle de *mean squared error*, ce qui explique la notation.

Décomposition du MSE

Le MSE se décompose en termes de biais et de variance:

$$\begin{aligned}
\mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] &= \mathrm{E}_\theta^{(n)}[(T(X^{(n)}) - g(\theta))^2] \\
&= \left(\mathrm{E}_\theta^{(n)}[T(X^{(n)}) - g(\theta)]\right)^2 + \mathrm{Var}_\theta^{(n)}[T(X^{(n)}) - g(\theta)] \\
&= \left(\mathrm{E}_\theta^{(n)}[T(X^{(n)})] - g(\theta)\right)^2 + \mathrm{Var}_\theta^{(n)}[T(X^{(n)})] \\
&= (b_\theta^{(n)}(T(X^{(n)})))^2 + \mathrm{Var}_\theta^{(n)}[T(X^{(n)})].
\end{aligned}$$

C'est cette décomposition qui fait qu'on préfère souvent le MSE à d'autres mesures de précision, comme par exemple l'*erreur absolue moyenne*

$$\mathrm{MAE}_\theta^{(n)} = \mathrm{E}_\theta^{(n)}[|T(X^{(n)}) - g(\theta)|]$$

(MAE tient ici pour *mean absolute error*).

Lien avec la convergence faible

Si

- $T(X^{(n)})$ est asymptotiquement sans biais pour $g(\theta)$
- $\mathrm{Var}_\theta[T(X^{(n)})] \to 0$ pour tout θ,

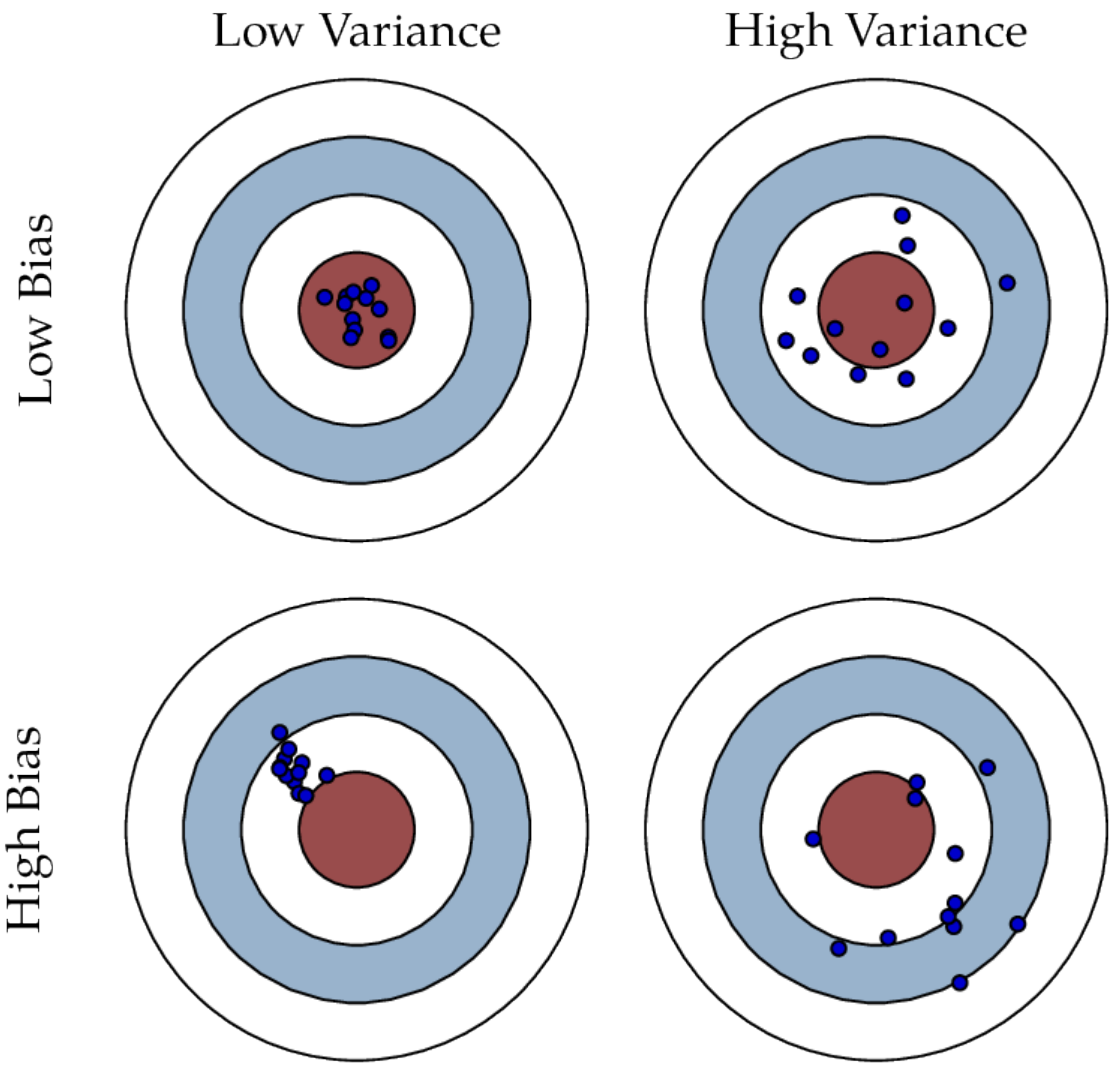
alors $T(X^{(n)})$ est faiblement convergent.

En effet, la décomposition ci-dessus implique alors que, pour tout θ,

$$\mathrm{E}_\theta^{(n)}[(T(X^{(n)}) - \theta)^2] \to 0 \text{ quand } n \to \infty,$$

ce qui établit que $T(X^{(n)}) \xrightarrow{L_2} g(\theta)$ sous $\mathrm{P}_\theta^{(n)}$, donc aussi que $T(X^{(n)}) \xrightarrow{\mathrm{P}} g(\theta)$ sous $\mathrm{P}_\theta^{(n)}$ (pour tout θ).

Biais vs Variance



Comparaison d'estimateurs: exemple du bus

Le MSE permet de réaliser un arbitrage entre (i) minimisation du biais et (ii) minimisation de la variance, deux objectifs souvent antagonistes.

Dans l'exemple du bus, on vérifie que les estimateurs de θ

$$T_1(X^{(n)}) = X_{(n)} \quad \text{et} \quad T_2(X^{(n)}) = \frac{n+1}{n} X_{(n)}$$

satisfont

$$\mathrm{MSE}_\theta^{(n)}[T_1(X^{(n)})] = \frac{2\theta^2}{(n+1)(n+2)} \geq \frac{\theta^2}{n(n+2)} = \mathrm{MSE}_\theta^{(n)}[T_2(X^{(n)})]$$

pour tout $\theta \in \Theta$. On gagne donc à remplacer $T_1(X^{(n)})$ par $T_2(X^{(n)})$!

