

MAP 433 : Introduction aux méthodes statistiques. Cours 3

11 Septembre 2015

Aujourd'hui

- 1 Modélisation statistique
 - Expérience statistique
 - Expériences dominées
 - Modèle de densité

- 2 Méthodes d'estimation pour le modèle de densité
 - Méthode des moments
 - Z -estimation
 - M -estimation
 - Principe de maximum de vraisemblance

Expérience statistique

Consiste à identifier :

- Des observations

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$$

considérées comme des **réalisations** de variables aléatoires $Z = (X_1, \dots, X_n)$ de loi \mathbb{P}^Z .

- Une famille de lois

$$\{\mathbb{P}_\theta, \theta \in \Theta\}.$$

- Une **problématique** : retrouver le paramètre θ tel que $\mathbb{P}^Z = \mathbb{P}_\theta$ (estimation) ou bien prendre une décision sur une propriété relative à θ (test).

Expérience statistique

- Approche paramétrique : **on suppose** que F appartient à une **famille de lois connue** indexée par un paramètre θ de dimension finie : $\theta \in \Theta \subset \mathbb{R}^d$.
 - Exemple : $\Theta = \mathbb{R}$,

$$X_i = \theta + \xi_i, \quad i = 1, \dots, n,$$

ξ_i v.a. i.i.d. de densité **connue** f sur \mathbb{R} et $\mathbb{E}(X_i) = \theta$.

Question : en utilisant cette information supplémentaire, peut-on construire un estimateur plus performant que l'estimateur \bar{X}_n basé sur l'approche empirique ?

Expérience statistique

- En écrivant

$$X_i = \theta + \xi_i, \quad i = 1, \dots, n,$$

ξ_i v.a. i.i.d. de densité **connue** f , nous précisons la forme de la loi \mathbb{P}_θ de (X_1, \dots, X_n) :

$$\mathbb{P}_\theta [A] = \int_A \left(\prod_{i=1}^n f(x_i - \theta) \right) dx_1 \dots dx_n,$$

pour tout $A \in \mathcal{B}(\mathbb{R}^n)$.

Expérience statistique

Definition

Une expérience (un modèle) statistique \mathcal{E} est le triplet

$$\mathcal{E} = (\mathfrak{Z}, \mathcal{Z}, \{ \mathbb{P}_\theta, \theta \in \Theta \}),$$

avec

- $(\mathfrak{Z}, \mathcal{Z})$ espace mesurable (souvent $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$),
- $\{ \mathbb{P}_\theta, \theta \in \Theta \}$ famille de probabilités définies *simultanément* sur le même espace $(\mathfrak{Z}, \mathcal{Z})$,
- θ est le *paramètre inconnu*, et Θ est *l'ensemble des paramètres connu*.

Expérience engendrée par (X_1, \dots, X_n)

- **Traitement sur un exemple** : on observe

$$Z = (X_1, \dots, X_n), \quad X_i = \theta + \xi_i,$$

ξ_i v.a. i.i.d. de densité **connue** f .

- La famille de lois $\{\mathbb{P}_\theta^n, \theta \in \Theta = \mathbb{R}\}$ est définie sur $\mathcal{Z} = \mathbb{R}^n$ par

$$\mathbb{P}_\theta^n[A] = \int_A \left(\prod_{i=1}^n f(x_i - \theta) \right) dx_1 \dots dx_n,$$

pour $A \in \mathcal{Z} = \mathcal{B}(\mathbb{R}^n)$ (et \mathbb{P}^Z est l'une des \mathbb{P}_θ^n).

- Expérience **engendrée par l'observation** Z :

$$\mathcal{E}^n = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^n, \theta \in \Theta\}).$$

Expérience (modèle) paramétrique, non-paramétrique

- Si Θ peut être « pris » comme un sous-ensemble de \mathbb{R}^d :
expérience (=modèle) paramétrique.
- Sinon (par exemple si le paramètre θ est un élément d'un espace fonctionnel) : expérience (=modèle) non-paramétrique.

Expériences dominées

- On fait une hypothèse minimale de « complexité » sur le modèle statistique. **But** : ramener l'étude de la famille

$$\{\mathbb{P}_\theta, \theta \in \Theta\}$$

à l'étude d'une famille de fonctions

$$\{z \in \mathfrak{Z} \rightsquigarrow f(\theta, z) \in \mathbb{R}_+, \theta \in \Theta\}.$$

- Via la notion de **domination**. Si μ, ν sont deux mesures σ -finies sur \mathfrak{Z} , alors μ **domine** ν (notation $\nu \ll \mu$) si

$$\mu[A] = 0 \Rightarrow \nu[A] = 0.$$

Théorème de Radon-Nikodym

Théorème

Si $\nu \ll \mu$, il existe une fonction positive

$$z \rightsquigarrow p(z) \stackrel{\text{notation}}{=} \frac{d\nu}{d\mu}(z),$$

définie μ -p.p., μ -intégrable, telle que

$$\nu[A] = \int_A p(z) \mu(dz) = \int_A \frac{d\nu}{d\mu}(z) \mu(dz), \quad A \in \mathcal{Z}.$$

