Siahaan--Gensollen Rémy
Geron Alban

Statistique 1

Siahaan–Gensollen Rémy
Géron Alban

Statistique 1

Table des matières

Introduction

Ce cours est le résultat de notes prises lors du cours de Statistique 1 enseigné par Arnak Dalalyan au premier semestre de deuxième année du cycle ingénieur de l'ENSAE Paris, ainsi que du travail que nous avons effectué par la suite pour les réarranger en un document unique. Nous avons tenté d'être exhaustifs, sans nous attarder sur trop de détails ou nuances. Malgré tout, certaines sections sont peut-être trop peu fournies. Aussi, nous ne saurions que trop vous conseiller de prendre des notes de votre côté.

La structure visuelle est normalement assez claire, mais ne soyez pas surpris si elle change de temps à autre au fil du document : il s'agit d'un projet réalisé en parallèle des cours, et il n'est aucunement prévu que ce document soit publié (ailleurs que sur le drive commun de l'école). Ceci étant dit, nous espérons que ce cours vous sera utile. Si vous avez des remarques pour l'améliorer ou si vous repérez des erreurs, n'hésitez pas à nous contacter. Bonne lecture !

Convention. Sauf indication contraire, la loi $\mathcal{E}(\theta)$ pour $\theta > 0$ désigne la *loi exponentielle de moyenne* θ, c'est-à-dire la loi de densité

$$x \longmapsto \frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{[0,+\infty[}(x)$$

I Rappels de probabilités

Dans tout ce chapitre, d est un entier naturel non nul, $(X_n)_{n\geq 1}$ est une suite de vecteurs aléatoires dans $\mathbb{R}^d$ et X un vecteur aléatoire dans $\mathbb{R}^d$.

Définition I.1 Convergences

On dit que $(X_n)_{n\geq 1}$ converge...

- ... presque sûrement vers X lorsqu'il existe un événement $\Omega_0 \subset \Omega$ de probabilité 1 tel que

$$\forall \omega \in \Omega_0, \qquad X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega)$$

On note alors $X_n \xrightarrow[n\to+\infty]{\text{p.s.}} X$.

- ... dans L^p ($p \geq 1$) vers X lorsque X et tous les X_n admettent un moment d'ordre p et

$$\mathbb{E}[\|X_n - X\|^p] \xrightarrow[n\to+\infty]{} 0$$

On note alors $X_n \xrightarrow[n\to+\infty]{L^p} X$.

- ... en probabilité vers X lorsque

$$\forall \varepsilon > 0, \qquad P(\|X_n - X\| \geq \varepsilon) \xrightarrow[n\to+\infty]{} 0$$

On note alors $X_n \xrightarrow[n\to+\infty]{P} X$.

- ... en loi / en distribution vers X lorsque pour toute fonction $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ continue et bornée, on a

$$\mathbb{E}[f(X_n)] \xrightarrow[n\to+\infty]{} \mathbb{E}[f(X)]$$

On note alors $X_n \xrightarrow[n\to+\infty]{\text{loi}} X$.

Remarques.

— La Figure 1 rappelle les implications entre les différents modes de convergence.

— La définition de la convergence presque sûre est bien équivalente à celle du cours de probabilité de première année[1] :

$$X_n \xrightarrow[n\to+\infty]{\text{p.s.}} X \iff P\left(\left\{\omega \in \Omega : X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega)\right\}\right) = 1$$

En effet :

$\boxed{\Leftarrow}$ $\Omega_0 = \left\{\omega \in \Omega : X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega)\right\}$ convient, c'est bien un événement comme prouvé dans le cours de Brunel.

$\boxed{\Rightarrow}$ Puisque $\forall \omega \in \Omega_0$, $X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega)$, on a :

$$\Omega_0 \subset \left\{\omega \in \Omega : X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega)\right\}$$

[1] Voir cours de *Théories des probabilités*.

donc

$$1 = P(\Omega_0) \leq P\left(\left\{\omega \in \Omega : X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega)\right\}\right) \leq 1$$

— L'inégalité large dans la définition de la convergence en probabilité peut être remplacée par une égalité stricte. En d'autres termes, les deux assertions suivantes sont équivalentes :

(i) $\forall \varepsilon > 0, \qquad P(\|X_n - X\| \geq \varepsilon) \xrightarrow[n\to+\infty]{} 0\,;$

(ii) $\forall \varepsilon > 0, \qquad P(\|X_n - X\| > \varepsilon) \xrightarrow[n\to+\infty]{} 0.$

$$\begin{tikzcd}
X_n \xrightarrow[n\to+\infty]{\text{p.s.}} X \arrow[r, Rightarrow] \arrow[ddr, Rightarrow, "\textit{(avec hyp. du TCD)}"'] & X_n \xrightarrow[n\to+\infty]{P} X \arrow[r, Rightarrow, "\textit{(réciproque vraie si } X = c \textit{ p.s.)}"] & X_n \xrightarrow[n\to+\infty]{\text{loi}} X \\
& X_n \xrightarrow[n\to+\infty]{L^p} X \arrow[u, Rightarrow] & \\
& X_n \xrightarrow[n\to+\infty]{L^q} X \arrow[u, Rightarrow] &
\end{tikzcd}$$

Figure 1 *Liens entre les modes de convergence. Ici c est un vecteur fixé de $\mathbb{R}^d$, et p et q deux réels tels que $1 \leq p \leq q$.*

Lorsque $d = 1$, on n'utilise pas la plupart du temps la définition de la convergence en loi mais plutôt la propriété suivante :

Propriété I.2

Supposons que $d = 1$. Notons F_{X_n} (resp. F_X) la fonction de répartition de X_n (resp. X).
Les deux assertions suivantes sont équivalentes :

(i) $(X_n)_{n\geq 1}$ converge en loi vers X ;

(ii) Pour tout réel t en lequel F_X est continue (ce qui revient à dire que t n'est pas un atome de X), on a

$$F_{X_n}(t) \xrightarrow[n\to+\infty]{} F_X(t)$$

Théorème I.3 Loi forte des grands nombres

Supposons que les $(X_n)_{n\geq 1}$ sont i.i.d. et que $\mathbb{E}[\|X_1\|] < +\infty$.

On note pour tout $n \geq 1$, $\bar{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. Alors

$$\bar{X}_n \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbb{E}[X_1]$$

```python
import numpy as np
f = lambda x: 1 / (x ** 5 + 1)
U = np.random.random(10_000)
m = np.mean([f(u) for u in U])
print(m)

>>> 0.888046176564409
```

Figure 2 *Approximation de l'intégrale $\int_0^1 \frac{1}{x^5+1}dx$ par la méthode de Monte-Carlo.*

Application : méthode de Monte-Carlo. Si $f : [0,1]^d \longrightarrow \mathbb{R}$ est intégrable, on peut approximer $\int_{[0,1]^d} f(x)dx$ facilement. En effet cette intégrale est égale à $\mathbb{E}[f(U)]$ où $U \sim \mathcal{U}([0,1]^d)$ donc il suffit de générer un grand nombre n de réalisations $U_1, \ldots, U_n$ i.i.d. de la loi $\mathcal{U}([0,1]^d)$, et une approximation de l'intégrale est $\frac{1}{n}\sum_{i=1}^{n} f(U_i)$.

Théorème I.4 Théorème central limite

Supposons que les $(X_n)_{n\geq 1}$ sont i.i.d. et que $\mathbb{E}[\|X_1\|^2] < +\infty$.
Alors
$$\sqrt{n}\big(\bar{X}_n - \mathbb{E}[X_1]\big) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \operatorname{Var}(X_1))$$

Interprétation. Si la matrice $\operatorname{Var}(X_1)$ est inversible,[2] alors cette convergence se réécrit :
$$\sqrt{n}\operatorname{Var}(X_1)^{-1/2}\big(\bar{X}_n - \mathbb{E}[X_1]\big) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, I_d)$$

Autrement dit, lorsqu'on centre et qu'on réduit $\bar{X}_n$, la suite obtenue converge en loi vers une gaussienne centrée réduite. De façon plus informelle, lorsque n est grand, la loi de $\bar{X}_n$ peut être approchée par $\mathcal{N}\Big(\mathbb{E}[X_1], \frac{\operatorname{Var}(X_1)}{n}\Big)$ (voir Figure 3) : la moyenne empirique se concentre autour de la moyenne théorique à une vitesse $1/\sqrt{n}$. Peut se formaliser sous forme d'une inégalité :

[2] Correspond à l'hypothèse d'une variance non nulle dans le cas $d = 1$.

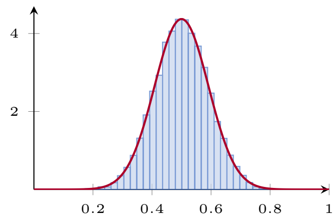

Figure 3 *Histogramme de 10 000 réalisations de la variable* $\bar{X}_{10}$ *où* $X_1, \dots, X_{10} \overset{i.i.d.}{\sim} \mathcal{U}([0,1])$. *On voit que l'histogramme se superpose avec la densité de la loi* $\mathcal{N}\big(\frac{1}{2}, \frac{1}{120}\big) = \mathcal{N}\Big(\mathbb{E}[X_1], \frac{\operatorname{Var}(X_1)}{10}\Big)$.

Théorème I.5 Inégalité de Berry-Esseen (hors programme)

Supposons que $d = 1$. Si $X_1, X_2, \dots$ sont i.i.d. admettant un moment jusqu'à l'ordre 3, et si Φ est la f.d.r. de la loi $\mathcal{N}(0,1)$, alors
$$\forall t \in \mathbb{R}, \quad |F_{\bar{X}_n}(t) - \Phi(t)| \leq \frac{0.5\mathbb{E}[|X_1 - \mu|^3]}{\sqrt{n}\operatorname{Var}(X_1)^{3/2}}$$

QCM de convergence.