Ceci tient pour tout n fixé (et l'inégalité est stricte pour $n \geq 2$)

Comparaison d'estimateurs: exemple du verre de bière

Dans l'exemple du verre de bière, on vérifie que les estimateurs de σ^2

$$T_1(X^{(n)}) = s^2 \quad \text{et} \quad T_2(X^{(n)}) = S^2$$

satisfont

$$\mathrm{MSE}_\theta^{(n)}[T_1(X^{(n)})] = \frac{(2n-1)\sigma^4}{n^2} < \frac{2\sigma^4}{n-1} = \mathrm{MSE}_\theta^{(n)}[T_2(X^{(n)})]$$

pour tout $\theta \in \Theta$ (pour tout $n \geq 2$).

On perd donc à remplacer $T_1(X^{(n)})$ par $T_2(X^{(n)})$!

Comparaison générale

Comparer des estimateurs $T_1(X^{(n)})$ et $T_2(X^{(n)})$ en termes de MSE ne peut se faire de façon conclusive que si $\mathrm{MSE}_\theta^{(n)}[T_1(X^{(n)})] \leq \mathrm{MSE}_\theta^{(n)}[T_2(X^{(n)})]$ ou $\mathrm{MSE}_\theta^{(n)}[T_1(X^{(n)})] \geq \mathrm{MSE}_\theta^{(n)}[T_2(X^{(n)})]$ pour tout $\theta \in \Theta$.

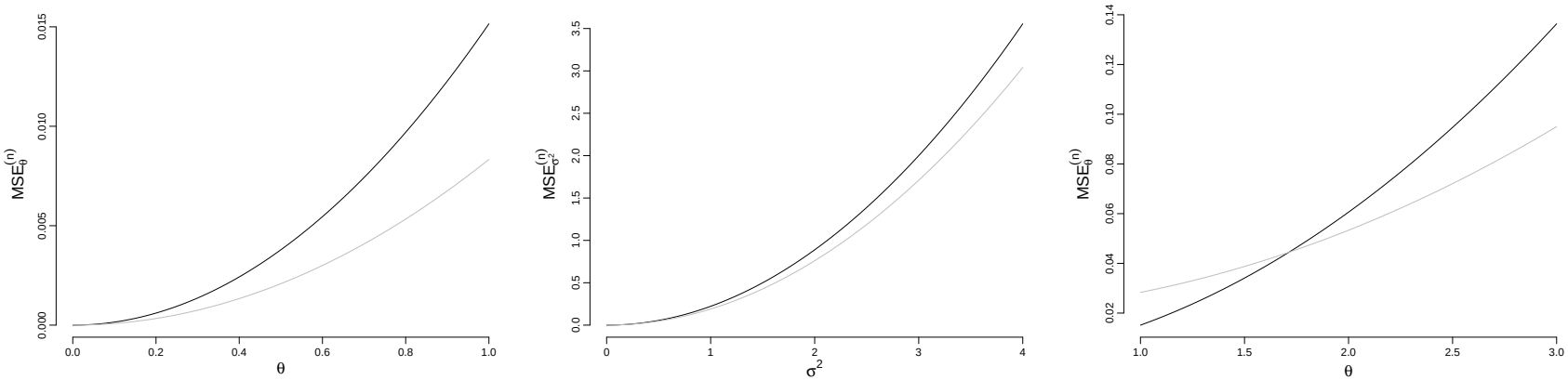


Figure 1: Courbes des MSE dans l'exemple du bus (gauche) et du verre de bière (milieu). A droite, un cas dans lequel aucun des deux estimateurs ne serait "meilleur" que l'autre.

Estimateur à erreur quadratique minimale

Définition 10

Soient un modèle statistique paramétrique et une fonction mesurable $g : \Theta \to \mathbb{R}$. *Soit* $\mathcal{C}$ *une classe d'estimateurs de* $g(\theta)$. *Alors* $T_*(X^{(n)})$ *est à erreur quadratique minimum dans* $\mathcal{C}$ *si et seulement si*

1. $T_*(X^{(n)}) \in \mathcal{C}$, *et*
2. *pour tout* $T(X^{(n)}) \in \mathcal{C}$,

$$\mathrm{MSE}_\theta^{(n)}[T_*(X^{(n)})] \leq \mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] \text{ pour tout } \theta \in \Theta.$$

Remarque : ci-dessus, les MSE sont relatifs à l'estimation de $g(\theta)$
($\mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] = \mathrm{E}_\theta^{(n)}[(T(X^{(n)}) - g(\theta))^2]$)

Principe du non-biais

Pourquoi se restreindre à une classe $\mathcal{C}$ d'estimateurs?

Soit $T_*(X^{(n)})$ à erreur quadratique minimum par rapport à tous les estimateurs. Fixons $\theta_0 \in \Theta$ arbitrairement et posons $T_{\theta_0}(x^{(n)}) = g(\theta_0)$ pour tout $x^{(n)}$. Alors

$$\begin{aligned} 0 \leq \mathrm{E}_{\theta_0}^{(n)}[(T_*(X^{(n)}) - g(\theta_0))^2] &\leq \mathrm{E}_{\theta_0}^{(n)}[(T_{\theta_0}(X^{(n)}) - g(\theta_0))^2] \\ &= \mathrm{E}_{\theta_0}^{(n)}[(g(\theta_0) - g(\theta_0))^2] \\ &= 0, \end{aligned}$$

ce qui implique que $T_*(X^{(n)}) = g(\theta_0)$ $\mathrm{P}_{\theta_0}^{(n)}$-p.s. Puisque θ_0 est arbitraire, on doit avoir $T_*(X^{(n)}) = g(\theta)$ $\mathrm{P}_{\theta}^{(n)}$-p.s. $\forall \theta \in \Theta$, ce qui est impossible.

Le principe du non-biais préconise de se restreindre à la classe $\mathcal{C}$ des estimateurs sans biais pour $g(\theta)$.

Moments de vecteurs aléatoires

Soit

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix}$$

un vecteur aléatoire à valeurs dans $\mathbb{R}^d$.