Expérience dominée

Definition

Une expérience statistique $\mathcal{E} = (\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ est *dominée* par la mesure σ -finie μ définie sur \mathfrak{Z} si

$$\forall \theta \in \Theta : \mathbb{P}_\theta \ll \mu.$$

On appelle *densités* de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ la famille de fonctions (définies μ -p.p.)

$$z \rightsquigarrow \frac{d\mathbb{P}_\theta}{d\mu}(z), \quad z \in \mathfrak{Z}, \quad \theta \in \Theta.$$

Densité, régression

Deux classes d'expériences statistiques **dominées** fondamentales :

- Le modèle de **densité**
- Le modèle de **régression**

Modèle de densité (paramétrique)

- On observe un n -échantillon de v.a.r. X_1, \dots, X_n .
- La loi des X_i appartient à $\{\mathbb{P}_\theta, \theta \in \Theta\}$, famille de **probabilités sur \mathbb{R}** , **dominée** par une mesure (σ -finie) $\mu(dx)$ sur \mathbb{R} .
- La loi de (X_1, \dots, X_n) s'écrit

$$\begin{aligned}\mathbb{P}_\theta^n(dx_1 \cdots dx_n) &= \mathbb{P}_\theta(dx_1) \otimes \cdots \otimes \mathbb{P}_\theta(dx_n) \\ &\ll \mu(dx_1) \otimes \cdots \otimes \mu(dx_n) \\ &\stackrel{\text{notation}}{=} \mu^n(dx_1 \cdots dx_n)\end{aligned}$$

Modèle de densité (paramétrique)

- Densité du modèle : on part de

$$f(\theta, x) = \frac{d\mathbb{P}_\theta}{d\mu}(x), \quad x \in \mathbb{R}$$

et

$$\frac{d\mathbb{P}_\theta^n}{d\mu^n}(x_1, \dots, x_n) = \prod_{i=1}^n f(\theta, x_i), \quad x_1, \dots, x_n \in \mathbb{R}.$$

- L'expérience statistique engendrée par (X_1, \dots, X_n) s'écrit :

$$\mathcal{E}^n = \left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{ \mathbb{P}_\theta^n, \theta \in \Theta \} \right), \quad \Theta \subset \mathbb{R}^d.$$

Exemple 1 : modèle de densité gaussienne univariée

- $X_i \sim \mathcal{N}(m, \sigma^2)$, avec

$$\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}.$$

$$\mathbb{P}_\theta(dx) = f(\theta, x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)dx \\ \ll \mu(dx) = dx.$$

- Puis

$$\frac{d\mathbb{P}_\theta^n}{d\mu^n}(x_1, \dots, x_n) = \prod_{i=1}^n f(\theta, x_i) \\ = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2\right),$$

avec $x_1, \dots, x_n \in \mathbb{R}$.

Exemple 2 : modèle de Bernoulli

- $X_i \sim \text{Bernoulli}(\theta)$, avec $\theta \in \Theta = [0, 1]$.

$$\mathbb{P}_\theta(dx) = (1 - \theta) \delta_0(dx) + \theta \delta_1(dx) \\ \ll \mu(dx) = \delta_0(dx) + \delta_1(dx) \text{ (mesure de comptage).}$$

- Puis

$$\frac{d\mathbb{P}_\theta}{d\mu}(x) = (1 - \theta) \mathbb{1}_{\{x=0\}} + \theta \mathbb{1}_{\{x=1\}} = \theta^x (1 - \theta)^{1-x}$$

avec $x \in \{0, 1\}$ (et 0 sinon), et

$$\frac{d\mathbb{P}_\theta^n}{d\mu^n}(x_1 \cdots x_n) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i},$$

avec $x_i \in \{0, 1\}$ (et 0 sinon).

Exemple 3 : temps de panne « arrêtés »

- On observe X_1, \dots, X_n , où $X_i = Y_i \wedge T$, avec Y_i lois exponentielles de paramètre θ et T temps fixe (censure).
- Cas 1 : $T = \infty$ (pas de censure). Alors $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et

$$\mathbb{P}_\theta(dx) = \theta \exp(-\theta x) 1_{\{x \geq 0\}} dx \ll \mu(dx) = dx$$

et

$$\frac{d\mathbb{P}_\theta^n}{d\mu^n}(x_1, \dots, x_n) = \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right),$$

avec $x_i \in \mathbb{R}_+$ (et 0 sinon).

- Cas 2 : Comment s'écrit le modèle dans la cas où $T < \infty$ (présence de censure) ? Comment choisir μ ?

Exemple : temps de panne « arrêtés »

- Loi $\mathbb{P}_\theta(dx)$ de $X = Y \wedge T$: $Y \sim$ exponentielle de paramètre θ :

$$X = Y1_{\{Y < T\}} + T1_{\{Y \geq T\}}$$

d'où

$$\begin{aligned}\mathbb{P}_\theta(dx) &= \theta e^{-\theta x} 1_{\{0 \leq x < T\}} dx + \left(\int_T^{+\infty} \theta e^{-\theta y} dy \right) \delta_T(dx) \\ &= \theta e^{-\theta x} 1_{\{0 \leq x < T\}} dx + e^{-\theta T} \delta_T(dx) \\ &\ll \mu(dx) = dx + \delta_T(dx) \quad (\text{par exemple}).\end{aligned}$$

Exemple : temps de panne « arrêtés » (fin)