1. Est-ce que si X_n converge p.s. vers X alors
$$\forall A \in \mathcal{B}(\mathbb{R}), \quad P(X_n \in A) \xrightarrow[n\to+\infty]{} P(X \in A) \quad ?$$
Non. Considérer $X_n = 1/n$ et $A =]0, +\infty[$.
En revanche ça marche si X_n converge en loi vers X et qu'on impose à A d'être un segment de $\mathbb{R}$.
2. Est-ce que si X_n converge p.s. vers X et si les X_n sont discrètes alors X aussi ?
Non. Prendre $X_n \sim \mathcal{U}\big(\{0, \frac{1}{n}, \dots, 1\}\big)$.
3. Est-ce que si X_n converge p.s. vers X et si les X_n sont à densité alors X aussi ?
Non. Prendre $X_n \sim \mathcal{E}(1/n)$.
4. Est-il vrai que, de manière informelle, $\mathcal{B}(n,\theta) \overset{\text{loi}}{\approx} \mathcal{N}(n\theta, n\theta(1-\theta))$?
Oui. En effet, si $S_n \sim \mathcal{B}(n,\theta)$, alors S_n a la même loi que $X_1 + \dots + X_n$ où les X_i sont indépendantes de loi $\mathcal{B}(\theta)$. Par le théorème central limite,
$$Y_n := \sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}[X_1]\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \theta(1-\theta))$$
Ainsi, la loi de Y_n pour des grandes valeurs de n est proche de la loi $\mathcal{N}(0, \theta(1-\theta))$. Comme $S_n = n\theta + \sqrt{n}Y_n$, on a $\mathcal{B}(n,\theta) \overset{\text{loi}}{\approx} \mathcal{N}(n\theta, n\theta(1-\theta))$.

Théorème I.6 1er théorème de continuité

La composition par une fonction continue conserve la convergence presque sûre, en probabilité et en loi.

Preuve. Soit $g : \mathbb{R}^d \longrightarrow \mathbb{R}^q$ une fonction continue.

1. Supposons que $X_n \xrightarrow[n\to+\infty]{\text{p.s.}} X$.
   Alors, il existe $\Omega_0 \subset \Omega$ tel que $P(\Omega_0) = 1$ et
   $$\forall \omega \in \Omega_0, \quad X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega)$$
   Donc $P(\Omega_0) = 1$ et pour tout $\omega \in \Omega_0$, $g(X_n(\omega)) \xrightarrow[n\to+\infty]{} g(X(\omega))$ par continuité de g.
   Ainsi $g(X_n) \xrightarrow[n\to+\infty]{\text{p.s.}} g(X)$.
2. Cas particulier. On suppose que $X_n \xrightarrow[n\to+\infty]{P} a$ où a est *déterministe.*
   Soit $\delta > 0$. Montrons que
   $$P(\|g(X_n) - g(a)\| > \delta) \xrightarrow[n\to+\infty]{} 0$$
   Comme g est continue en a, il existe $\varepsilon > 0$ tel que
   $$\|x - a\| \leq \varepsilon \implies \|g(x) - g(a)\| \leq \delta$$
   i.e.
   $$\|g(x) - g(a)\| > \delta \implies \|x - a\| > \varepsilon$$
   donc
   $$\big\{\omega \in \Omega : \|g(X_n(\omega)) - g(a)\| > \delta\big\} \subset \big\{\omega \in \Omega : \|X_n(\omega) - a\| > \varepsilon\big\}$$
   d'où
   $$P(\|g(X_n) - g(a)\| > \delta) \leq P(\|X_n - a\| > \varepsilon) \xrightarrow[n\to+\infty]{} 0$$
   Donc $g(X_n) \xrightarrow[n\to+\infty]{P} g(a)$.
3. On suppose que $X_n \xrightarrow[n\to+\infty]{\text{loi}} X$.
   Soit h continue bornée, alors $h \circ g$ est continue et bornée donc
   $$\mathbb{E}[h \circ g(X_n)] \xrightarrow[n\to+\infty]{} \mathbb{E}[h \circ g(X)]$$
   d'où $g(X_n) \xrightarrow[n\to+\infty]{\text{loi}} g(X)$.

■

Propriété I.7 Convergence coordonnée par coordonnée

Les deux assertions suivantes sont équivalentes :

(i) $X_n \xrightarrow[n\to+\infty]{\text{p.s.}} X$;

(ii) Pour chaque $j = 1, \ldots, d$, on a $X_n^{(j)} \xrightarrow[n\to+\infty]{\text{p.s.}} X^{(j)}$.

On a le même résultat en remplaçant les convergences presque sûres par des convergences en probabilité. Idem pour la convergence L^p (sous réserve que X et les X_n admettent des moment d'ordre p).

En revanche, pour la convergence en loi, on a seulement l'implication

$$X_n \xrightarrow[n\to+\infty]{\text{loi}} X \quad \implies \quad \forall j = 1, \ldots, d,\ X_n^{(j)} \xrightarrow[n\to+\infty]{\text{loi}} X^{(j)}$$

qui provient du 1^er^ théorème de continuité appliqué à la fonction continue $x \longmapsto x^{(j)}$. La réciproque n'est pas vraie : en général, on ne peut pas déduire de la convergence en loi coordonnée par coordonnée une convergence en loi du vecteur aléatoire.[3]

[3] Considérer par exemple $X_n = X_n' = Y_n = -Y_n' = X$ où $X \sim \mathcal{N}(0,1)$. Ces quatre suites convergent en loi vers $\mathcal{N}(0,1)$, mais si $((X_n, Y_n))_{n\geq1}$ et $((X_n', Y_n'))_{n\geq1}$ convergeaient en loi vers le même vecteur aléatoire, alors $(X_n + Y_n)_{n\geq1}$ et $(X_n' + Y_n')_{n\geq1}$ auraient la même limite en loi. Or on voit bien que ce n'est pas le cas, puisqu'elles convergent en loi respectivement vers $2X$ et 0.

Néanmoins il existe des cas particulier où on peut quand même le faire. Le plus important est le suivant :

Théorème I.8 Slutsky

Soient $(X_n)_{n\geq 1}$ (*resp.* $(Y_n)_{n\geq 1}$) une suite de variables aléatoires dans $\mathbb{R}^d$ (*resp.* $\mathbb{R}^{d'}$), et X (*resp.* Y) une variable aléatoire dans $\mathbb{R}^d$ (*resp.* $\mathbb{R}^{d'}$).
Si $X_n \xrightarrow[n\to+\infty]{\text{loi}} X$ et $Y_n \xrightarrow[n\to+\infty]{\text{loi}} c \in \mathbb{R}^{d'}$ fixé, alors $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow[n\to+\infty]{\text{loi}} \begin{pmatrix} X \\ c \end{pmatrix}$.

Théorème I.9 2ème théorème de continuité (« méthode δ »)

On suppose qu'il existe $a \in \mathbb{R}^d$ tel que[4]

$$\sqrt{n}(X_n - a) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \Sigma)$$

Soit $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ une fonction **continûment dérivable** (i.e. C^1), alors

$$\sqrt{n}(g(X_n) - g(a)) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \nabla g(a)^\top \Sigma \nabla g(a))$$

[4] On dit que la distribution de X_n est *asymptotiquement normale*, et Σ est appelée *variance asymptotique* de $(X_n)_{n\geq 1}$.

Remarque. Si les X_i sont i.i.d. et $\mathbb{E}[\|X_1\|^2] < +\infty$, alors par le TCL :

$$\sqrt{n}(\bar{X}_n - \mathbb{E}[X_1]) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \mathrm{Var}(X_1))$$

On peut donc appliquer la méthode δ à la suite $(\bar{X}_n)_{n\geq 1}$ (prendre $a = \mathbb{E}[X_1]$ et $\Sigma = \mathrm{Var}(X_1)$). On a donc la normalité asymptotique de $g(\bar{X}_n)$, et l'expression de sa variance asymptotique.

Preuve. Cas particulier. On fait la preuve en supposant que $d = 1$. (Du coup on note σ^2 la variance asymptotique de X_n plutôt que Σ.)
On pose

$$h(x) = \begin{cases} \frac{g(x)-g(a)}{x-a} & \text{si } x \neq a \\ g'(a) & \text{si } x = a \end{cases}$$

h est continue étant donnée l'hypothèse sur g.
De plus,

$$X_n - a = \frac{1}{\sqrt{n}}\sqrt{n}(X_n - a) \xrightarrow[n\to+\infty]{\text{loi}} 0$$

donc

$$X_n - a \xrightarrow[n\to+\infty]{P} 0$$

donc

$$X_n \xrightarrow[n\to+\infty]{P} a$$

Par le premier théorème de continuité,

$$h(X_n) \xrightarrow[n\to+\infty]{P} h(a)$$

D'autre part,

$$\sqrt{n}(g(X_n) - g(a)) = \sqrt{n}(X_n - a) \times h(X_n)$$

Par Slutsky, si on note $Z \sim \mathcal{N}(0, \sigma^2)$, alors

$$\sqrt{n}(g(X_n) - g(a)) \xrightarrow[n\to+\infty]{\text{loi}} Z \times h(a)$$

Il ne reste plus qu'à remarquer que $Z \times h(a) \sim \mathcal{N}(0, \sigma^2(h(a))^2)$ et que $h(a) = g'(a)$. ■

II Échantillonnage

II.1 Introduction : hypothèse d'échantillonnage

Hypothèse E. On observe une réalisation de n variables aléatoires **indépendantes** et de **même loi** P^*.

- $X_1, \dots, X_n$ s'appelle l'échantillon.
- On suppose l'existence d'un espace probabilisé $(\Omega, \mathcal{F}, P)$ tel que pour tout i, $X_i : \Omega \longrightarrow \mathbb{R}^m$.
- On suppose que les vraies données observées $x_1, \dots, x_n$ peuvent s'écrire $X_1(\omega), \dots, X_n(\omega)$ pour un certain $\omega \in \Omega$.[5]

[5] En gros : les données qu'on observe sur un tableur Excel par exemple sont des nombres $x_1, \dots, x_n$, qui correspondent à la réalisation de l'échantillon pour un ω particulier. Il y a autant de mondes parallèles que d'éléments de Ω et le monde dans lequel nous vivons correspond à ce ω.

Définition II.1 Loi d'une v.a.

Soit X est une variable aléatoire dans $\mathbb{R}^m$. On appelle loi de X la mesure de probabilité sur $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$, définie par :

$$\begin{array}{rcl} \mathcal{B}(\mathbb{R}^m) & \longrightarrow & \mathbb{R} \\ A & \longmapsto & P(X \in A) \end{array}$$

que l'on peut assimiler à l'ensemble :

$$\{P(X \in A) : A \in \mathcal{B}(\mathbb{R}^m)\}$$

Problème stat. En utilisant l'échantillon $X_1, \dots, X_n$, retrouver certaines propriétés de P^*. Ce problème se divise en plusieurs parties :

1. *Estimation* : soit $T : \mathcal{P} \longrightarrow \mathbb{R}^p$ une fonction quelconque, où $\mathcal{P}$ désigne l'ensemble des lois sur $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$. On cherche à approximer ou à estimer $T(P^*)$.
2. *Ensembles de confiance* : on se donne un $\alpha \in]0, 1[$ et on cherche un ensemble $E = E(X_1, \dots, X_n) \subset \mathbb{R}^p$ tel que $P(E \ni T(P^*)) \geq 1 - \alpha$.
3. *Test d'hypothèses* : on se donne un sous-ensemble $\mathcal{P}_0$ de l'ensemble $\mathbb{P}$ et on souhaite savoir si $P^* \in \mathcal{P}_0$ ou pas.