Si $\mathrm{E}[|Z_r|] < \infty$ pour tout $r = 1, \ldots, d$, le vecteur moyen de Z est

$$\mu_Z = \mathrm{E}[Z] \stackrel{\text{def}}{=} \begin{pmatrix} \mathrm{E}[Z_1] \\ \vdots \\ \mathrm{E}[Z_d] \end{pmatrix}.$$

On vérifiera directement que pour toute matrice $m \times d$ et tout m-vecteur b,

$$\mathrm{E}[AZ + b] = A\mathrm{E}[Z] + b.$$

Moments de vecteurs aléatoires

Si $\mathrm{E}[Z_r^2] < \infty$ pour tout $r = 1, \ldots, d$, la matrice de variance-covariance de Z est

$$\Sigma_Z = \mathrm{Var}[Z] \stackrel{\text{def}}{=} \mathrm{E}[(Z - \mu_Z)(Z - \mu_Z)']$$

$$= \begin{pmatrix} \mathrm{Var}[Z_1] & \mathrm{Cov}[Z_1, Z_2] & \ldots & \mathrm{Cov}[Z_1, Z_d] \\ \mathrm{Cov}[Z_1, Z_2] & \mathrm{Var}[Z_2] & \ldots & \mathrm{Cov}[Z_2, Z_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[Z_1, Z_d] & \mathrm{Cov}[Z_2, Z_d] & \ldots & \mathrm{Var}[Z_d] \end{pmatrix}.$$

On vérifiera facilement que

$$\mathrm{Var}[AZ + b] = A\mathrm{Var}[Z]A'$$

pour toute matrice $m \times d$ et tout m-vecteur b.

Moments de vecteurs aléatoires

Cette matrice est bien entendu symétrique. Elle est aussi semi-définie positive, puisque, pour tout d-vecteur v, on a

$$0 \leq \mathrm{Var}[v'Z] = v'\mathrm{Var}[Z](v')' = v'\Sigma_Z v$$

Par ailleurs, la matrice de variance-covariance se réécrit

$$\begin{aligned}
\Sigma_Z = \mathrm{Var}[Z] &= \mathrm{E}[(Z-\mu_Z)(Z-\mu_Z)'] \\
&= \mathrm{E}[ZZ' - Z\mu_Z' - \mu_Z Z' + \mu_Z\mu_Z'] \\
&= \mathrm{E}[ZZ'] - \mathrm{E}[Z]\mu_Z' - \mu_Z(\mathrm{E}[Z])' + \mu_Z\mu_Z' \\
&= \mathrm{E}[ZZ'] - \mu_Z\mu_Z'.
\end{aligned}$$

Moments de vecteurs aléatoires

Enfin, si

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix} \quad \text{et} \quad Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix}$$

admettent des matrices de variance-covariance, alors

$$\begin{aligned} \mathrm{Cov}[Y, Z] &\stackrel{\text{def}}{=} \mathrm{E}[(Y - \mu_Y)(Z - \mu_Z)'] \\ &= \begin{pmatrix} \mathrm{Cov}[Y_1, Z_1] & \mathrm{Cov}[Y_1, Z_2] & \dots & \mathrm{Cov}[Y_1, Z_d] \\ \mathrm{Cov}[Y_2, Z_1] & \mathrm{Cov}[Y_2, Z_2] & \dots & \mathrm{Cov}[Y_2, Z_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[Y_r, Z_1] & \mathrm{Cov}[Y_r, Z_2] & \dots & \mathrm{Cov}[Y_r, Z_d] \end{pmatrix} \end{aligned}$$

est la covariance entre Y et Z. On vérifiera facilement que

- $\mathrm{Cov}[Z, Y] = (\mathrm{Cov}[Y, Z])'$
- $\mathrm{Cov}[Y, Z] = \mathrm{E}[Y Z'] - \mu_Y \mu_Z'$
- $\mathrm{Cov}[A_1 Y + b_1, A_2 Z + b_2] = A_1 \mathrm{Cov}[Y, Z] A_2'$ pour toute matrice A_1, A_2 et vecteurs b_1, b_2 de dimensions compatibles.

Erreur quadratique moyenne multivariée

Définition 11

Soient un modèle statistique paramétrique et $g : \Theta \to \mathbb{R}^m$ *mesurable. Alors l'erreur quadratique moyenne, sous* $\mathrm{P}_\theta^{(n)}$*, de l'estimateur* $T(X^{(n)})$ *de* $g(\theta)$ *est*

$$\mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] = \mathrm{E}_\theta^{(n)}[(T(X^{(n)}) - g(\theta))(T(X^{(n)}) - g(\theta))'].$$

Le MSE est une matrice réelle $m \times m$.

On pourra comparer deux MSE matriciels au moyen de l'ordre induit par les matrices semi-définies positives: on dira que

$$A \leq B$$

si et seulement si $B - A \geq 0$, au sens où $B - A$ est semi-définie positive.

Estimateur à erreur quadratique minimale (version multivariée)

Définition 12

Soient un modèle statistique paramétrique et $g : \Theta \to \mathbb{R}^m$ *mesurable. Soit* $\mathcal{C}$ *une classe d'estimateurs de* $g(\theta)$*. Alors* $T_*(X^{(n)})$ *est à erreur quadratique minimum dans* $\mathcal{C}$ *si et seulement si*

1. $T_*(X^{(n)}) \in \mathcal{C}$ *et*
2. *pour tout* $T(X^{(n)}) \in \mathcal{C}$*,*

$$\mathrm{MSE}_\theta^{(n)}[T_*(X^{(n)})] \leq \mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] \text{ pour tout } \theta \in \Theta.$$

Lien entre multivarié et univarié

Supposons que $T_1(X^{(n)})$ domine $T_2(X^{(n)})$ en termes de MSE multivarié: $v'(\mathrm{MSE}_\theta^{(n)}[T_2(X^{(n)})] - \mathrm{MSE}_\theta^{(n)}[T_1(X^{(n)})])v \geq 0$ pour tout m-vecteur v.

Puisque

$$\begin{aligned} v'\mathrm{MSE}_\theta^{(n)}[T(X^{(n)})]v &= \mathrm{E}_\theta^{(n)}[v'(T(X^{(n)}) - g(\theta))(T(X^{(n)}) - g(\theta))'v] \\ &= \mathrm{E}_\theta^{(n)}[\{v'(T(X^{(n)}) - g(\theta))\}^2] \\ &= \mathrm{E}_\theta^{(n)}[\{v'T(X^{(n)}) - v'g(\theta)\}^2] \\ &= \mathrm{MSE}_\theta^{(n)}[v'T(X^{(n)})], \end{aligned}$$

où le dernier MSE est relatif à l'estimation de $v'g(\theta)$, on a donc

$$\mathrm{MSE}_\theta^{(n)}[v'T_2(X^{(n)})] \geq \mathrm{MSE}_\theta^{(n)}[v'T_1(X^{(n)})]$$

(domination scalaire de $v'T_1(X^{(n)})$ sur $v'T_2(X^{(n)})$) pour tout m-vecteur v!

Modèle régulier

Définition 13

Soit un modèle statistique paramétrique et soit $L_\theta^{(n)}(\cdot)$ sa fonction de vraisemblance. Ce modèle est régulier $\Leftrightarrow$ les hypothèses suivantes sont satisfaites :

(H1) Θ *est ouvert.*

(H2) $\mathcal{X}^{(n)} = \{x^{(n)} \in \mathbb{R}^n : L_\theta^{(n)}(x^{(n)}) > 0\}$ *ne dépend pas de* θ.