- Alors, pour ce choix de mesure dominante

$$\frac{d\mathbb{P}_\theta}{d\mu}(x) = \theta e^{-\theta x} 1_{\{0 \leq x < T\}} + e^{-\theta T} 1_{\{x=T\}}$$

- Finalement,

$$\mathbb{P}_\theta^n(dx_1, \dots, dx_n) \ll \mu^n(dx_1 \dots dx_n) = \bigotimes_{i=1}^n [dx_i + \delta_T(dx_i)]$$

et

$$\begin{aligned} \frac{d\mathbb{P}_\theta^n}{d\mu^n}(x_1, \dots, x_n) &= \prod_{i=1}^n (\theta e^{-\theta x_i} 1_{\{0 \leq x_i < T\}} + e^{-\theta T} 1_{\{x_i=T\}}) \\ &= \theta^{N_n(T)} e^{-\theta \sum_{i=1}^n x_i 1_{\{x_i < T\}}} e^{-\theta T(n - N_n(T))}, \end{aligned}$$

avec $0 \leq x_i \leq T$ et 0 sinon, et $N_n(T) = \sum_{i=1}^n 1_{\{x_i < T\}}$

Méthodes d'estimation

- Méthode de substitution (ou des moments)
- Z -estimation
- M -estimation
- Le principe du **maximum de vraisemblance**

Méthode des moments : dimension 1

- $X_1, \dots, X_n \sim_{\text{i.i.d.}} \mathbb{P}_\theta$, avec $\theta \in \Theta \subset \mathbb{R}$.
- Principe : trouver $g : \mathbb{R} \rightarrow \mathbb{R}$ (en général $g(x) = x^k$) et $h : \mathbb{R} \rightarrow \mathbb{R}$ régulières de sorte que

$$\theta = h(\mathbb{E}_\theta [g(X)]) = h\left(\int_{\mathbb{R}} g(x) dF_\theta(x)\right) = T(F_\theta)$$

et T fonctionnelle régulière de la distribution inconnue F_θ .

- Estimateur : « plug-in »

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

Méthode des moments

- Précision d'estimation via les techniques empiriques :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, h'(\mathbb{E}_{\theta}[g(X)])^2 \text{Var}_{\theta}[g(X)])$$

en loi sous \mathbb{P}_{θ} et la variance asymptotique dépend en général de $\theta \rightarrow$ élimination par estimation préliminaire licite via le lemme de Slutsky.

- Exemple : $X_1, \dots, X_n \sim_{\text{i.i.d.}}$ exponentielle de paramètre θ . On a

$$\mathbb{E}_{\theta}[X] = \frac{1}{\theta},$$

l'estimateur par moment associé s'écrit

$$\hat{\theta}_n = \frac{1}{\bar{X}_n}.$$

Exemple en dimension $d > 1$

- $X_1, \dots, X_n \sim_{\text{i.i.d.}} \text{Béta}(\alpha, \beta)$, de densité

$$x \rightsquigarrow \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbf{1}_{\{0 < x < 1\}},$$

- Le paramètre est $\theta = (\alpha, \beta) \in \Theta = \mathbb{R}_+ \setminus \{0\} \times \mathbb{R}_+ \setminus \{0\}$.
- On a

$$\mathbb{E}_\theta [X] = \frac{\alpha}{\alpha + \beta}, \quad \mathbb{E}_\theta [X^2] = \frac{\alpha(\alpha + 1)}{(\alpha + \beta + 1)(\alpha + \beta)}$$

Exemple en dimension $d > 1$

- L'estimateur par moment $\hat{\theta}_n = (\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)})$ associé est défini par

$$\begin{cases} \bar{X}_n &= \frac{\hat{\theta}_n^{(1)}}{\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)}} \\ \frac{1}{n} \sum_{i=1}^n X_i^2 &= \frac{\hat{\theta}_n^{(1)}(\hat{\theta}_n^{(1)} + 1)}{(\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)} + 1)(\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)})}. \end{cases}$$

- Etude asymptotique via le TCL multidimensionnel et la méthode « delta » multidimensionnelle.

Limites de la méthode des moments

- Méthode **non systématique**
- Représentation pas toujours explicite
- Choix de la fonction g , notion d'optimalité parmi une classe d'estimateurs...
- **Généralisation** : Z -estimation (ou estimation par méthode des moments généralisés, GMM = *generalized method of moments*).

Z-estimation

- La méthode des moments (en dimension 1) est basée sur l'inversibilité de la fonction

$$m_g(\theta) = \int_{\mathbb{R}} g(x) \mathbb{P}_{\theta}(dx)$$

i.e. pour tout $\theta \in \Theta$

$$\int_{\mathbb{R}} (m_g(\theta) - g(x)) \mathbb{P}_{\theta}(dx) = 0.$$

- Principe de construction d'un Z-estimateur : remplacer $m_g(\theta) - g(x)$ par une fonction $\phi(\theta, x) : \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ arbitraire telle que

$$\forall \theta \in \Theta, \int_{\mathbb{R}} \phi(\theta, x) \mathbb{P}_{\theta}(dx) = 0.$$

Z-estimation

- Résoudre l'équation **empirique** associée :

$$\frac{1}{n} \sum_{i=1}^n \phi(a, X_i) = 0 \text{ pour } a \in \Theta.$$

Definition

On appelle **Z-estimateur** associé à ϕ tout estimateur $\hat{\theta}_n$ satisfaisant

$$\sum_{i=1}^n \phi(\hat{\theta}_n, X_i) = 0$$

- Il n'y a pas unicité de $\hat{\theta}_n$ (à ce niveau).
- Programme **Etablir des conditions** sur ϕ et sur la famille

Z-estimation : à quoi ça sert ?