II.2 Fonction de répartition empirique

Contexte. On considère $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P^*$ à valeurs dans $\mathbb{R}^m$.

Il faut distinguer la fonction de répartition théorique[6] $F^*(t) = P(X_1 \leq t)$, $t \in \mathbb{R}^m$, de la fonction de répartition empirique :

[6] On note

$$\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \leq \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

lorsque $u_i \leq v_i$ pour tout $i \in [\![1, n]\!]$.

Définition II.2 Fonction de répartition empirique

On appelle fonction de répartition empirique de l'échantillon $X_1, \dots, X_n$ la fonction **aléatoire** définie par :

$$\begin{aligned} \hat{F}_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq t) \\ &= \frac{1}{n} \operatorname{Card}\{i \in [\![1, n]\!] : X_i \leq t\} \end{aligned}$$

Remarque. À t fixé, $\hat{F}_n(t)$ est une **variable aléatoire** ! $\hat{F}_n$ est donc à considérer comme une fonction à deux variables :

$$\begin{array}{rcl} \mathbb{R}^m \times \Omega & \longrightarrow & \mathbb{R} \\ (t, \omega) & \longmapsto & \hat{F}_n(t, \omega) \end{array}$$

Propriété II.3

Soient $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P^*$ à valeurs dans $\mathbb{R}^m$. Alors pour tout $t \in \mathbb{R}^m$:

1. $n\hat{F}_n(t) \sim \mathcal{B}(n, F^*(t))$;
2. $\hat{F}_n(t) \xrightarrow[n\to+\infty]{\text{p.s.}} F^*(t)$;
3. $\sqrt{n}(\hat{F}_n(t) - F^*(t)) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, p(1-p))$ avec $p = F^*(t)$.

Preuve.

1. Posons $Z_i = \mathbb{1}(X_i \leq t)$ pour tout i.
   Les Z_i sont indépendantes (car les X_i le sont) et suivent toutes la loi de Bernoulli de paramètre $p = F^*(t)$.
   On a $n\hat{F}_n(t) = Z_1 + \cdots + Z_n \sim \mathcal{B}(n, p)$.
2. On a $\mathbb{E}[|Z_1|] \leq 1$ donc la LFGN s'applique :
   $$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n Z_i \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbb{E}[Z_1] = F^*(t)$$
3. On a
   $$\sqrt{n}(\hat{F}_n(t) - F^*(t)) = \sqrt{n}(\bar{Z}_n - \mathbb{E}[Z_1])$$
   Or $\mathbb{E}[Z_1^2] \leq 1$ donc d'après le TCL,
   $$\sqrt{n}(\bar{Z}_n - \mathbb{E}[Z_1]) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \mathrm{Var}(Z_1))$$
   Il ne reste plus qu'à remarquer que $\mathrm{Var}(Z_1) = p(1-p)$.

■

Théorème II.4 Glivenko-Cantelli

Soient $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P^*$ à valeurs dans $\mathbb{R}^m$. Alors :

$$\sup_{t\in\mathbb{R}^m} |\hat{F}_n(t) - F^*(t)| \xrightarrow[n\to+\infty]{\text{p.s.}} 0$$

Remarque. Posons $\xi_n = \sup_{t\in\mathbb{R}^m} |\hat{F}_n(t) - F^*(t)|$.

— C'est une variable aléatoire et on a $\xi_n = \|\hat{F}_n - F^*\|_\infty$.

— Pour presque tout $\omega \in \Omega$, $(\hat{F}_n(\cdot, \omega))_{n\geq 1}$ converge uniformément vers $F^*(\cdot)$. Le théorème de Glivenko-Cantelli est donc une loi forte des grands nombres **uniforme**.

Preuve. Cas particulier. On se restreint au cas $m = 1$ et F^* est continue. Soit $k \in \mathbb{N}^*$. On pose $x_0 = -\infty$, $x_k = +\infty$ et pour tout $i \in [\![1, k-1]\!]$:

$$x_i \in (F^*)^{-1}\left(\frac{i}{k}\right) \quad \text{i.e.} \quad F^*(x_i) = \frac{i}{k}$$

Par croissance de F^*, on a

$$x_0 < x_1 < \cdots < x_{k-1} < x_k$$

Soit $t \in [x_i, x_{i+1}]$. En utilisant la croissance de F^* et celle de $\hat{F}_n(\cdot, \omega)$, on a :

$$\begin{aligned}\hat{F}_n(t) - F^*(t) \leq \hat{F}_n(x_{i+1}) - F^*(x_i) &= \hat{F}_n(x_{i+1}) - F^*(x_{i+1}) + \underbrace{F^*(x_{i+1}) - F^*(x_i)}_{\leq 1/k} \\ &\leq F_n(x_{i+1}) - F^*(x_{i+1}) + \frac{1}{k} \\ &\leq |F_n(x_{i+1}) - F^*(x_{i+1})| + \frac{1}{k}\end{aligned}$$

On prouve de même que

$$F^*(t) - \hat{F}_n(t) \leq F^*(x_{i+1}) - \hat{F}_n(x_i) \leq |F_n(x_i) - F^*(x_i)| + \frac{1}{k}$$

donc

$$|\hat{F}_n(t) - F^*(t)| \leq \max_{1\leq i\leq k-1} |\hat{F}_n(x_i) - F^*(x_i)| + \frac{1}{k}$$

d'où

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F^*(t)| \leq \underbrace{\max_{1\leq i\leq k-1} |\hat{F}_n(x_i) - F^*(x_i)|}_{\xrightarrow[n\to+\infty]{\text{p.s.}} 0} + \frac{1}{k} \tag{1}$$

Considérons, pour tout $k \in \mathbb{N}^*$,

$$\Omega_k = \left\{\omega \in \Omega : \max_{1\leq i\leq k-1} |\hat{F}_n(x_i, \omega) - F^*(x_i)| \xrightarrow[n\to+\infty]{} 0\right\}$$

On a $P(\Omega_k) = 1$ pour tout $k \in \mathbb{N}^*$, donc l'événement $A = \bigcap_{k\in\mathbb{N}^*} \Omega_k$ est de probabilité 1.
Pour tout $\omega \in A$ on a, d'après (1) :

$$\forall k \in \mathbb{N}^*, \quad \xi_n(\omega) \leq \max_{1\leq i\leq k-1} |\hat{F}_n(x_i, \omega) - F^*(x_i)| + \frac{1}{k}$$

donc

$$\forall k \in \mathbb{N}^*, \quad 0 \leq \limsup_{n\to+\infty} \xi_n(\omega) \leq 0 + \frac{1}{k}$$

donc

$$\limsup_{n\to+\infty} \underbrace{\xi_n(\omega)}_{\geq 0} = 0$$

d'où $\lim_{n\to+\infty} \xi_n(\omega) = 0$.
Donc $\xi_n \xrightarrow[n\to+\infty]{\text{p.s.}} 0$. ∎

II.3 Théorèmes de Kolmogorov et de Donsker

Note : ce sont des sujets un peu plus avancés, pas forcément utiles pour ce semestre mais le seront en Stat 2.

Par Glivenko-Cantelli, $(\xi_n)_{n\geq 1}$ converge presque sûrement vers 0. On peut se demander à quelle vitesse a lieu cette convergence. Le théorème de Kolmogorov donne une réponse surprenante : en multipliant par $\sqrt{n}$, on obtient une suite qui converge en loi et sa limite en loi peut être exprimée explicitement, et en plus, indépendamment de F^*. Plus fort que ça : en fait même à n fixé la loi de ξ_n ne dépend pas de F^*.

Théorème II.5 Kolmogorov

Si $X_1, \ldots, X_n$ sont i.i.d. de f.d.r. F^*, alors

1. la loi de ξ_n ne dépend pas de F^* ;

2. $\sqrt{n}\xi_n \xrightarrow[n\to+\infty]{\text{loi}} K$, où K est une variable aléatoire dont la fonction de répartition est :

$$\begin{array}{rccl} F_K: & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & \begin{cases} 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2t^2} & \text{si } t > 0 \\ 0 & \text{si } t \leq 0 \end{cases} \end{array}$$

Preuve. Prouvons le premier point avec l'hypothèse supplémentaire F^* est une bijection de $\mathbb{R}$ dans $]0,1[$. On a

$$\xi_n = \sup_{x\in]0,1[} |\hat{F}_n(t)((F^*)^{-1}(x)) - F^*((F^*)^{-1}(x))|$$

Or

$$\hat{F}_n(t)((F^*)^{-1}(x)) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i \leq (F^*)^{-1}(x)) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(F^*(X_i) \leq x)$$

On prouve classiquement que $F^*(X_i) \sim \mathcal{U}([0,1])$. On a ainsi

$$\xi_n = \sup_{x\in]0,1[}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(U_i \leq x) - x\right| \qquad \text{où} \qquad U_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$$

ξ_n est une stat. libre, i.e. sa loi ne dépend pas de F^*. ■

Remarques.

1. Si l'on veut tester l'hypothèse $F^* = \Phi$, la f.d.r. de $\mathcal{N}(0,1)$, on peut alors définir $q_n(\alpha)$ comme le quantile d'ordre $1-\alpha$ de ξ_n. Comme on l'a vu dans la preuve précédente, ça revient à prendre le $(1-\alpha)$-quantile de la loi de :

$$\sup_{x\in]0,1[}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(U_i \leq x) - x\right|, \quad U_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$$

   Ensuite, on rejette l'hypothèse $F^* = \Phi$ si et seulement si :

$$\sup_{t\in\mathbb{R}}|\hat{F}_n(t) - \Phi(t)| \geq q_n(\alpha) \tag{2}$$

   Ce test, appelé test de Kolmogorov(-Smirnov), est de niveau α ; autrement dit, si $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, alors l'événement (2) est de probabilité α.
2. Lorsque n est grand, on remplace $q_n(\alpha)$ dans (2) par le quantile de la variable aléatoire K.

Trigger warning : giga tunnel ci-dessous.