(H3) *Pour tout* $x^{(n)} \in \mathcal{X}^{(n)}$, *la fonction* $\theta \mapsto L_\theta^{(n)}(x^{(n)})$ *est différentiable sur* Θ.

(H4) *L'expression* $\int_{\mathcal{X}^{(n)}} L_\theta^{(n)}(x^{(n)})\, dx^{(n)}$ *est dérivable sous le signe, au sens où*

$$\nabla_\theta \int_{\mathcal{X}^{(n)}} L_\theta^{(n)}(x^{(n)})\, dx^{(n)} = \int_{\mathcal{X}^{(n)}} \nabla_\theta L_\theta^{(n)}(x^{(n)})\, dx^{(n)}.$$

(H5) *Pour tout* $\theta \in \Theta$, *la matrice réelle* $k \times k$

$$I^{(n)}(\theta) \stackrel{\text{def}}{=} \int_{\mathcal{X}^{(n)}} \underbrace{\nabla_\theta \ln L_\theta^{(n)}(x^{(n)})(\nabla_\theta \ln L_\theta^{(n)}(x^{(n)}))'}_{\text{une matrice } k\times k} L_\theta^{(n)}(x^{(n)})\, dx^{(n)}$$

existe, est finie et est inversible.

Remarques

- La définition est dans le cas continu. Le cas discret est similaire.
- L'hypothèse (H2) écarte l'exemple du modèle d'échantillonnage uniforme, pour lequel le support de la vraisemblance $L_\theta^{(n)}(\cdot)$ est

$$\mathcal{X}_\theta^{(n)} = [0,\theta] \times [0,\theta] \times \ldots \times [0,\theta] \quad (n \text{ fois}),$$

  donc dépend de θ.
- Il découle de (H4) que

$$\mathrm{E}_\theta^{(n)}[\nabla_\theta \ln L_\theta^{(n)}(X^{(n)})] = \int_{\mathcal{X}^{(n)}} \left(\nabla_\theta \ln L_\theta^{(n)}(x^{(n)})\right) L_\theta^{(n)}(x^{(n)})\, dx^{(n)}$$

$$= \int_{\mathcal{X}^{(n)}} \nabla_\theta L_\theta^{(n)}(x^{(n)})\, dx^{(n)} = \nabla_\theta \int_{\mathcal{X}^{(n)}} L_\theta^{(n)}(x^{(n)})\, dx^{(n)} = 0$$

Information de Fisher

En (H5), l'information de Fisher en θ, $I^{(n)}(\theta)$, se réécrit donc

$$\begin{aligned} I^{(n)}(\theta) &= \mathrm{E}_\theta^{(n)}[\nabla_\theta \ln L_\theta^{(n)}(X^{(n)})(\nabla_\theta \ln L_\theta^{(n)}(X^{(n)}))'] \\ &= \mathrm{Var}_\theta^{(n)}[\nabla_\theta \ln L_\theta^{(n)}(X^{(n)})]. \end{aligned}$$

Dans un modèle d'échantillonnage paramétrique,

$$\begin{aligned} I^{(n)}(\theta) &= \mathrm{Var}_\theta^{(n)}\left[\nabla_\theta \ln \prod_{i=1}^n L_\theta^{(1)}(X_i)\right] = \mathrm{Var}_\theta^{(n)}\left[\sum_{i=1}^n \nabla_\theta \ln L_\theta^{(1)}(X_i)\right] \\ &= \sum_{i=1}^n \mathrm{Var}_\theta^{(n)}[\nabla_\theta \ln L_\theta^{(1)}(X_i)] = n\mathrm{Var}_\theta^{(n)}[\nabla_\theta \ln L_\theta^{(1)}(X_1)] \\ &= nI^{(1)}(\theta). \end{aligned}$$

Information de Fisher : interprétation

Supposons que $k = 1$ et $I^{(n)}(\theta) = 0 \; \forall \theta$ (ce qui est incompatible avec (H5)). Puisqu'on a alors

$$\nabla_\theta \ln L_\theta^{(n)}(X^{(n)}) = 0 \quad \mathrm{P}_\theta^{(n)}\text{-p.s.},$$

on a aussi

$$\nabla_\theta L_\theta^{(n)}(X^{(n)}) = 0 \quad \mathrm{P}_\theta^{(n)}\text{-p.s.}$$

Par conséquent, la fonction de vraisemblance $\theta \mapsto L_\theta^{(n)}(x^{(n)})$ est constante (presque partout en $x^{(n)}$ pour tout θ), ce qui implique que les probabilités

$$\mathrm{P}_\theta^{(n)}[X^{(n)} \in B] = \int_B L_\theta^{(n)}(x^{(n)})\, dx^{(n)}, \quad B \in \mathcal{B}^n,$$

ne dépendent pas de θ. Dans cette situation, il est impossible de conduire l'inférence sur θ ($X^{(n)}$ *ne porte pas d'information sur* θ).

Information de Fisher : exemple

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$, avec $\mu \in \mathbb{R}$ (σ_0^2 connu).

Puisque

$$L_\mu^{(1)}(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(x-\mu)^2\right),$$

on a

$$\frac{d}{d\mu} \ln L_\mu^{(1)}(x) = \frac{d}{d\mu}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right) - \frac{1}{2\sigma_0^2}(x-\mu)^2\right] = \frac{1}{\sigma_0^2}(x-\mu),$$

ce qui donne

$$I^{(1)}(\mu) = \mathrm{Var}_\mu\left[\frac{d}{d\mu} \ln L_\mu^{(1)}(X_1)\right] = \mathrm{Var}_\mu\left[\frac{1}{\sigma_0^2}(X_1-\mu)\right] = \frac{1}{\sigma_0^4}\mathrm{Var}_\mu[X_1] = \frac{1}{\sigma_0^2},$$

donc

$$I^{(n)}(\mu) = nI^{(1)}(\mu) = \frac{n}{\sigma_0^2}.$$

Ici, l'information ne dépend pas de μ et est décroissante en σ_0^2.

Estimateur régulier

Définition 14

Soit un modèle statistique paramétrique et notons $L_\theta^{(n)}(\cdot)$ *sa fonction de vraisemblance. Soit* $T(X^{(n)})$ *un estimateur de* $g(\theta)$*, où* $g : \Theta \to \mathbb{R}^m$ *est mesurable. Cet estimateur est régulier* $\Leftrightarrow$ *les hypothèses suivantes sont satisfaites :*

(H1) $\mathrm{E}_\theta^{(n)}[\|T(X^{(n)})\|^2]$ *existe et est finie pour tout* $\theta \in \Theta$.