- Exemple. $\Theta = \mathbb{R}$, $\mathbb{P}_\theta(dx) = f(x - \theta)dx$, et f symétrique : $\overline{f(-x)} = f(x)$, $\forall x \in \mathbb{R}$.
- Il n'y a pas de bornitude des moments !
- On pose

$$\phi(a, x) = \text{Arctg}(x - a).$$

- La fonction

$$a \rightsquigarrow \mathbb{E}_\theta [\phi(a, X)] = \int_{\mathbb{R}} \text{Arctg}(x - a) f(x - \theta) dx$$

est strictement décroissante et s'annule seulement en $a = \theta$.

- Z-estimateur associé : solution $\hat{\theta}_n$ de

$$\sum_{i=1}^n \text{Arctg}(X_i - \hat{\theta}_n) = 0$$

(unicité).

Le cas multidimensionnel

Si $\Theta \subset \mathbb{R}^d$ avec $d > 1$, la fonction ϕ est remplacée par

$$\Phi = (\phi_1, \dots, \phi_d) : \Theta \times \mathbb{R} \rightarrow \mathbb{R}^d.$$

Definition

On appelle Z-estimateur associé à Φ tout estimateur $\hat{\theta}_n$ satisfaisant

$$\sum_{i=1}^n \phi_\ell(\hat{\theta}_n, X_i) = 0, \quad \ell = 1, \dots, d.$$

Z-estimation \rightarrow M-estimation

- En dimension 1 : si

$$\phi(\theta, x) = \partial_{\theta} \psi(\theta, x)$$

pour une certaine fonction ψ , résoudre $\sum_{i=1}^n \phi(\theta, X_i) = 0$
revient à **chercher un point critique** de

$$\theta \rightsquigarrow \sum_{i=1}^n \psi(\theta, X_i).$$

- En dimension $d \geq 1$, il faut $\phi(\theta, x) = \nabla_{\theta} \psi(\theta, x)$ (moins facile à obtenir).
- **Invite à généraliser** la recherche d'estimateurs via la maximisation d'un critère \rightarrow M-estimation.

M-estimation

- Principe : Se donner une application $\psi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}_+$ telle que, pour tout $\theta \in \Theta \subset \mathbb{R}^d$,

$$a \rightsquigarrow \mathbb{E}_\theta [\psi(a, X)] = \int \psi(a, x) \mathbb{P}_\theta(dx)$$

admet un maximum en $a = \theta$.

Definition

On appelle *M-estimateur* associé à ψ tout estimateur $\hat{\theta}_n$ satisfaisant

$$\sum_{i=1}^n \psi(\hat{\theta}_n, X_i) = \max_{a \in \Theta} \sum_{i=1}^n \psi(a, X_i).$$

- Il n'y a pas unicité de $\hat{\theta}_n$ (à ce niveau).

Un exemple classique : paramètre de localisation

- $\Theta = \mathbb{R}$, $\mathbb{P}_\theta(dx) = f(x - \theta)dx$, et $\int_{\mathbb{R}} xf(x)dx = 0$,
 $\int_{\mathbb{R}} x^2 \mathbb{P}_\theta(dx) < +\infty$ pour tout $\theta \in \mathbb{R}$. On pose

$$\psi(a, x) = -(a - x)^2$$

- La fonction

$$a \rightsquigarrow \mathbb{E}_\theta [\psi(a, X)] = - \int_{\mathbb{R}} (a - X)^2 f(x - \theta) dx$$

admet un **maximum** en $a = \mathbb{E}_\theta [X] = \int_{\mathbb{R}} xf(x - \theta)dx = \theta$.

- **M-estimateur associé :**

$$\sum_{i=1}^n (X_i - \hat{\theta}_n)^2 = \min_{a \in \mathbb{R}} \sum_{i=1}^n (X_i - a)^2.$$

Paramètre de localisation

- C'est **aussi** un Z -estimateur associé à $\phi(a, x) = 2(x - a)$: on résout

$$\sum_{i=1}^n (a - X_i) = 0 \quad \text{d'où} \quad \hat{\theta}_n = \bar{X}_n.$$

- Dans cet **exemple très simple**, tous les points de vue coïncident.
- Si, dans le même contexte, $\int_{\mathbb{R}} x^2 \mathbb{P}_{\theta}(dx) = +\infty$ et $f(x) = f(-x)$, on peut utiliser Z -estimateur avec $\phi(a, x) = \text{Arctg}(x - a)$. Méthode robuste, mais est-elle optimale ? Peut-on faire mieux **si f est connue ? A suivre...**