Théorème II.6 Donsker

Considérons :

$$\begin{array}{rccl} G_n: & \mathbb{R}\times\Omega & \longrightarrow & \mathbb{R} \\ & (t,\omega) & \longmapsto & \sqrt{n}(\hat{F}_n(t,\omega) - F^*(t)) \end{array}$$

(À $t \in \mathbb{R}$ fixé, $G_n(t,\cdot)$ est une variable aléatoire, tandis qu'à $\omega \in \Omega$ fixé, $G_n(\cdot,\omega)$ est une fonction. De manière équivalente, on peut voir G_n à la fois comme une application de $\mathbb{R}$ dans $\mathbb{R}^\Omega$ ou comme une application de Ω dans $\mathbb{R}^\mathbb{R}$. G_n est ce qu'on appelle un processus aléatoire.)

Alors
$$G_n \xrightarrow[n\to+\infty]{\text{loi}} G$$
où G est un « processus gaussien » de moyenne 0 et d'opérateur de covariance
$$\operatorname{cov}(G(t), G(s)) = \min(F^*(t), F^*(s)) - F^*(t)F^*(s)$$

Conséquence. $\|G_n\|_\infty \xrightarrow[n\to+\infty]{\text{loi}} \|G\|_\infty$. Donc la variable aléatoire K définie précédemment est égale en loi à $\|G\|_\infty = \sup_{t\in\mathbb{R}} |G(t)|$.

II.4 Méthode de substitution

Comme précédemment, on considère les données $X_1, \ldots, X_n$ de loi inconnue P^*. On cherche à estimer $\theta^* = T(P^*)$,[7] avec
$$T : \mathcal{P} \longrightarrow \mathbb{R}$$
où l'on a noté $\mathcal{P}$ l'ensemble des mesures de probabilité sur $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^*))$. On sait qu'il existe une bijection
$$\Psi : \mathcal{F} \longrightarrow \mathcal{P}$$
où $\mathcal{F}$ est l'ensemble des fonctions de répartitions des v.a. à valeurs dans $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^*))$. On rappelle que la fonction de répartition empirique $\hat{F}_n$ peut être vue :

1. comme une fonction à deux arguments : $\hat{F}_n : \mathbb{R}^m \times \Omega \longrightarrow \mathbb{R}$, $(t, \omega) \longmapsto \hat{F}_n(t, \omega)$;
2. comme une fonction qui prend en argument un vecteur de $\mathbb{R}^m$ et renvoie une variable aléatoire réelle : $\hat{F}_n : \mathbb{R}^m \longrightarrow \mathbb{R}^\Omega$, $t \longmapsto \hat{F}_n(t, \cdot)$;
3. comme une variable aléatoire à valeurs dans l'espace des fonctions de répartition : $\hat{F}_n : \Omega \longrightarrow \mathcal{F}$, $\omega \longmapsto \hat{F}_n(\cdot, \omega)$.

On adopte plutôt le troisième point de vue dans la définition qui suit.

Définition II.7 Méthode de substitution

La méthode de substitution consiste à estimer $\theta^* = T(P^*)$ par
$$\hat{\theta}_n = T(\Psi(\hat{F}_n))$$
La loi $\hat{P}_n := \Psi(\hat{F}_n)$ est appelée mesure empirique associée à $X_1, \ldots, X_n$.

Exemple. Soient P une loi sur $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m)$ et F la fonction de répartition associée. Supposons que $T(P) = P([1, +\infty[) = \lim_{\varepsilon\to 0}(1 - F(1-\varepsilon))$: $T(P)$ correspond à la probabilité qu'une variable aléatoire de loi P appartienne à $[1, +\infty[$. Alors la méthode de substitution consiste à estimer $T(P^*)$ par la variable aléatoire :
$$\hat{\theta}_n = \lim_{\varepsilon\to 0}(1 - \hat{F}_n(1 - \varepsilon))$$

Remarque. On peut prouver que
$$\hat{P}_n(A) = \frac{\operatorname{Card}\{i \in [\![1, n]\!] : X_i \in A\}}{n} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\in A} \qquad \text{i.e.} \qquad \hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$
Donc pour toute fonction $h : \mathbb{R} \longrightarrow \mathbb{R}$ mesurable,
$$\int_{\mathbb{R}} h(x)\hat{P}_n(dx) = \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}} h d\delta_{X_i} = \frac{1}{n}\sum_{i=1}^n h(X_i)$$

[7] Par exemple, si $m = 1$ et si on veut estimer l'espérance de la loi des données, T pourrait être l'application qui à une loi P associe le réel $\int_{\mathbb{R}^m} x dP(x)$.

Propriété II.8

Si l'application $T \circ \Psi$ est continue, alors $\hat{\theta}_n \xrightarrow[n\to+\infty]{\text{p.s.}} \theta^*$.

Preuve. On sait que $\theta^* = T \circ \Psi(F^*)$ et $\hat{\theta}_n = T \circ \Psi(\hat{F}_n)$ d'après le théorème de Glivenko-Cantelli. Comme $T \circ \Psi$ est continue, $T \circ \Psi(\hat{F}_n) \xrightarrow[n\to+\infty]{\text{p.s.}} T \circ \Psi(F^*)$. ∎

II.5 Moyenne et variance empirique

Supposons que $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P^*$ sont tels que $\mathbb{E}[\|X_1\|^2] < +\infty$. Posons $\mu^* = \mathbb{E}[X_1] \in \mathbb{R}^m$ et $\Sigma^* = \mathbb{E}[X_1 X_1^\top] - \mu^*(\mu^*)^\top$, qui est une matrice positive de taille $m \times m$. On note aussi $\widehat{\Sigma}_n$ la matrice de covariance empirique :

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \bar{X}_n \bar{X}_n^\top$$

que l'on note plus souvent S_n^2 lorsque $m = 1$. Enfin, on pose :

$$\mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$$

Propriété II.9

Soient $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ et $\sigma > 0$.

1. $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
2. $\frac{nS_n^2}{\sigma^2} \sim \chi^2_{n-1}$.
3. $\bar{X}_n \perp\!\!\!\perp S_n^2$.

Preuve.

1. Soit $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$. Les X_i sont i.i.d. gaussiennes donc $\vec{X}$ est un vecteur gaussien. Plus précisément,

$$\vec{X} \sim \mathcal{N}_n\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \sigma^2 I_n\right)$$

Toute transformation affine de $\vec{X}$ est gaussienne. Donc $\bar{X}_n = \frac{1}{n}\mathbb{1}_n^\top \vec{X}$ suit une loi gaussienne, avec

$$\mathbb{E}[\bar{X}_n] = \mu \qquad \text{et} \qquad \operatorname{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

2. Posons $Z = \frac{1}{\sigma}(\vec{X} - \mu\mathbb{1}_n)$. Alors $Z \sim \mathcal{N}_n(0, I_n)$.

De plus,

$$\begin{aligned}\frac{nS_n^2}{\sigma^2} &= \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n}\left(\frac{X_i - \mu - (\bar{X}_n - \mu)}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2\end{aligned}$$

Astuce : ici on remarque que $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top Z = \begin{pmatrix}\bar{Z}_n \\ \vdots \\ \bar{Z}_n\end{pmatrix}$ d'où :

$$\begin{aligned}\frac{nS_n^2}{\sigma^2} &= \left\|Z - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top Z\right\|^2 \\ &= \left\|\left(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top\right) Z\right\|^2\end{aligned}$$

Or $\Pi_1 := I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$ vérifie $\Pi_1^\top = \Pi_1 = \Pi_1^2$: c'est une matrice de projection orthogonale, de rang/trace $n-1$.

Donc $\frac{nS_n^2}{\sigma^2} = \|\Pi_1 Z\|^2 \sim \chi^2_{n-1}$.

3. On a

$$\begin{pmatrix}\Pi_1 Z \\ (I-\Pi_1)Z\end{pmatrix} = \underbrace{\begin{pmatrix}\Pi_1 \\ I-\Pi_1\end{pmatrix}}_{\text{déterministe}} Z$$

donc $\begin{pmatrix}\Pi_1 Z \\ (I-\Pi_1)Z\end{pmatrix}$ est un **vecteur gaussien**. Donc pour avoir l'indépendance de $\Pi_1 Z$ et $(I-\Pi_1)Z$, il suffit que leur covariance soit nulle.

On calcule :

$$\begin{aligned}\operatorname{cov}(\Pi_1 Z, (I-\Pi_1)Z) &= \mathbb{E}[(\Pi_1 Z)((I-\Pi_1)Z)^\top] \\ &= \Pi_1\mathbb{E}[ZZ^\top](I-\Pi_1)\end{aligned}$$

Or $Z \sim \mathcal{N}_n(0, I_n)$ donc $\mathbb{E}[ZZ^\top] = \operatorname{Var}(Z) = I_n$. Puisque $\Pi_1^2 = \Pi_1$, on obtient :

$$\operatorname{cov}(\Pi_1 Z, (I-\Pi_1)Z) = \Pi_1(I-\Pi_1) = 0$$

D'où :

$$\Pi_1 Z \perp\!\!\!\perp (I-\Pi_1)Z$$

or $S_n^2 = \frac{\sigma^2}{n}\|\Pi_1 Z\|^2$ comme on l'a vu au point précédent, donc :

$$S_n^2 \perp\!\!\!\perp (I-\Pi_1)Z = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top Z = \begin{pmatrix}\bar{Z}_n \\ \vdots \\ \bar{Z}_n\end{pmatrix}$$

d'où

$$S_n^2 \perp\!\!\!\perp \frac{\bar{X}_n - \mu}{\sigma}$$

et finalement :

$$S_n^2 \perp\!\!\!\perp \bar{X}_n$$

■

III Estimation des paramètres

III.1 Modèle statistique

Soient $X_1, \dots, X_n$ des variables aléatoires i.i.d. de loi inconnue P^* sur $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$.

Définition III.1 Modèle

On appelle modèle statistique un triplet

$$(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$$

où :

- ▹ $\mathcal{X}$ est un espace d'états ;
- ▹ $\mathcal{A}$ est une tribu sur $\mathcal{X}$;
- ▹ chaque P_θ est une probabilité sur $(\mathcal{X}, \mathcal{A})$.

Hypothèse P : la loi inconnue P^* appartient à $\{P_\theta : \theta \in \Theta\}$.

Il faut voir $(\mathcal{X}, \mathcal{A})$ comme l'espace mesurable dans lequel les X_i *prennent leurs valeurs*, et chaque P_θ est une potentielle loi pour les X_i, et on cherche laquelle est la bonne. Dans ce cours, $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$. On va donc souvent identifier le modèle statistique à la famille $\{P_\theta : \theta \in \Theta\}$.