(H2) *L'expression*

$$\psi(\theta) \stackrel{\text{def}}{=} \mathrm{E}_\theta^{(n)}[T(X^{(n)})] = \int_{\mathcal{X}^{(n)}} T(x^{(n)})\, L_\theta^{(n)}(x^{(n)})\, dx^{(n)}$$

est dérivable sous le signe, au sens où

$$\nabla_\theta \int_{\mathcal{X}^{(n)}} T(x^{(n)})\, L_\theta^{(n)}(x^{(n)})\, dx^{(n)} = \int_{\mathcal{X}^{(n)}} T(x^{(n)})\, \nabla_\theta L_\theta^{(n)}(x^{(n)})\, dx^{(n)}.$$

Borne de Cramér–Rao

Théorème 15 (Borne de Cramér–Rao)

Soit un modèle statistique paramétrique régulier. Soit $T(X^{(n)})$ un estimateur régulier de $g(\theta)$, où $g : \Theta \to \mathbb{R}^m$ est mesurable. Alors, pour tout $\theta \in \Theta$,

$$\begin{aligned} \mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] &= \mathrm{E}_\theta^{(n)}[(T(X^{(n)}) - g(\theta))(T(X^{(n)}) - g(\theta))'] \\ &\geq \Delta_\theta (I^{(n)}(\theta))^{-1} \Delta_\theta', \end{aligned}$$

où $\Delta_\theta = (\frac{\partial \psi_i}{\partial \theta_j}(\theta))_{i=1,\ldots,m,\, j=1,\ldots,k}$ est la matrice jacobienne de $\psi(\cdot)$ et où "$\geq$" est l'ordre associé aux matrices semi-définies positives.

Preuve: au tableau.

Réécriture

Si $g(\theta) = \theta$ et si l'estimateur $T(X^{(n)})$ de θ est sans biais ($\psi(\theta) = g(\theta) = \theta$), alors la borne de Cramér–Rao prend la forme

$$(I^{(n)}(\theta))^{-1} :$$

plus grande est l'information, meilleure est la précision d'un estimateur optimal.

Dans un modèle d'échantillonnage, la borne de Cramér–Rao se réécrit

$$\frac{1}{n}\, \Delta_\theta (I^{(1)}(\theta))^{-1} \Delta_\theta'.$$

La MSE d'un estimateur (régulier) tend vers 0 à la vitesse $\frac{1}{n}$ au mieux.

Estimateur efficace

Définition 16

Soit un modèle statistique paramétrique régulier. Soit $T(X^{(n)})$ un estimateur régulier de $g(\theta)$, où $g : \Theta \to \mathbb{R}^m$ est mesurable. L'estimateur $T(X^{(n)})$ est efficace pour $g(\theta)$ si et seulement si

$$\mathrm{MSE}_\theta^{(n)}[T(X^{(n)})] = \Delta_\theta (I^{(n)}(\theta))^{-1} \Delta_\theta'$$

pour tout $\theta \in \Theta$.

La preuve du théorème précédent révèle que ceci a lieu si et seulement si

1. $\psi(\cdot) = g(\cdot)$, et
2. $\mathrm{Var}_\theta^{(n)}[T(X^{(n)})] = \Delta_\theta (I^{(n)}(\theta))^{-1} \Delta_\theta'$ pour tout $\theta \in \Theta$.

Estimateur efficace

Un estimateur ne peut donc être efficace que pour l'estimation de $g(\theta) = \psi(\theta)$ ($\rightsquigarrow$ le non-biais est une condition nécessaire d'efficacité!)

Définition 17

Soit un modèle statistique paramétrique régulier. Soit $T(X^{(n)})$ un estimateur régulier de $g(\theta)$, où $g : \Theta \to \mathbb{R}^m$ est mesurable. L'estimateur $T(X^{(n)})$ est efficace (automatiquement pour sa moyenne $\psi(\theta)$) si et seulement si

$$\mathrm{Var}_\theta^{(n)}[T(X^{(n)})] = \Delta_\theta (I^{(n)}(\theta))^{-1} \Delta_\theta'$$

pour tout $\theta \in \Theta$.

Estimateur efficace: exemple du verre de bière

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$, avec $\mu \in \mathbb{R}$ (σ_0^2 connu).

Puisque

$$I^{(n)}(\mu) = \frac{n}{\sigma_0^2},$$

un estimateur $T(X^{(n)})$ sans biais de μ est efficace si

$$\mathrm{Var}_\mu^{(n)}[T(X^{(n)})] = \frac{1}{I^{(n)}(\mu)} = \frac{\sigma_0^2}{n}.$$

Puisque ceci est la variance de l'estimateur sans biais $\bar{X}$, celui-ci est efficace!

Estimateur efficace: exemple électoral

Soit $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. Bern(p), avec $p \in (0, 1)$.

Puisque $L_p^{(1)}(x) = p^x(1-p)^{1-x}$, on a

$$\frac{d}{dp} \ln L_p(x) = \frac{d}{dp}\Big[x(\ln p)+(1-x)\ln(1-p)\Big] = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x}{p(1-p)} - \frac{1}{1-p},$$

ce qui donne

$$I^{(1)}(p) = \mathrm{Var}_p\left[\frac{d}{dp} \ln L_p^{(1)}(X_1)\right] = \mathrm{Var}_p\left[\frac{X_1}{p(1-p)} - \frac{1}{1-p}\right]$$

$$= \frac{1}{p^2(1-p)^2}\mathrm{Var}_p[X_1] = \frac{1}{p^2(1-p)^2}\, p(1-p) \;= \frac{1}{p(1-p)}.$$

Un estimateur $T(X^{(n)})$ sans biais de p est donc efficace si

$$\mathrm{Var}_p^{(n)}[T(X^{(n)})] = \frac{1}{I^{(n)}(p)} = \frac{1}{nI^{(1)}(p)} = \frac{p(1-p)}{n}.$$

Puisque ceci est la variance de l'estimateur sans biais $\bar{X}$, celui-ci est efficace!

Critère d'efficacité

Nous concluons avec les deux résultats suivants.

Théorème 18 (Critère d'efficacité, dans le cas où $m = k$)

S'il existe une matrice $A(\theta)$ de rang maximal et une fonction $\phi(\theta)$ telles que

$$\nabla_\theta \ln L_\theta^{(n)}(X^{(n)}) = A(\theta)\big(T(X^{(n)}) - \phi(\theta)\big) \quad P_\theta^{(n)}\text{-p.s.}$$

pour tout θ, alors $T(X^{(n)})$ est efficace pour $\psi(\theta) = \mathrm{E}_\theta[T(X^{(n)})]$ (en fait, il s'agit d'un "si et seulement si"). Dans ce cas, on a

$$\begin{aligned} &(i) \quad \phi(\theta) = \psi(\theta), \\ &(ii) \quad \mathrm{Var}_\theta^{(n)}[T(X^{(n)})] = \Delta_\theta(A'(\theta))^{-1} \\ &(iii) \quad I^{(n)}(\theta) = A(\theta)\Delta_\theta \end{aligned}$$

pour tout θ.