Lien entre Z - et M - estimateurs

- Pas d'inclusion entre ces deux classes d'estimateurs en général :
 - Si ψ non-régulière, M -estimateur \nRightarrow Z -estimateur
 - Si une équation d'estimation admet plusieurs solutions distinctes, Z -estimateur \nRightarrow M -estimateur (cas d'un extremum local).
- Toutefois, si ψ est régulière, les M -estimateurs sont des Z -estimateurs : si $\Theta \subset \mathbb{R}$ ($d = 1$), en posant

$$\phi(a, x) = \partial_a \psi(a, x),$$

on a

$$\sum_{i=1}^n \partial_a \psi(\theta, X_i) \Big|_{a=\hat{\theta}_n} = \sum_{i=1}^n \phi(\hat{\theta}_n, X_i) = 0.$$

Maximum de vraisemblance

- Principe **fondamental** et **incontournable** en statistique. Cas particuliers connus depuis le XVIIIème siècle. Définition générale : Fisher (1922).
- Fournit une première **méthode systématique** de construction d'un M -estimateur (souvent un Z -estimateur, souvent aussi *a posteriori* un estimateur par substitution simple).
- Procédure **optimale** (dans quel sens ?) sous des hypothèses de **régularité** de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ (Cours 6).
- Parfois difficile à mettre en oeuvre en pratique → **méthodes numériques**, statistique computationnelle.

Fonction de vraisemblance

- La famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ est dominée par une mesure σ -finie μ .
On se donne, pour $\theta \in \Theta$

$$f(\theta, x) = \frac{d\mathbb{P}_\theta}{d\mu}(x), \quad x \in \mathbb{R}.$$

Definition

Fonction de vraisemblance du n -échantillon associée à la famille $\{f(\theta, \cdot), \theta \in \Theta\}$:

$$\theta \rightsquigarrow \mathcal{L}_n(\theta, X_1, \dots, X_n) = \prod_{i=1}^n f(\theta, X_i)$$

- C'est une fonction aléatoire (définie μ -presque partout).

Exemples

- Exemple 1 : **Modèle de Poisson**. On observe

$$X_1, \dots, X_n \sim_{\text{i.i.d.}} \text{Poisson}(\theta),$$

$\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et prenons $\mu(dx) = \sum_{k \in \mathbb{N}} \delta_k(dx)$.

- La densité de \mathbb{P}_θ par rapport à μ est

$$f(\theta, x) = e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, 2, \dots$$

- La **fonction de vraisemblance** associée s'écrit

$$\begin{aligned} \theta \rightsquigarrow \mathcal{L}_n(\theta, X_1, \dots, X_n) &= \prod_{i=1}^n e^{-\theta} \frac{\theta^{X_i}}{X_i!} \\ &= \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}. \end{aligned}$$

Exemples

- Exemple 2 **Modèle de Cauchy**. On observe

$$X_1, \dots, X_n \sim_{\text{i.i.d.}} \text{Cauchy},$$

$$\theta \in \Theta = \mathbb{R} \text{ et } \mu(dx) = dx \text{ (par exemple).}$$

- On a alors

$$\mathbb{P}_{\theta}(dx) = f(\theta, x)dx = \frac{1}{\pi(1 + (x - \theta)^2)}dx.$$

- La **fonction de vraisemblance** associée s'écrit

$$\theta \rightsquigarrow \mathcal{L}_n(\theta, X_1, \dots, X_n) = \frac{1}{\pi^n} \prod_{i=1}^n (1 + (X_i - \theta)^2)^{-1}.$$

Principe de maximum de vraisemblance

- Cas d'une famille de lois **restreinte à deux points**

$$\Theta = \{\theta_1, \theta_2\} \subset \mathbb{R},$$

avec \mathbb{P}_{θ_i} discrète et $\mu(dx)$ la mesure de comptage.

- **A priori**, pour tout (x_1, \dots, x_n) , et pour $\theta \in \{\theta_1, \theta_2\}$,

$$\begin{aligned}\mathbb{P}_{\theta} [X_1 = x_1, \dots, X_n = x_n] &= \prod_{i=1}^n \mathbb{P}_{\theta} [X_i = x_i] \\ &= \prod_{i=1}^n f(\theta, x_i).\end{aligned}$$

La probabilité d'avoir la réalisation fixée (x_1, \dots, x_n) .

Principe de maximum de vraisemblance

- A posteriori, on observe (X_1, \dots, X_n) . L'événement

$$\left\{ \prod_{i=1}^n f(\theta_1, X_i) > \prod_{i=1}^n f(\theta_2, X_i) \right\} \quad (\text{Cas 1})$$

ou bien l'événement

$$\left\{ \prod_{i=1}^n f(\theta_2, X_i) > \prod_{i=1}^n f(\theta_1, X_i) \right\} \quad (\text{Cas 2})$$

est réalisé. (On ignore le cas d'égalité.)

- Principe de maximum de vraisemblance :

$$\hat{\theta}_n^{\text{mv}} = \theta_1 1_{\{\text{Cas 1}\}} + \theta_2 1_{\{\text{Cas 2}\}}.$$

Estimateur du maximum de vraisemblance

- On généralise le principe précédent pour une famille de lois et un ensemble de paramètres **quelconques**.
- Situation : $X_1, \dots, X_n \sim_{\text{i.i.d.}} \mathbb{P}_\theta$, $\{\mathbb{P}_\theta, \theta \in \Theta\}$ dominée, $\Theta \subset \mathbb{R}^d$, $\theta \rightsquigarrow \mathcal{L}_n(\theta, X_1, \dots, X_n)$ vraisemblance associée.