Exemple. On considère les données $x_1, \dots, x_n$ où x_i représente la durée de vue d'une composante électrique. Dans ce cas, on peut considérer le modèle $\{P_\theta : \theta \in]0, +\infty[\}$ où $P_\theta = \mathcal{E}(\theta)$ donnée par la densité

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{[0,+\infty[}(x)$$

Définition III.2 Modèle paramétrique

Le modèle $\{P_\theta : \theta \in \Theta\}$ est dit paramétrique lorsqu'il existe $k \in \mathbb{N}$ tel que $\Theta \subset \mathbb{R}^k$.[8]

[8] En clair : un modèle est paramétrique s'il est décrit par un nombre fini de paramètres réels.

Exemples.

- ▹ Le modèle $\{\mathcal{E}(\theta) : \theta \in \mathbb{R}_+^*\}$ est paramétrique car $\mathbb{R}_+^* \subset \mathbb{R}$.
- ▹ Le modèle $\{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$ est paramétrique car $\mathbb{R} \times \mathbb{R}_+ \subset \mathbb{R}^2$.
- ▹ Le modèle $\{P_f : f \in \Theta\}$ où Θ est l'ensemble des densités sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ qui sont continues sur $[0, 1]$ et P_f est la loi associée à la densité f, n'est pas paramétrique.

Dans un modèle paramétrique, sous l'hypothèse P, il existe $\theta^* \in \Theta$ pour lequel $P^* = P_{\theta^*}$. On dit que θ^* est $\boxed{\textbf{une}}$ vraie valeur de paramètre, car parfois il peut arriver qu'il y ait plusieurs valeurs du paramètre telle que la loi associée dans le modèle coïncide avec P^*. Néanmoins cela n'arrive pas lorsque les P_θ sont deux-à-deux distinctes : on *identifie* chaque paramètre θ à une unique loi P_θ du modèle, c'est ce qui amène à la définition suivante :

Définition III.3 Modèle identifiable

Le modèle $\{P_\theta : \theta \in \Theta\}$ est dit identifiable lorsque l'application $\theta \longmapsto P_\theta$ est injective.

Exemples.

1. Supposons que $P_\theta = \mathcal{E}(\theta)$, $\Theta =]0, +\infty[$. Alors $\{P_\theta : \theta \in \Theta\}$ est identifiable : en effet,
$$P_\theta = P_{\theta'} \implies \mathbb{E}_\theta[X_1] = \mathbb{E}_{\theta'}[X_1] \implies \theta = \theta'$$
où l'on note $\mathbb{E}_\theta[h(X_1)]$ l'espérance de $h(X_1)$ où X_1 est n'importe quelle v.a. de loi P_θ.[9]
2. Supposons que $P_\theta = \mathcal{N}(0, \theta^2)$, $\theta \in \Theta = \mathbb{R}$. Alors $\{P_\theta : \theta \in \Theta\}$ n'est pas identifiable. Par exemple, $P_1 = P_{-1}$ pourtant $1 \neq -1$.

[9] La notation complète serait $\mathbb{E}_{X_1 \sim P_\theta}[h(X_1)]$ mais on lui préfère ici sa version abrégée.

Définition III.4 Rappels de théorie de la mesure

Soit $(E, \mathcal{E})$ un espace mesurable.

— Une mesure μ sur $(E, \mathcal{E})$ est dite σ-finie lorsqu'il existe une famille $(A_n)_{n\in\mathbb{N}}$ d'éléments de $\mathcal{E}$ qui vérifient $\mu(A_n) < +\infty$ pour tout $n \in \mathbb{N}$, tels que $\bigcup_{n\in\mathbb{N}} A_n = E$.

— Soient μ et ν deux mesures σ-finies sur $(E, \mathcal{E})$. On dit que ν est absolument continue par rapport à μ, et on note $\nu \ll \mu$, lorsque tout ensemble mesurable μ-négligeable est aussi ν-négligeable, i.e.,
$$\forall A \in \mathcal{E}, \quad \mu(A) = 0 \implies \nu(A) = 0$$

Théorème III.5 Radon-Nikodym

$\nu \ll \mu$ si et seulement si ν admet une densité par rapport à μ, i.e. il existe une fonction f mesurable positive telle que
$$\forall A \in \mathcal{E}, \quad \nu(A) = \int_A f d\mu$$

En particulier, les lois qui admettent une densité par rapport à la mesure de Lebesgue sont les lois absolument continues par rapport à la mesure de Lebesgue.

Définition III.6 Modèle dominé

Le modèle $\{P_\theta : \theta \in \Theta\}$ est dit dominé lorsqu'il existe une mesure σ-finie μ telle que pour tout $\theta \in \Theta$, P_θ est absolument continue par rapport à μ.

Définition III.7 Modèle à densité, modèle discret

Le modèle $\{P_\theta : \theta \in \Theta\}$ est dit à densité (ou continu) lorsqu'il est dominé par la mesure de Lebesgue.
Le modèle $\{P_\theta : \theta \in \Theta\}$ est dit discret lorsque
$$\exists A \in \mathcal{B}(\mathbb{R}^m) \text{ dénombrable}, \qquad \forall \theta \in \Theta, \qquad P_\theta(A) = 1$$

Attention ! L'ordre des quantificateurs est important. On peut avoir un modèle $\{P_\theta : \theta \in \Theta\}$ qui n'est pas discret et où pourtant chaque loi P_θ est discrète. Par exemple, prendre $\Theta = [0, 1]$ et $P_\theta = \delta_\theta$, la masse de Dirac en θ. Il est clair que la loi P_θ est discrète[10] mais le modèle $\{P_\theta : \theta \in \Theta\}$ n'est pas discret ! En effet, pour tout $A \in \mathcal{B}([0, 1])$ dénombrable, $[0, 1] \setminus A \neq \emptyset$. Considérons $\theta \in [0, 1] \setminus A$, alors on a $P_\theta(A) = 0 \neq 1$.

[10] Soit voir que c'est la loi d'une variable aléatoire constante, soit voir que $P_\theta(\{\theta\}) = 1$.

Remarque. Si le modèle $\{P_\theta : \theta \in \Theta\}$ est discret, alors il est dominé par la mesure de comptage d'un ensemble dénombrable.

En effet, s'il est discret, alors il existe A dénombrable tel que pour tout $\theta \in \Theta$, $P_\theta(A) = 1$.
Posons $\mu = \sum_{a \in A} \delta_a$, la mesure de comptage sur A. Autrement dit, pour tout B, $\mu(B) = \text{Card}(A \cap B)$. Alors μ est σ-finie et $P_\theta = \int p_\theta d\mu$ où $p_\theta(a) = P_\theta(\{a\})$ pour tout $a \in A$.
Si $\mu(B) = 0$, alors $B \cap A = \emptyset$ donc $P_\theta(B) = P_\theta(B \cap A) + P_\theta(B \cap A^c) \leq 0 + P(A^c) = 0$ donc $P_\theta \ll \mu$.

On admet que la réciproque est vraie, autrement dit :

Propriété III.8

$\{P_\theta : \theta \in \Theta\}$ est discret si et seulement s'il est dominé par la mesure de comptage d'un ensemble dénombrable.

En bref :

Nom	Signification
Modèle	$(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$
Paramétrique	$\Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$
Identifiable	$P_\theta = P_{\theta'} \implies \theta = \theta'$
Dominé	$P_\theta \ll \mu$, $\forall \theta \in \Theta$
À densité	Dominé par $\mu = \text{Leb}$
Discret	Dominé par $\mu =$ mes. comptage d'un ens. dén.

III.2 Estimateur

On considère $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, $\theta^* \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$.

Définition III.9 Estimateur

Toute statistique, i.e. variable aléatoire de la forme $\bar{\theta}_n = g(X_1, \ldots, X_n)$ avec $g : (\mathbb{R}^m)^n \longrightarrow \mathbb{R}^k$ mesurable, est appelée estimateur de θ lorsque g ne dépend pas de θ.
Un estimateur $\bar{\theta}_n$ de θ est dit :[11]

— sans biais, si
$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta[\bar{\theta}_n] = \theta$$

— (fortement) consistant, si
$$\forall \theta \in \Theta, \quad \bar{\theta}_n \xrightarrow[n \to +\infty]{P_\theta\text{-p.s.}} \theta$$

— faiblement consistant, si
$$\forall \theta \in \Theta, \quad \bar{\theta}_n \xrightarrow[n \to +\infty]{P_\theta} \theta$$

— asymptotiquement normal, s'il existe $v : \Theta \longrightarrow [0, +\infty[$ telle que
$$\forall \theta \in \Theta, \quad \sqrt{n}(\bar{\theta}_n - \theta) \xrightarrow[n \to +\infty]{\text{loi}} \mathcal{N}(0, v(\theta))$$

[11] Remarquer que dans les quatre définitions on dit : pour tout θ, si les données sont de loi P_θ, alors on a telle égalité / convergence. C'est logique : on ne sait pas quel est le vrai paramètre θ^*, donc on veut que pour tous les potentiels paramètres θ, *sous* P_θ, l'estimateur ait telle propriété qui permet de se faire une idée de θ via des simulations (en R ou Python par exemple, en approximant son espérance, ou en simulant l'estimateur pour un grand n...). Cela garantit que quand on fait ces simulations avec nos vraies données, on va bien pouvoir se faire une idée de qui est θ^*.

Exemple. Soient $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{E}(\theta^*)$, $\theta^* > 0$.

1. Soit $\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$ estimateur de θ^*. Un calcul rapide montre que $\bar{\theta}_n$ est sans biais. De plus, par la LFGN, $\bar{\theta}_n$ est consistant. Enfin, en prenant

$v(\theta) = \theta^2$ on constate que $\bar{\theta}_n$ est asymptotiquement normal.

2. Soit $\hat{\theta}_n = S_n = \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2}$ (écart-type empirique). On peut vérifier que
$$\hat{\theta}_n^2 \xrightarrow[n\to+\infty]{P} \mathrm{Var}_{\theta^*}(X_1) = (\theta^*)^2$$
donc $\hat{\theta}_n$ est faiblement consistant.[12]

[12] En vrai il l'est aussi fortement.

3. $\hat{\theta}'_n = X_1$ est sans biais mais non consistant.
4. $\hat{\theta}''_n = \frac{1}{n+1}\sum_{i=1}^n X_i$ est biaisé, mais $\hat{\theta}''_n = \frac{n}{n+1}\bar{\theta}_n$ avec $\frac{n}{n+1} \xrightarrow[n\to+\infty]{} 1$ donc $\hat{\theta}''_n$ est fortement consistant (par Slutsky, ou bien en revenant à la définition de la convergence presque sûre).