Expression alternative pour l'information de Fisher

On verra que, si la vraisemblance est C^2 en un paramètre scalaire θ, alors

$$I^{(n)}(\theta) = -\mathrm{E}\left[\frac{d^2}{d\theta^2} \ln L_\theta^{(n)}(X^{(n)})\right].$$

Si $\theta \in \Theta \subset \mathbb{R}^k$, ceci se généralise en $\big(I^{(n)}(\theta)\big)_{ij} = -\mathrm{E}\big[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln L_\theta^{(n)}(X^{(n)})\big]$.

Dans l'exemple du verre de bière,

$$\frac{d^2}{d\mu^2} \ln L_\mu^{(1)}(X_1) = \frac{d}{d\mu}\left(\frac{d}{d\mu} \ln L_\mu^{(1)}(X_1)\right) = \frac{d}{d\mu}\left(\frac{1}{\sigma_0^2}(X_1 - \mu)\right) = -\frac{1}{\sigma_0^2},$$

ce qui livre bien $I^{(1)}(\mu) = -\mathrm{E}\big[\frac{d^2}{d\mu^2} \ln L_\mu^{(n)}(X_1)\big] = \frac{1}{\sigma_0^2}$.

Contenu du chapitre

Introduction

En pratique, personne ne fournit des estimateurs à évaluer sur la base des critères ci-dessus. Il nous faut donc des méthodes pour construire des estimateurs.

Dans cette section, on présente deux méthodes générales:

1. la méthode des moments
2. la méthode du maximum de vraisemblance

Méthode des moments

Soit un modèle statistique d'échantillonnage paramétrique

$$\left(\mathbb{R}^n, \mathcal{B}^n, \mathcal{P}^{(n)} = \left\{ \mathrm{P}_\theta^{(n)} : \theta \in \Theta \subset \mathbb{R}^k \right\}\right),$$

où les mesures de probabilité $\mathrm{P}_\theta^{(n)}$ représentent les distributions possibles de $X^{(n)} = (X_1, \ldots, X_n)$.

La méthode des moments suppose que la fonction

$$\begin{array}{rcl} M : \Theta & \to & M(\Theta) \\ \theta & \mapsto & \begin{pmatrix} \mu'_1(\theta) \\ \vdots \\ \mu'_k(\theta) \end{pmatrix}, \end{array}$$

est bien définie et inversible (pour rappel, $\mu'_r(\theta) = \mathrm{E}_\theta[X_1^r]$).

Méthode des moments

La méthode des moments consiste à considérer l'estimateur $T(X^{(n)})$ qui est la solution en θ du système

$$\begin{cases} \mu_1'(\theta) = m_1' \\ \qquad \vdots \\ \mu_k'(\theta) = m_k', \end{cases}$$

où $m_r' = \frac{1}{n}\sum_{i=1}^n X_i^r$. L'hypothèse d'inversibilité assure l'existence d'une solution unique, pour autant que $(m_1', \ldots, m_k') \in M(\Theta)$.

Méthode des moments : exemple du verre de bière

Soit $X^{(n)} = (X_1, \ldots, X_n)$, avec $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, où $\theta = (\mu, \sigma^2)$ appartient à $\Theta = \mathbb{R} \times \mathbb{R}_0^+ \subset \mathbb{R}^2$. Puisque

$$\begin{aligned} \mu_1'(\theta) &= \mathrm{E}_\theta[X_1] = \mu \\ \mu_2'(\theta) &= \mathrm{E}_\theta[X_1^2] = \sigma^2 + \mu^2, \end{aligned}$$

l'estimateur $T(X^{(n)})$ des moments de θ est la solution en θ du système

$$\left\{ \begin{array}{l} \mu = \frac{1}{n} \sum_{i=1}^n X_i \\ \sigma^2 + \mu^2 = \frac{1}{n} \sum_{i=1}^n X_i^2. \end{array} \right.$$

Ceci livre

$$T(X^{(n)}) = \left(\begin{array}{c} \bar{X} \\ \left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) - \bar{X}^2 \end{array} \right) = \left(\begin{array}{c} \bar{X} \\ s^2 \end{array} \right).$$

Méthode des moments : exemple électoral

Soit $X^{(n)} = (X_1, \ldots, X_n)$, avec $X_1, \ldots, X_n$ i.i.d. $\mathrm{Bern}(p)$, où $\theta = p$ appartient à $\Theta = [0, 1] \subset \mathbb{R}$.

Puisque $\mu_1'(p) = \mathrm{E}_p[X_1] = p$, l'estimateur $T(X^{(n)})$ des moments de p est la solution en p de l'équation

$$p = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Il s'agit donc de l'estimateur

$$T(X^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

Méthode des moments : exemple du bus

Soit $X^{(n)} = (X_1, \ldots, X_n)$, avec $X_1, \ldots, X_n$ i.i.d. $\mathrm{Unif}([0, \theta])$, où $\theta \in \Theta = \mathbb{R}_0^+$.

Comme $\mu_1'(\theta) = \mathrm{E}_\theta[X_1] = \frac{\theta}{2}$, l'estimateur $T(X^{(n)})$ des moments de θ est la solution en θ de l'équation

$$\frac{\theta}{2} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Ceci livre $T(X^{(n)}) = 2\bar{X}$.

Remarque : puisqu'on a vu que $\mathrm{MSE}_\theta^{(n)}[X_{(n)}] = O(\frac{1}{n^2})$ quand $n \to \infty$ et que

$$\mathrm{MSE}_\theta^{(n)}[2\bar{X}] = \mathrm{Var}_\theta^{(n)}[2\bar{X}] + 0^2 = \frac{4}{n}\mathrm{Var}_\theta[X_1] = \frac{\theta^2}{3n},$$

l'estimateur des moments $2\bar{X}$ ne peut pas être à erreur quadratique minimum pour n grand.

Méthode des moments : convergence

Fixons $\theta \in \Theta$ arbitrairement. Sous l'hypothèse de moments finis d'ordre k, la loi forte des grands nombres assure que $m'_r \overset{\text{p.s.}}{\to} \mu'_r(\theta)$ sous $\mathrm{P}_\theta^{(n)}$, $r = 1, \ldots, k$.

Si la fonction réciproque de $\theta \mapsto M(\theta) = (\mu'_1(\theta), \ldots, \mu'_k(\theta))$ est continue, il en découle que, sous $\mathrm{P}_\theta^{(n)}$,

$$T(X^{(n)}) = M^{-1}(m'_1, \ldots, m'_k) \overset{\text{p.s.}}{\to} M^{-1}(\mu'_1(\theta), \ldots, \mu'_k(\theta)) = \theta.$$

Donc les estimateurs des moments sont fortement convergents.