Definition

On appelle **estimateur du maximum de vraisemblance** tout estimateur $\hat{\theta}_n^{\text{mv}}$ satisfaisant

$$\mathcal{L}_n(\hat{\theta}_n^{\text{mv}}, X_1, \dots, X_n) = \max_{\theta \in \Theta} \mathcal{L}_n(\theta, X_1, \dots, X_n).$$

- Existence, unicité...

Remarques

■ Log-vraisemblance :

$$\begin{aligned}\theta \rightsquigarrow \ell_n(\theta, X_1, \dots, X_n) &= \log \mathcal{L}_n(\theta, X_1, \dots, X_n) \\ &= \sum_{i=1}^n \log f(\theta, X_i).\end{aligned}$$

Bien défini si $f(\theta, \cdot) > 0$ μ -pp.

Max. vraisemblance = max. log-vraisemblance.

- L'estimateur du maximum de vraisemblance **ne dépend pas** du choix de la mesure dominante μ .
- Notion de **racine de l'équation de vraisemblance** : tout estimateur $\hat{\theta}_n^{\text{rv}}$ vérifiant

$$\nabla_{\theta} \ell_n(\hat{\theta}_n^{\text{rv}}, X_1, \dots, X_n) = 0.$$

Modèle binomial

L'expérience statistique est générée par un n -échantillon de loi de Bernoulli de paramètre $\theta \in \Theta = [0, 1]$.

■ Vraisemblance

$$\mathcal{L}_n(\theta) = \prod_{i=1} \theta^{X_i} (1 - \theta)^{1-X_i} = \theta^{n\bar{X}_n} (1 - \theta)^{n(1-\bar{X}_n)}.$$

où $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ est la moyenne empirique.

■ Log-vraisemblance

$$\ell_n(\theta) = n\bar{X}_n \log(\theta) + n(1 - \bar{X}_n) \log(1 - \theta).$$

Modèle binomial

- Equations de vraisemblance : pour $\theta \in (0, 1)$,

$$\frac{n\bar{X}_n}{\theta} - \frac{n(1 - \bar{X}_n)}{1 - \theta} = 0$$

- Si $0 < \bar{X}_n < 1$, cette équation admet une solution unique, \bar{X}_n .
- Si $\bar{X}_n = 0$, alors $\mathcal{L}_n(\theta) = (1 - \theta)^n$: la vraisemblance est maximum en $\theta = 0$.
- Si $\bar{X}_n = 1$ alors $\mathcal{L}_n(\theta) = \theta^n$: la vraisemblance est maximum en $\theta = 1$.

Exemple : modèle normal

L'expérience statistique est engendrée par un n -échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, le paramètre est $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$.

■ Vraisemblance

$$\mathcal{L}_n((\mu, \sigma^2), X_1, \dots, X_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\right).$$

■ Log-vraisemblance

$$\ell_n((\mu, \sigma^2), X_1, \dots, X_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

Exemple : modèle normal

Equation(s) de vraisemblance

$$\begin{cases} \partial_{\mu} \ell_n((\mu, \sigma^2), X_1, \dots, X_n) &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \\ \partial_{\sigma^2} \ell_n((\mu, \sigma^2), X_1, \dots, X_n) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2. \end{cases}$$

Solution de ces équations (pour $n \geq 2$) :

$$\hat{\theta}_n^{\text{rv}} = \left(\bar{X}_n, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right)$$

et on vérifie que $\hat{\theta}_n^{\text{rv}} = \hat{\theta}_n^{\text{mv}}$.

Exemple : modèle de Poisson

■ Vraisemblance

$$\mathcal{L}_n(\theta, X_1, \dots, X_n) = \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}.$$

■ Log-vraisemblance

$$\ell_n(\theta, X_1, \dots, X_n) = c(X_1, \dots, X_n) - n\theta + \sum_{i=1}^n X_i \log \theta.$$

■ Equation de vraisemblance

$$-n + \sum_{i=1}^n X_i \frac{1}{\theta} = 0, \text{ soit } \boxed{\hat{\theta}_n^{\text{rv}} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n}$$

et on vérifie que $\hat{\theta}_n^{\text{rv}} = \hat{\theta}_n^{\text{mv}}$.

Exemple : modèle de Laplace

L'expérience statistique est engendrée par un n -échantillon de loi de Laplace de paramètre $\theta \in \Theta = \mathbb{R}$. La densité par rapport à la mesure de Lebesgue :

$$f(\theta, x) = \frac{1}{2\sigma} \exp\left(-\frac{|x - \theta|}{\sigma}\right),$$

où $\sigma > 0$ est **connu**.

■ Vraisemblance

$$\mathcal{L}_n(\theta, X_1, \dots, X_n) = (2\sigma)^{-n} \exp\left(-\frac{1}{\sigma} \sum_{i=1}^n |X_i - \theta|\right)$$

■ Log-vraisemblance

$$\ell_n(\theta, X_1, \dots, X_n) = -n \log(2\sigma) - \frac{1}{\sigma} \sum_{i=1}^n |X_i - \theta|.$$

Exemple : modèle de Laplace

Maximiser $\mathcal{L}_n(\theta, X_1, \dots, X_n)$ revient à minimiser la fonction $\theta \rightsquigarrow \sum_{i=1}^n |X_i - \theta|$, dérivable presque partout de dérivée constante par morceaux. **Equation de vraisemblance :**

$$\sum_{i=1}^n \text{sign}(X_i - \theta) = 0.$$

Soit $X_{(1)} \leq \dots \leq X_{(n)}$ la statistique d'ordre.