III.3 Risque d'estimation

On considère $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, $\theta^* \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$.

Définition III.10 Risque

Soit $\bar{\theta}_n$ un estimateur de θ^*. On appelle risque de $\bar{\theta}_n$ la fonction

$$\begin{array}{rccl} R_{\bar{\theta}_n} : & \Theta & \longrightarrow & \mathbb{R}_+ \\ & \theta & \longmapsto & \mathbb{E}_\theta[\|\bar{\theta}_n - \theta\|^2] \end{array}$$

où $\|\cdot\|$ est la norme euclidienne.

Lemme III.11 Décomposition biais-variance

On a
$$R_{\bar{\theta}_n}(\theta) = \|\mathbb{E}_\theta[\bar{\theta}_n] - \theta\|^2 + \mathbb{E}_\theta[\|\bar{\theta}_n - \mathbb{E}_\theta[\bar{\theta}_n]\|^2]$$

Preuve. On pose $m = \mathbb{E}_\theta[\bar{\theta}_n]$. On a

$$\begin{aligned} \|\mathbb{E}_\theta[\bar{\theta}_n] - \theta\|^2 + \mathbb{E}_\theta[\|\bar{\theta}_n - \mathbb{E}_\theta[\bar{\theta}_n]\|^2] &= \|m\|^2 - 2\theta^\top m + \|\theta\|^2 \\ &\quad + \mathbb{E}_\theta[\|\bar{\theta}_n\|^2] - 2\underbrace{\mathbb{E}_\theta[\bar{\theta}_n^\top m]}_{=\|m\|^2} + \|m\|^2 \\ &= \mathbb{E}_\theta\left[\|\bar{\theta}_n\|^2 - 2\theta^\top\bar{\theta}_n + \|\theta\|^2\right] \\ &= R_{\bar{\theta}_n}(\theta) \end{aligned}$$

■

Remarques.

— Par définition, $R_{\bar{\theta}_n}(\theta) \xrightarrow[n\to+\infty]{} 0$ équivaut à $\bar{\theta}_n \xrightarrow[n\to+\infty]{L^2} \theta$.

Donc si $R_{\bar{\theta}_n}(\theta) \xrightarrow[n\to+\infty]{} 0$ (pour tout $\theta \in \Theta$), alors $\bar{\theta}_n$ est faiblement consistant.

— Lorsque $k = 1$, cette relation s'écrit :
$$\text{risque} = \text{biais}^2 + \text{variance}$$
Pour que $R_{\bar{\theta}_n}(\theta) \xrightarrow[n\to+\infty]{} 0$, il *suffit* donc que le biais et la variance tendent tous les deux vers 0.

Comment comparer deux estimateurs ?

▷ Approche naïve. Il serait naturel de dire que $\bar{\theta}_n$ est préférable à $\hat{\theta}_n$ si

$$R_{\bar{\theta}_n}(\theta^*) < R_{\hat{\theta}_n}(\theta^*)$$

L'écriture ci-dessus est incomplète. En effet, θ^* est inconnu. Par conséquent, on devrait écrire

$$\forall \theta^* \in \Theta, \quad R_{\bar{\theta}_n}(\theta^*) < R_{\hat{\theta}_n}(\theta^*)$$

En fait on est en train de comparer deux **fonctions**, ce qui n'est pas toujours possible.[13]

▷ 1er remède : approche minimax / worse-case risk. On définit

$$R^*_{\bar{\theta}_n} := \sup_{\theta \in \Theta} R_{\bar{\theta}_n}(\theta)$$

et on dit que θ_n est préférable à $\hat{\theta}_n$ si $R^*_{\bar{\theta}_n} < R^*_{\hat{\theta}_n}$. On dit que $\hat{\theta}_n$ est *minimax* (ou optimal) si $\hat{\theta}_n \in \arg\min_{\bar{\theta}_n} R^*_{\bar{\theta}_n}$.

▷ 2ème remède : approche bayésienne.[14] On introduit une fonction de poids $w : \Theta \longrightarrow [0,1]$. On définit le risque intégré :

$$R^{\mathsf{B}}_{\bar{\theta}_n} = \int_{\Theta} R_{\bar{\theta}_n}(\theta) w(\theta) d\theta$$

et on dit que $\bar{\theta}_n$ est préférable à $\hat{\theta}_n$ lorsque $R^{\mathsf{B}}_{\bar{\theta}_n} < R^{\mathsf{B}}_{\hat{\theta}_n}$. Si $\hat{\theta}^{\mathsf{B}}_n \in \arg\min_{\bar{\theta}_n} R^{\mathsf{B}}_{\bar{\theta}_n}$, alors $\hat{\theta}^{\mathsf{B}}_n$ est appelé estimateur Bayésien. Affaire à suivre...

[13] Si f et g sont deux fonctions à valeurs dans $\mathbb{R}$, on peut avoir $f(x) < g(x)$ et $f(y) > g(y)$ pour $x \neq y$.

[14] Approche moins pessimiste : là où dans la 1ère méthode on prenait la pire valeur possible, ici on se dit que cette pire valeur est peut-être moins probable et on tient compte de cela.

IV Méthode des moments

On considère $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, $\theta^* \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$. Soient $\varphi_1, \dots, \varphi_k$ des fonctions de $\mathbb{R}^m$ dans $\mathbb{R}$. On suppose que

$$\forall j \in [\![1, k]\!], \quad \forall \theta \in \Theta, \qquad \mathbb{E}_\theta[|\varphi_j(X_1)|] < +\infty.$$

On pose alors

$$m_j(\theta) = \mathbb{E}_\theta[\varphi_j(X_1)] \qquad \text{et} \qquad \hat{m}_j = \frac{1}{n}\sum_{i=1}^n \varphi_j(X_i).$$

Soit $\Phi : \mathbb{R}^m \longrightarrow \mathbb{R}^k$ définie par $\Phi(x) = (\varphi_1(x), \dots, \varphi_k(x))^\top$. Soit :

$$\begin{array}{rccl} M : & \Theta & \to & \mathbb{R}^k \\ & \theta & \mapsto & \mathbb{E}_\theta[\Phi(X_1)] = (m_1(\theta), \dots, m_k(\theta))^\top \end{array}$$

On pose enfin :

$$\hat{M}_n = \frac{1}{n}\sum_{i=1}^n \Phi(X_i) = (\hat{m}_1, \dots, \hat{m}_k)^\top$$

Définition IV.1 Estimateur par la méthode des moments

On dit que $\hat{\theta}_n^{\mathsf{MM}}$ est un estimateur de $\hat{\theta}^*$ par la méthode des moments lorsqu'il vérifie

$$\forall j \in [\![1, k]\!], \qquad m_j(\hat{\theta}_n^{\mathsf{MM}}) = \hat{m}_j$$

soit de manière équivalente

$$M\left(\hat{\theta}_n^{\mathsf{MM}}\right) = \hat{M}_n$$

Remarques.

— L'estimateur par la méthode des moments peut ne pas exister ou ne pas être unique : si $\hat{M}_n \notin \mathrm{Im}(M)$, alors $\hat{\theta}_n^{\mathsf{MM}}$ n'existe pas.

— Si M est injective et $\hat{M}_n \in \mathrm{Im}(M)$, alors $\hat{\theta}_n^{\mathsf{MM}}$ existe et est unique :

$$\hat{\theta}_n^{\mathsf{MM}} = M^{-1}(\hat{M}_n)$$

— Dans la version classique de la méthode des moments, on a généralement $m = 1$ et pour tout $j \in [\![1, k]\!]$, $\varphi_j(x) = x^j$.

Exemple. Soient $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.
Ici on a donc $k = 2$, $\varphi_1(x) = x$ et $\varphi_2(x) = x^2$, et pour tout $\theta = (\mu, \sigma^2)$, $M(\theta) = (\mu, \mu^2 + \sigma^2)^\top$.
On pose $\hat{\theta}_n^{\mathsf{MM}} = (\hat{\theta}_{n,1}^{\mathsf{MM}}, \hat{\theta}_{n,2}^{\mathsf{MM}})^\top$ et on résout le système :

$$\begin{aligned} M(\hat{\theta}_n^{\mathsf{MM}}) = \hat{M}_n &\iff \begin{cases} \hat{\theta}_{n,1}^{\mathsf{MM}} = \bar{X}_n \\ (\hat{\theta}_{n,1}^{\mathsf{MM}})^2 + \hat{\theta}_{n,2}^{\mathsf{MM}} = \overline{X_n^2} \end{cases} \\ &\iff \begin{cases} \hat{\theta}_{n,1}^{\mathsf{MM}} = \bar{X}_n \\ \hat{\theta}_{n,2}^{\mathsf{MM}} = S_n^2 \end{cases} \end{aligned}$$

Propriété IV.2

Si $M : \Theta \longrightarrow \mathbb{R}^k$ est continue et injective et si $\hat{\theta}_n^{\mathsf{MM}}$ existe, alors

$$\hat{\theta}_n^{\mathsf{MM}} \xrightarrow[n\to+\infty]{\text{p.s.}} \theta^*$$

Preuve. Sous ces hypothèses, $M : \Theta \longrightarrow \mathrm{Im}(M)$ est bijective et continue donc $M^{-1} : \mathrm{Im}(M) \longrightarrow \Theta$ est continue.
Par la LFGN,

$$\hat{M}_n \xrightarrow[n\to+\infty]{\text{p.s.}} M(\theta^*)$$

donc par le premier théorème de continuité,

$$M^{-1}(\hat{M}_n) \xrightarrow[n\to+\infty]{\text{p.s.}} M^{-1}(M(\theta^*))$$

ce qui achève la preuve. ■

Propriété IV.3

Ici on se place dans le cas $k = 1$. On suppose que :

- ▷ $\mathbb{E}_\theta[\|\Phi(X_1)\|^2] < +\infty$;
- ▷ M est C^1 et injective ;
- ▷ $\hat{\theta}_n^{\mathsf{MM}}$ existe.

Alors $\hat{\theta}_n^{\mathsf{MM}}$ est asymptotiquement normal.
Plus précisément, on a

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{MM}} - \theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \sigma_{MM}^2)$$

où $\sigma_{MM}^2 = (M^{-1})'(M(\theta^*))^2 \operatorname{Var}_{\theta^*}(\varphi(X_1))$.