Mais il n'y a aucune garantie qu'ils soient efficaces ou à erreur quadratique minimum, pas même asymptotiquement (voir le slide précédent).

Méthode du maximum de vraisemblance

Considérons le modèle statistique paramétrique

$$\left(\mathbb{R}^n, \mathcal{B}^n, \mathcal{P}^{(n)} = \left\{ \mathrm{P}_\theta^{(n)} : \theta \in \Theta \subset \mathbb{R}^k \right\}\right)$$

et notons $L_\theta^{(n)}(\cdot)$ la fonction de vraisemblance associée.

Dans un premier temps, considérons l'estimation du paramètre θ uniquement (plutôt que celle de $g(\theta)$).

La méthode du maximum de vraisemblance consiste à estimer θ par "la" valeur $T(X^{(n)})$ de θ qui, à $X^{(n)}$ fixé, maximise la vraisemblance $L_\theta^{(n)}(X^{(n)})$.

Méthode du maximum de vraisemblance

Parce que l'unicité n'est pas garantie, nous adoptons la définition suivante.

Définition 19

Soit un modèle statistique paramétrique. La statistique $T(X^{(n)})$ à valeurs dans Θ est un estimateur du maximum de vraisemblance de $\theta \Leftrightarrow$ pour tout $\theta \in \Theta$,

$$L^{(n)}_{T(X^{(n)})}(X^{(n)}) \geq L^{(n)}_{\theta}(X^{(n)}) \quad \mathrm{P}^{(n)}_{\theta}\text{-}p.s.,$$

au sens où $\mathrm{P}^{(n)}_{\theta}[\{x^{(n)} : L^{(n)}_{T(x^{(n)})}(x^{(n)}) \geq L^{(n)}_{\theta}(x^{(n)})\}] = 1.$

Pour faire court, on écrira souvent ci-dessous que $T(X^{(n)})$ est un MLE (maximum likelihood estimator) de θ.

Méthode du maximum de vraisemblance : exemple du bus

Soit $X^{(n)} = (X_1, \ldots, X_n)$, avec $X_1, \ldots, X_n$ i.i.d. $\mathrm{Unif}([0, \theta])$, où $\theta \in \Theta = \mathbb{R}_0^+$.

La vraisemblance correspondante est

$$L_\theta^{(n)}(X^{(n)}) = \frac{1}{\theta^n} \mathbb{I}[X_{(1)} \geq 0] \mathbb{I}[X_{(n)} \leq \theta].$$

$\rightsquigarrow T(X^{(n)}) = X_{(n)}$ est le MLE de θ.

Non-unicité de l'estimateur maximum de vraisemblance

Soit $X^{(n)} = (X_1, \ldots, X_n)$, avec $X_1, \ldots, X_n$ i.i.d. $\mathrm{Unif}([\theta - \frac{1}{2}, \theta + \frac{1}{2}])$, où le paramètre $\theta \in \Theta = \mathbb{R}$.

La vraisemblance associée est

$$L_\theta^{(n)}(X^{(n)}) = \prod_{i=1}^{n} \mathbb{I}[\theta - \tfrac{1}{2} \leq X_i \leq \theta + \tfrac{1}{2}] = \mathbb{I}[\theta - \tfrac{1}{2} \leq X_1, \ldots, X_n \leq \theta + \tfrac{1}{2}]$$

$$= \mathbb{I}[\theta - \tfrac{1}{2} \leq X_{(1)}, X_{(n)} \leq \theta + \tfrac{1}{2}] = \mathbb{I}[X_{(n)} - \tfrac{1}{2} \leq \theta \leq X_{(1)} + \tfrac{1}{2}].$$

Toute statistique $T(X^{(n)})$ qui, avec probabilité 1 (sous n'importe quel $\mathrm{P}_\theta^{(n)}$), prend ses valeurs dans l'intervalle $[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}]$ est un MLE de θ.

$\rightsquigarrow$ $X_{(n)} - \frac{1}{2}$, $X_{(1)} + \frac{1}{2}$, et $(X_{(1)} + X_{(n)})/2$ sont des MLE de θ.

Lien avec l'exhaustivité

Supposons que $S(X^{(n)})$ est une statistique exhaustive.

Par le critère de factorisation de Neyman–Fisher, pour tout $\theta \in \Theta$,

$$L_\theta^{(n)}(X^{(n)}) = g_\theta(S(X^{(n)}))\, h(X^{(n)}) \quad \mathrm{P}_\theta^{(n)}\text{-p.s.}$$

Clairement, $T(X^{(n)})$ maximise la vraisemblance $L_\theta^{(n)}(X^{(n)})$ en θ si et seulement si $T(X^{(n)})$ maximise $g_\theta(S(X^{(n)}))$ en θ.

$\rightsquigarrow$ Tout MLE de θ est de la forme

$$T(X^{(n)}) = m(S(X^{(n)}))$$

pour une certaine fonction m.

(Illustration sur la base des exemples ci-dessus)

Vraisemblance différentiable

Si, pour tout $x^{(n)}$, la vraisemblance $\theta \mapsto L_\theta^{(n)}(x^{(n)})$ est différentiable, alors toute valeur du paramètre θ (appartenant à l'intérieur de Θ) maximisant la vraisemblance doit être solution du système d'équations

$$\nabla_\theta L_\theta^{(n)}(X^{(n)}) = 0,$$

ou, de façon équivalente, du système des équations de vraisemblance

$$\nabla_\theta \ln L_\theta^{(n)}(X^{(n)}) = 0.$$

On devra s'assurer que les solutions de ce système correspondent bien à des maxima globaux et à étudier séparément l'éventuel bord de Θ.

Maximum de vraisemblance : exemple électoral

Soit le vecteur aléatoire $X^{(n)} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. $\mathrm{Bern}(p)$, avec $\theta = p \in \Theta = [0,1] \subset \mathbb{R}$. Puisque, pour tout $p \in]0,1[$,

$$L_p^{(n)}(X^{(n)}) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} = p^{n\bar{X}}(1-p)^{n-n\bar{X}} \quad \mathrm{P}_p^{(n)}\text{-p.s.},$$

on trouve

$$\begin{aligned}\frac{d}{dp}\ln L_p^{(n)}(X^{(n)}) &= n\bar{X} \times \frac{1}{p} - (n - n\bar{X}) \times \frac{1}{1-p} \\ &= \frac{n}{p(1-p)}\big(\bar{X}(1-p) - (1-\bar{X})p\big) = \frac{n}{p(1-p)}(\bar{X} - p).\end{aligned}$$

Les équations de vraisemblance prennent donc la forme

$$\frac{n}{p(1-p)}(\bar{X} - p) = 0.$$

Maximum de vraisemblance : exemple électoral

La seule solution en p est bien entendu

$$T(X^{(n)}) = \bar{X}.$$

Indépendamment du fait que $\bar{X}$ prenne ou pas une valeur dans l'intérieur de Θ, il s'agit bien d'un maximum (pourquoi?)