- n pair : $\hat{\theta}_n^{\text{mv}}$ **n'est pas unique** ; tout point de l'intervalle $[X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}]$ est un EMV.
- n impair : $\hat{\theta}_n^{\text{mv}} = X_{(\frac{n+1}{2})}$, l'EMV est unique. Mais $\hat{\theta}_n^{\text{rv}}$ n'existe pas.
- **pour tout** n , la médiane empirique est un EMV.

Exemple : modèle de Cauchy

■ Vraisemblance

$$\mathcal{L}_n(\theta, X_1, \dots, X_n) = \pi^{-n} \prod_{i=1}^n \frac{1}{1 + (X_i - \theta)^2}.$$

■ Log-vraisemblance

$$\ell_n(\theta, X_1, \dots, X_n) = -n \log \pi - \sum_{i=1}^n \log (1 + (X_i - \theta)^2)$$

■ Equation de vraisemblance

$$\sum_{i=1}^n \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0$$

pas de solution explicite et admet en général plusieurs solutions.

Maximum de vraisemblance = M -estimateur

- Une inégalité de convexité : μ mesure σ -finie sur \mathbb{R} ; f, g deux densités de probabilités par rapport à μ . Alors

$$\int_{\mathbb{R}} f(x) \log f(x) \mu(dx) \geq \int_{\mathbb{R}} f(x) \log g(x) \mu(dx)$$

(si les intégrales sont finies) avec égalité ssi $f = g$ μ -pp.

- Preuve : à montrer

$$\int_{\mathbb{R}} f(x) \log \frac{g(x)}{f(x)} \mu(dx) \leq 0.$$

(avec une convention de notation appropriée)

Une inégalité de convexité

- On a $\log(1 + x) \leq x$ pour $x \geq -1$ avec égalité ssi $x = 0$.
- Donc

$$\log \frac{g(x)}{f(x)} = \log \left(1 + \left(\frac{g(x)}{f(x)} - 1 \right) \right) \leq \frac{g(x)}{f(x)} - 1$$

(avec égalité ssi $f(x) = g(x)$).

- Finalement

$$\begin{aligned} \int_{\mathbb{R}} f(x) \log \frac{g(x)}{f(x)} \mu(dx) &\leq \int_{\mathbb{R}} f(x) \left(\frac{g(x)}{f(x)} - 1 \right) \mu(dx) \\ &= \int_{\mathbb{R}} g(x) \mu(dx) - \int_{\mathbb{R}} f(x) \mu(dx) \\ &= 0. \end{aligned}$$

Conséquence pour l'EMV

- On pose

$$\psi(a, x) := \log f(a, x), \quad a \in \Theta, \quad x \in \mathbb{R}$$

(avec une convention pour le cas où on n'a pas $f(a, \cdot) > 0$.)

- La fonction

$$a \rightsquigarrow \mathbb{E}_\theta [\psi(a, X)] = \int_{\mathbb{R}} \log f(a, x) f(\theta, x) \mu(dx)$$

a un maximum en $a = \theta$ d'après **l'inégalité de convexité**.

- Le M -estimateur associé à ψ maximise la fonction

$$a \rightsquigarrow \sum_{i=1}^n \log f(a, X_i) = \ell_n(a, X_1, \dots, X_n)$$

c'est-à-dire la **log-vraisemblance**. C'est l'**estimateur du maximum de vraisemblance**.

- C'est aussi un Z -estimateur si la fonction $\theta \rightsquigarrow \log f(\theta, \cdot)$ est régulière, associé à la fonction

$$\phi(\theta, x) = \partial_{\theta} \log f(\theta, x) = \frac{\partial_{\theta} f(\theta, x)}{f(\theta, x)}, \quad \theta \in \Theta, x \in \mathbb{R}$$

lorsque $\Theta \subset \mathbb{R}$, à condition que le maximum de log-vraisemblance n'est pas atteint sur la frontière de Θ . (Se généralise en dimension d .)

Un M -estimateur qui n'est pas un Z -estimateur

- On observe $X_1, \dots, X_n \sim_{\text{i.i.d.}}$ uniformes sur $[0, \theta]$,
 $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$.

- On a

$$\mathbb{P}_\theta(dx) = \theta^{-1} 1_{[0, \theta]}(x) dx$$

et

$$\begin{aligned} \mathcal{L}_n(\theta, X_1, \dots, X_n) &= \theta^{-n} \prod_{i=1}^n 1_{[0, \theta]}(X_i) \\ &= \theta^{-n} 1_{\{\max_{1 \leq i \leq n} X_i \leq \theta\}} \end{aligned}$$

- La fonction de vraisemblance n'est pas régulière.
- L'estimateur du maximum de vraisemblance est
 $\hat{\theta}_n^{\text{mv}} = \max_{1 \leq i \leq n} X_i$.