Preuve. $M \in C^1$ donc $M^{-1} \in C^1$.
On a

$$\hat{\theta}_n^{\mathsf{MM}} = M^{-1}\left(\frac{1}{n}\sum_{i=1}^n \underbrace{\varphi(X_i)}_{Z_i}\right) = M^{-1}(\bar{Z}_n)$$

où $Z_1, \ldots, Z_n$ sont i.i.d. Le 2^{ème} théorème de continuité implique :

$$\sqrt{n}(M^{-1}(\bar{Z}_n) - M^{-1}(\underbrace{\mathbb{E}_{\theta^*}[Z_1]}_{a})) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \sigma_{MM}^2)$$

où $\sigma_{MM}^2 = (M^{-1})'(a)^2 \operatorname{Var}_{\theta^*}(Z_1)$. De plus,

$$a = \mathbb{E}_{\theta^*}[\varphi(X_1)] = M(\theta^*)$$

d'où le résultat. ■

Qu'en est-il des lois qui n'admettent pas d'espérance ?

Exemple. Soient $X_1, \ldots, X_n$ i.i.d. admettant pour densité

$$f(x,\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}, \quad x \in \mathbb{R}$$

Ici on a $\theta \in \mathbb{R} = \Theta$, $k = 1$. On *ne peut pas* choisir $\varphi(x) = x$ comme précédemment, car $\mathbb{E}_\theta[|\varphi(X_1)|] = +\infty$ pour tout $\theta \in \Theta$. On va plutôt prendre une fonction bornée. Par exemple :

$$\varphi(x) = \mathbb{1}_{]-\infty,0]}(x)$$

Alors

$$\begin{aligned} M(\theta) &= \mathbb{E}_\theta[\mathbb{1}(X_1 \leq 0)] \\ &= \int_{-\infty}^{0} \frac{dx}{\pi(1+(x-\theta)^2)} \\ &= \frac{1}{\pi}\Big(\frac{\pi}{2} - \arctan(\theta)\Big) \end{aligned}$$

Pour trouver l'estimateur des moments, on va donc résoudre :

$$\frac{1}{\pi}\Big(\frac{\pi}{2} - \arctan(\hat{\theta}_n^{\mathsf{MM}})\Big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq 0)$$

L'équation en $\hat{\theta}_n^{\mathsf{MM}}$ ci-dessus admet une unique solution :

$$\hat{\theta}_n^{\mathsf{MM}} = \tan\left(\frac{\pi}{2} - \frac{\pi}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq 0)\right)$$

qui nous donne ainsi l'existence, l'unicité et l'expression de $\hat{\theta}_n^{\mathsf{MM}}$.

V Méthode du maximum de vraisemblance

V.1 Définitions

On considère $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, $\theta^* \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$. Dans ce chapitre on suppose que **le modèle est dominé** : il existe une mesure σ-finie μ telle que pour tout $\theta \in \Theta$, $P_\theta \ll \mu$. On note :

$$f(x, \theta) = \frac{dP_\theta}{d\mu}(x)$$

i.e. $f(\cdot, \theta)$ est la densité de la loi P_θ par rapport à μ.[15] On pose également :

$$\ell(x, \theta) = \log(f(x, \theta))$$

[15] Dont l'existence est assurée par Radon-Nykodim.

Deux cas particuliers importants :

— Si le modèle est à densité, on prend $\mu =$ la mesure de Lebesgue et $f(\cdot, \theta) =$ densité de P_θ par rapport à la mesure de Lebesgue.

— Si le modèle est discret : il existe une famille $(a_j)_{j \in \mathbb{N}} \in (\mathbb{R}^m)^{\mathbb{N}}$ telle que

$$\sum_{j \in \mathbb{N}} P_\theta(X = a_j) = 1$$ [16]

[16] On s'autorisera souvent comme ici à écrire X au lieu de X_i, car les X_i ont de toute façon tous la même loi.

Dans ce cas, on prend :

$$\mu = \sum_{j \in \mathbb{N}} \delta_{a_j}$$

i.e. μ est la mesure de comptage sur $A = \{a_j : j \in \mathbb{N}\}$.
On a alors $f(x, \theta) = P_\theta(X = x) = \sum_{j \in \mathbb{N}} p_j(\theta) \mathbb{1}_{x = a_j}$, où $p_j(\theta) = P_\theta(X = a_j)$.

> En effet, pour tout $B \in \mathcal{B}(\mathbb{R}^m)$,
>
> $$\begin{aligned} P_\theta(X \in B) &= P_\theta(X \in B \cap A) + \underbrace{P_\theta(X \in B \setminus A)}_{\leq P_\theta(X \notin A) = 0} \\ &= P_\theta(X \in B \cap A) \\ &= \sum_{x \in B \cap A} P_\theta(X = x) \\ &= \int_B P_\theta(X = x) d\mu(x) \end{aligned}$$
>
> donc $x \longmapsto P_\theta(X = x)$ est la densité de X_1.

Exemples.

1. Si $P_\theta = \mathcal{B}(\theta)$ alors

$$\begin{aligned} f(x, \theta) &= \theta \mathbb{1}(x = 1) + (1 - \theta) \mathbb{1}(x = 0) && \text{pour tout réel } x \\ &= \theta^x (1 - \theta)^{1 - x} && \text{si } x \in \{0, 1\} \end{aligned}$$

2. Si $P_\theta = \mathcal{E}(\theta)$ alors

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{[0, +\infty[}(x)$$

Remarque. Lorsque $P_\theta = \sum_{j \in \mathbb{N}} p_j(\theta) \delta_{a_j}$, on a

$$f(x, \theta) = \prod_{j \in \mathbb{N}} p_j(\theta)^{\mathbb{1}(x = a_j)}$$

Définition V.1 Vraisemblance

On appelle vraisemblance du modèle $\{P_\theta : \theta \in \Theta\}$ la fonction

$$\begin{array}{rccl} L_n : & \mathcal{X}^n \times \Theta & \longrightarrow & [0, +\infty[\\ & (x_1, \dots x_n, \theta) & \longmapsto & \prod_{i=1}^{n} f(x_i, \theta) \end{array}$$

Remarque. Si les X_i ne sont pas indépendants,[17] on définira plutôt L_n comme la densité jointe de $X_1, \dots, X_n$.

[17] Lorsqu'il y a indépendance les deux définitions coïncident.

Définition V.2 EMV

On dit que $\hat{\theta}_n^{\mathsf{MV}}$ est un estimateur du maximum de vraisemblance lorsque

$$L_n(X_1, \dots, X_n, \hat{\theta}_n^{\mathsf{MV}}) = \max_{\theta \in \Theta} L_n(X_1, \dots, X_n, \theta)$$

i.e.

$$\hat{\theta}_n^{\mathsf{MV}} \in \arg\max_{\theta \in \Theta} L_n(X_1, \dots, X_n, \theta)$$

Remarques.

— Dans certains modèles, l'EMV n'existe pas ou n'est pas unique.

— Avec la convention $\log 0 = -\infty$, on a

$$\hat{\theta}_n^{\mathsf{MV}} \in \arg\min_{\theta \in \Theta} \ell_n(\theta)$$

où l'on note :

$$\begin{aligned} \ell_n(\theta) &= -\frac{1}{n} \log(L_n(X_1, \dots, X_n, \theta)) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \log(f(X_i, \theta)) \end{aligned}$$

V.2 Motivation de la méthode

Condition C : pour tout $\theta, \theta^* \in \Theta$, $\displaystyle\int_{\mathbb{R}^m} |\log(f(x, \theta))| f(x, \theta^*) d\mu(x) < +\infty.$

Posons $Z_i = -\log(f(X_i, \theta))$.[18] Les $X_1, \dots, X_n$ étant i.i.d., les $Z_1, \dots, Z_n$ le sont aussi. La condition C entraîne $\mathbb{E}_{\theta^*}[|Z_1|] < +\infty$. Par la LFGN, $\ell_n(\theta) = \bar{Z}_n \xrightarrow[n \to +\infty]{\text{p.s.}} \mathbb{E}_{\theta^*}[Z_1]$.

[18] La loi de Z_i dépend donc de θ, mais aussi de θ^* puisque $X_i \sim P_{\theta^*}$.

Posons :

$$\begin{aligned} J(\theta) &= \mathbb{E}_{\theta^*}[Z_1] \\ &= -\int_{\mathbb{R}^m} \log(f(x, \theta)) f(x, \theta^*) d\mu(x) \end{aligned}$$

Théorème V.3

Supposons que la condition C est satisfaite. Pour tout $\theta \in \Theta$:

$$J(\theta) \geq J(\theta^*)$$

L'égalité a lieu si et seulement si $P_\theta = P_{\theta^*}$.

Preuve. On a

$$\begin{aligned}
J(\theta) &= \mathbb{E}_{\theta^*}[-\log f(X_1,\theta)] \\
&= \underbrace{\mathbb{E}_{\theta^*}[-\log f(X_1,\theta^*)]}_{=J(\theta^*)} + \mathbb{E}_{\theta^*}\left[-\log \frac{f(X_1,\theta)}{f(X_1,\theta^*)}\right] \\
&\geq J(\theta^*) - \log\left(\mathbb{E}_\theta\left[\frac{f(X_1,\theta)}{f(X_1,\theta^*)}\right]\right) \qquad \text{par Jensen}
\end{aligned}$$

Or $\mathbb{E}_\theta\left[\frac{f(X_1,\theta)}{f(X_1,\theta^*)}\right] = \int_{\mathbb{R}^m} \frac{f(x,\theta)}{f(x,\theta^*)} f(x,\theta^*)d\mu(x) = 1$ donc $J(\theta) \geq J(\theta^*)$.

Étude des cas d'égalité. Par ce qui précède, on a

$$J(\theta) - J(\theta^*) = -\mathbb{E}_{\theta^*}\left[\log \frac{f(X_1,\theta)}{f(X_1,\theta^*)}\right] = -\mathbb{E}_{\theta^*}\left[\log \frac{f(X_1,\theta)}{f(X_1,\theta^*)} - \frac{f(X_1,\theta)}{f(X_1,\theta^*)} + 1\right]$$

Or pour tout $x \in \mathbb{R}_+^*$, $\log x \leq x - 1$. La v.a. $Y = \log \frac{f(X_1,\theta)}{f(X_1,\theta^*)} - \frac{f(X_1,\theta)}{f(X_1,\theta^*)} + 1$ est donc négative.

Supposons que de plus $J(\theta) = J(\theta^*)$, alors Y est centrée et négative donc $Y = 0$ P_{θ^*}-p.s., et comme $\log x = x - 1$ si et seulement si $x = 1$, on obtient

$$f(X_1,\theta) = f(X_1,\theta^*) \quad P_{\theta^*}\text{-p.s.}$$

donc (exercice) $P_\theta = P_{\theta^*}$. ■

Corollaire V.4

Supposons que la condition C est satisfaite et $\{P_\theta : \theta \in \Theta\}$ est identifiable. Alors

$$J(\theta) = J(\theta^*) \iff \theta = \theta^*$$

ce qui veut dire que θ^* est l'unique minimiseur de J.