Donc $\bar{X}$ est le MLE de p.

Estimation de $g(\theta)$

Passons à l'estimation de $g(\theta)$.

Si la fonction $g : \Theta \mapsto g(\Theta)$ est bijective, θ et $\tilde{\theta} = g(\theta)$ fournissent deux "adressages" valides du modèle, menant aux vraisemblances liées par

$$L_\theta^{(n)}(X^{(n)}) = L_{g^{-1}(\tilde{\theta})}^{(n)}(X^{(n)}) = \tilde{L}_{\tilde{\theta}}^{(n)}(X^{(n)})$$

($\tilde{L}^{(n)}$ désigne la vraisemblance associée à la paramétrisation par $\tilde{\theta}$).

$L_\theta^{(n)}(X^{(n)})$ est maximisée pour la valeur $T(X^{(n)})$ de θ
$\Leftrightarrow \tilde{L}_{\tilde{\theta}}^{(n)}(X^{(n)})$ est maximisée pour la valeur $g(T(X^{(n)}))$ de $\tilde{\theta} = g(\theta)$.

$\rightsquigarrow$ Si $T(X^{(n)})$ est un MLE de θ, alors $g(T(X^{(n)}))$ est un MLE de $g(\theta)$.

Estimation de $g(\theta)$

Si $g : \Theta \to g(\Theta)$ n'est pas bijective, la fonction

$$\begin{array}{ccc} G : \Theta & \to & G(\Theta) \\ \theta & \mapsto & \begin{pmatrix} g(\theta) \\ \theta \end{pmatrix} \end{array}$$

l'est, de sorte que, par le slide précédent, $G(T(X^{(n)}))$ est un MLE de $G(\theta)$.

Se restreindre à la première composante de $G(\cdot)$ mène à considérer que $g(T(X^{(n)}))$ est un MLE de $g(\theta)$.

On parle d'invariance du maximum de vraisemblance.

Comportement asymptotique : hypothèses

Comme déjà mentionné, il n'est aucunement garanti que la méthode des moments fournisse des estimateurs satisfaisants en termes d'efficacité.

Par contre, sous certaines hypothèses de régularité, la méthode du maximum de vraisemblance livre des estimateurs asymptotiquement efficaces.

Afin d'énoncer le théorème (dans le cadre des modèles d'échantillonnage à paramètre scalaire: $\Theta \subset \mathbb{R}$), nous adoptons les hypothèses suivantes :

(H1) Θ est ouvert.

(H2) $\mathcal{X} = \{x \in \mathbb{R} : L_\theta^{(1)}(x) > 0\}$ ne dépend pas de θ.

(H3) Pour tout $x \in \mathcal{X}$, la fonction $\theta \mapsto L_\theta^{(1)}(x)$ est dérivable deux fois sur Θ.

Comportement asymptotique : hypothèses

(H4) L'expression $\int_{\mathcal{X}} L_\theta^{(1)}(x)\,dx$ est dérivable deux fois sous le signe:

$$\frac{d^r}{d\theta^r}\int_{\mathcal{X}} L_\theta^{(1)}(x)\,dx = \int_{\mathcal{X}} \frac{d^r}{d\theta^r} L_\theta^{(1)}(x)\,dx, \quad r = 1, 2.$$

(H5) La fonction $\theta \mapsto B_\theta^{(1)}(x) \stackrel{\text{def}}{=} \frac{d^2}{d\theta^2} \ln L_\theta^{(1)}(x)$ est continue uniformément en x: si $(\theta_n) \to \theta$, alors $\sup_{x\in\mathcal{X}} |B_{\theta_n}^{(1)}(x) - B_\theta^{(1)}(x)| \to 0$.

(H6) Pour tout $\theta \in \Theta$, le réel

$$I^{(1)}(\theta) \stackrel{\text{def}}{=} \int_{\mathcal{X}} \left(\frac{d}{d\theta} \ln L_\theta^{(1)}(x)\right)^2 L_\theta^{(1)}(x)\,dx$$

existe et appartient à $]0, \infty[$.

Comportement asymptotique

Théorème 20

Soit un modèle statistique d'échantillonnage paramétrique, avec $\Theta \subset \mathbb{R}$.
Notons $L_\theta^{(1)}(\cdot)$ *la fonction de vraisemblance associée à un échantillon de taille* 1.
Sous les hypothèses (H1)–(H6),

1. *il existe une suite* $(T(X^{(n)}))$ *de solutions des équations de vraisemblance telle que, pour tout* $\theta \in \Theta$, $T(X^{(n)}) \overset{\text{p.s.}}{\to} \theta$ *sous* $\mathrm{P}_\theta^{(n)}$ *quand* $n \to \infty$.
2. *Pour toute telle suite* $(T(X^{(n)}))$, *on a que, pour tout* $\theta \in \Theta$,

$$\sqrt{n}(T(X^{(n)}) - \theta) \overset{\mathcal{L}}{\to} \mathcal{N}\left(0\,, \frac{1}{I^{(1)}(\theta)}\right)$$

sous $\mathrm{P}_\theta^{(n)}$ *quand* $n \to \infty$.

C'est la propriété BAN (pour Best Asymptotically Normal) du MLE.

Comportement asymptotique : interprétation

Un estimateur $T_{\text{eff}}(X^{(n)})$ efficace pour θ satisfait

$$\mathrm{E}_\theta^{(n)}[T_{\text{eff}}(X^{(n)})] = \theta \quad \text{et} \quad \mathrm{Var}_\theta^{(n)}[T_{\text{eff}}(X^{(n)})] = \frac{1}{I^{(n)}(\theta)} = \frac{1}{nI^{(1)}(\theta)},$$

ou, de façon équivalente,

$$\mathrm{E}_\theta^{(n)}[\sqrt{n}(T_{\text{eff}}(X^{(n)}) - \theta)] = 0 \quad \text{et} \quad \mathrm{Var}_\theta^{(n)}[\sqrt{n}(T_{\text{eff}}(X^{(n)}) - \theta)] = \frac{1}{I^{(1)}(\theta)}.$$

On peut donc considérer que la propriété BAN

$$\sqrt{n}(T(X^{(n)}) - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{I^{(1)}(\theta)}\right)$$

indique que la MLE $T(X^{(n)})$ est asymptotiquement efficace pour θ.

Résumé du chapitre

- Estimation ponctuelle : définitions d'estimateur convergent, exhaustif, (asymptotiquement) sans biais, à erreur quadratique minimum, efficace
- Critère de factorisation de Neyman–Fisher, décomposition du MSE, borne de Cramér–Rao
- Méthode des moments, méthode du maximum de vraisemblance
- Comportement asymptotique des estimateurs MLE