Estimation des paramètres de la loi Gamma

- Soit (X_1, X_2, \dots, X_n) n observations i.i.d. de loi Gamma($\theta = (\alpha, \beta) \in \Theta = (\mathbb{R}_+^* \times \mathbb{R}_+^*)$)

$$f_{\theta}(x) = \Gamma(\alpha)^{-1} \beta^{\alpha} x^{\alpha-1} e^{-\beta x}.$$

- On montre aisément que

$$\alpha = \frac{(\mathbb{E}_{\theta}[X_1])^2}{\text{Var}_{\theta}(X_1)} \quad \text{et} \quad \beta = \frac{\mathbb{E}_{\theta}[X_1]}{\text{Var}_{\theta}(X_1)}$$

- Estimateurs de moments

$$\hat{\alpha}_n = \frac{\bar{X}_n^2}{S_n^2} \quad \text{et} \quad \hat{\beta}_n = \frac{\bar{X}_n}{S_n^2}$$

où $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ est la **moyenne empirique** et $S_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ est la **variance empirique**.

Estimateur du maximum de vraisemblance

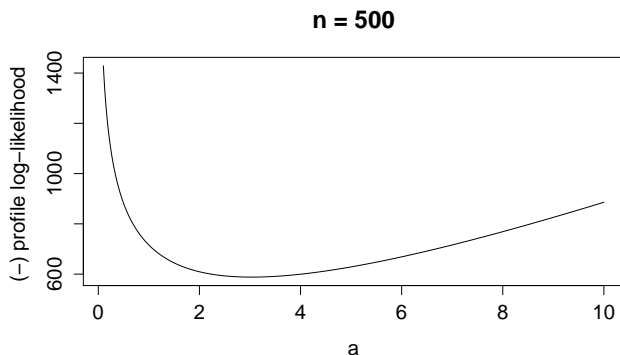
■ log-vraisemblance

$$\ell_n(\alpha, \beta) = -n \log \Gamma(\alpha) + n\alpha \log(\beta) + (\alpha - 1) \sum_{i=1}^n \log(X_i) - \beta \sum_{i=1}^n X_i.$$

Le maximum ne se calcule pas explicitement.

- La minimisation par rapport à β pour α fixée est explicite : $\hat{\beta}_n(\alpha) = \alpha / \bar{X}_n$. L'estimateur du MV est obtenu en maximisant par rapport à α la fonction $\alpha \mapsto \ell_n(\alpha, \hat{\beta}_n(\alpha))$.
- On apprendra bientôt que l'estimateur du maximum de vraisemblance est préférable à l'estimateur des moments.

Vraisemblance réduite



Boxplot

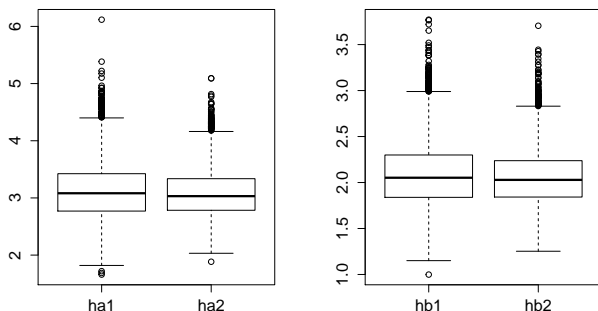


FIGURE – Boxplot : paramètres $\alpha = 3$, $\beta = 2$, $n = 100$, 5000 réplifications indépendantes

Distribution asymptotique - $n = 500$

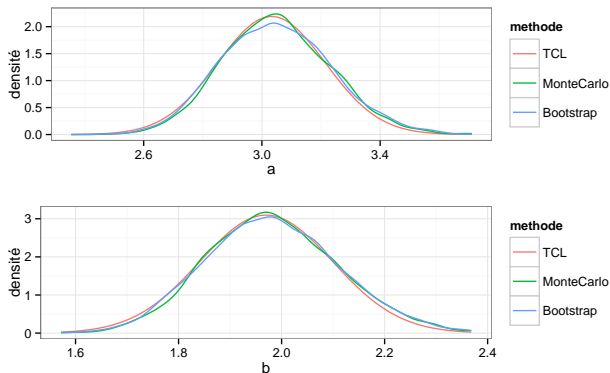


FIGURE – Boxplot : paramètres $\alpha = 3$, $\beta = 2$, $n = 500$, distribution asymptotique, bootstrap et bootstrap paramétrique

Distribution asymptotique - $n = 20$

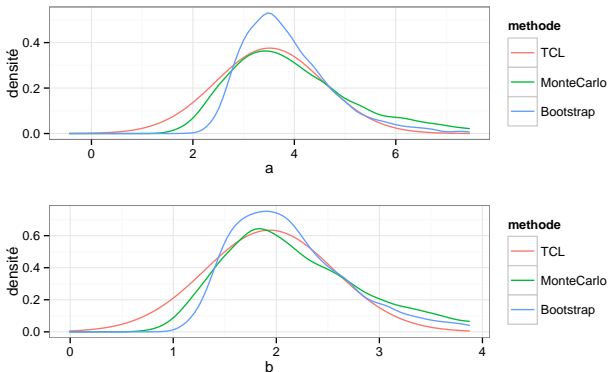


FIGURE – Boxplot : paramètres $\alpha = 3$, $\beta = 2$, $n = 500$, distribution asymptotique, bootstrap et bootstrap paramétrique