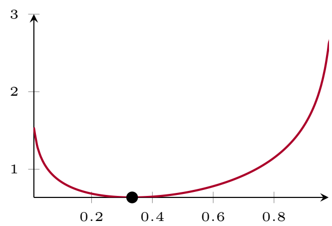

Figure 4 *Courbe de J sur $[0,1]$ lorsque $X_1,\ldots,X_n$ sont i.i.d. de loi Bernoulli de paramètre $\theta^* = 1/3$. On constate que* $\arg\min_{\theta\in[0,1]} J(\theta) = 1/3$.

Utilité du résultat ci-dessus en pratique. Si on connaît la « tête » de la densité $f(\cdot,\theta)$ mais pas le vrai paramètre θ^*, on peut simuler à partir des observations $X_1,\ldots,X_n$ une approximation de $J(\theta) = \mathbb{E}_{\theta^*}[Z_1]$, par exemple en estimant $J(\theta)$ par $\bar{Z}_n = \ell_n(\theta)$ (méthode des moments). En utilisant des méthodes d'optimisation on peut ensuite obtenir une approximation du minimiseur de J qui est donc, dans le cas d'un modèle identifiable, une approximation de θ^*.

V.3 Exemples

Étudions l'existence et, le cas échéant, l'expression de l'EMV dans les cas suivants.

1. **Modèle de Bernoulli :** $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\theta^*)$.

On prend $\Theta = [0,1]$, $f(x,\theta) = \theta^x(1-\theta)^{1-x}$. On a

$$\begin{aligned}
\ell_n(\theta) &= -\frac{1}{n}\sum_{i=1}^n \Big(X_i \log\theta + (1-X_i)\log(1-\theta)\Big) \\
&= -\bar{X}_n \log\theta - (1-\bar{X}_n)\log(1-\theta)
\end{aligned}$$

On obtient alors

$$\ell_n'(\theta) = \frac{\theta - \bar{X}_n}{\theta(1-\theta)}$$

On constate que l'EMV existe, est unique et vérifie $\hat{\theta}_n^{\mathsf{MV}} = \bar{X}_n$.

2. **Modèle exponentiel :** $X_1, \dots, X_n$ i.i.d. de densité $f(\cdot, \theta^*)$, où :

$$f(x,\theta) = \frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{[0,+\infty[}(x)$$

> On a
> $$\ell_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log\left(\frac{e^{-X_i/\theta}}{\theta}\right) = \log\theta + \frac{1}{\theta}\bar{X}_n$$
> puis
> $$\ell_n'(\theta) = \frac{1}{\theta} - \frac{\bar{X}_n}{\theta^2} = \frac{\theta - \bar{X}_n}{\theta^2}$$
> On constate que l'EMV existe, est unique et vérifie $\hat{\theta}_n^{\mathsf{MV}} = \bar{X}_n$.

3. On suppose $X_1, \dots, X_n$ i.i.d. de densité $f(\cdot, \theta^*)$, où :

$$f(x,\theta) = \begin{cases} \dfrac{C}{\sqrt{|x-\theta|}} e^{-|x-\theta|} & \text{si } x \neq \theta \\ 0 & \text{si } x = \theta \end{cases}$$

où $C = \dfrac{1}{\int_{\mathbb{R}} |x|^{-1/2} e^{-|x|} dx}$.

> On a
> $$\begin{aligned} \ell_n(\theta) &= -\frac{1}{n}\sum_{i=1}^n \left(\log C - \frac{1}{2}\log|X_i - \theta| - |X_i - \theta|\right) \\ &= -\log C + \frac{1}{2n}\sum_{i=1}^n \log|X_i - \theta| + \frac{1}{n}\sum_{i=1}^n |X_i - \theta| \end{aligned}$$
> On a $\lim_{\theta\to X_1} \ell_n(\theta) = -\infty$. En particulier $\theta \longmapsto \ell_n(\theta)$ n'est pas minorée donc $\hat{\theta}_n^{\mathsf{MV}}$ n'existe pas.

4. **Modèle double-exponentiel/modèle de Laplace :** $X_1, \dots, X_n$ i.i.d. de densité $f(\cdot, \theta^*)$, où :

$$f(x,\theta) = \frac{1}{2} e^{-|x-\theta|}$$

> On a
> $$\begin{aligned} \ell_n(\theta) &= -\frac{1}{n}\sum_{i=1}^n \left(\log\frac{1}{2} - |X_i - \theta|\right) \\ &= \log 2 + \frac{1}{n}\sum_{i=1}^n |X_i - \theta| \end{aligned}$$
> donc $\hat{\theta}_n^{\mathsf{MV}} \in \underset{\theta\in\mathbb{R}}{\arg\min} \underbrace{\frac{1}{n}\sum_{i=1}^n |X_i - \theta|}_{=:g(\theta)}$.
> On ordonne les X_i : on définit ainsi n nouvelles variables aléatoires $X_{(1)} \leq \dots \leq X_{(n)}$. On a
> $$g(\theta) = \frac{1}{n}\sum_{i=1}^n |X_{(i)} - \theta|$$
> En posant $X_{(0)} = -\infty$ et $X_{(n+1)} = +\infty$, on a donc $\theta \in]X_{(k)}, X_{(k+1)}[$ pour

un certain $k \in [\![0, n]\!]$. Donc

$$\begin{aligned} g(\theta) &= \frac{1}{n}\sum_{i=1}^{k} |X_{(i)} - \theta| + \frac{1}{n}\sum_{i=k+1}^{n} |X_{(i)} - \theta| \\ &= \frac{1}{n}\sum_{i=1}^{k} (\theta - X_{(i)}) + \frac{1}{n}\sum_{i=k+1}^{n} (X_{(i)} - \theta) \\ &= \frac{2k-n}{n}\theta + \text{cste} \end{aligned}$$

— <u>1^er^ cas :</u> $n = 2m$ est pair.
g est continue, décroissante sur $]-\infty, X_{(m)}]$, constante sur $]X_{(m)}, X_{(m+1)}[$ puis croissante sur $]X_{(m+1)}, +\infty[$.
Donc tout point de $[X_{(m)}, X_{(m+1)}]$ est un EMV.
Ici l'EMV existe mais n'est pas unique !

— <u>2^ème^ cas :</u> $n = 2m+1$ est impair.
Par un raisonnement similaire on prouve que $\hat{\theta}_n^{\mathsf{MV}} = X_{(m+1)}$, la médiane de $X_1, \ldots, X_n$.
Ici l'EMV existe et est unique !

V.4 Consistance de l'EMV

Théorème V.5

Soit $\Theta \subset \mathbb{R}$ un ouvert. On suppose que :

(i) pour tout x, l'application $\theta \longmapsto f(x, \theta)$ est continue ;

(ii) le modèle est identifiable ;

(iii) intégrabilité sous P_{θ^*} : $\displaystyle\int_{\mathbb{R}^m} |\log(f(x,\theta))| f(x,\theta^*) d\mu(x) < +\infty$;

(iv) tous les minima locaux de $\ell(\theta)$ forment un intervalle fermé de $\mathbb{R}$.

Alors :

$$\hat{\theta}_n^{\mathsf{MV}} \xrightarrow[n\to+\infty]{P_{\theta^*}} \theta^*$$

En pratique, on utilise peu ce théorème pour prouver la consistance car la 4^ème^ condition est archi chiante à prouver. Mais le théorème montre que la consistance est vraie dans un cadre assez général.

Preuve. Soit $\varepsilon > 0$. Montrons que

$$P_{\theta^*}(|\hat{\theta}_n^{\mathsf{MV}} - \theta^*| > \varepsilon) \xrightarrow[n\to+\infty]{} 0$$

Posons $A_n = \{|\hat{\theta}_n^{\mathsf{MV}} - \theta^*| \leq \varepsilon\}$.

On commence par montrer que :

$$A_n \supset \{\ell_n(\theta^* - \varepsilon) - \ell_n(\theta^*) > 0\} \cap \{\ell_n(\theta^* + \varepsilon) - \ell_n(\theta^*) > 0\}$$

1^ère^ **étape**. Montrons d'abord que tous les minima locaux de ℓ_n sont des minima globaux.
On note

$$I = \{\theta \in \Theta : \theta \text{ est un min. local}\}$$

I peut s'écrire sous la forme $[a, b]$ (4^ème^ condition du théorème).

Soit $\theta_{\max} \in \arg\max_{\theta \in I} \ell_n(\theta)$.
On introduit :

$$I_1 = \{\theta \in I : \theta \leq \theta_{\max}, \ell_n(\theta) < \ell_n(\theta_{\max})\}$$

Supposons que $I_1 \neq \emptyset$.
Alors I_1 admet une borne supérieure : notons-la a_1.
On sait que

$$\begin{cases} \forall \theta < a_1, & \ell_n(\theta) < \ell_n(\theta_{\max}) \quad (*) \\ \forall \theta \in \,]a_1, \theta_{\max}], & \ell_n(\theta) = \ell_n(\theta_{\max}) \end{cases}$$

Donc $\ell_n(a_1) = \ell_n(\theta_{\max})$.
Donc $\ell_n(a_1) = \ell_n(\theta_{\max})$.
Or $a_1 \in I$, donc a_1 est un min. local de ℓ_n.
Donc il existe $\delta > 0$ tel que $[a_1 - \delta, a_1] \subset I$ et pour tout $\theta \in [a_1 - \delta, a_1]$, on a

$$\ell_n(\theta) \geq \ell_n(a_1) = \ell_n(\theta_{\max})$$

Donc

$$\ell_n(\theta) = \ell_n(\theta_{\max}), \quad \forall \theta \in [a_1 - \delta, a_1]$$

ce qui contredit $(*)$.
Ainsi I_1 est vide, et donc pour toutu $\theta \in [a, \theta_{\max}]$ on a $\ell_n(\theta) = \ell_n(\theta_{\max})$.
De la même façon on pourrait montrer que

$$\forall \theta \in [\theta_{\max}, b], \quad \ell_n(\theta) = \ell_n(\theta_{\max})$$

Conclusion : ℓ_n est constante sur I. Comme I contient également les min. globaux de ℓ_n, on en déduit que tous les points de I sont des min. globaux.

2ème étape. Supposons que $\ell_n(\theta^* - \varepsilon) > \ell_n(\theta^*)$ et $\ell_n(\theta^* + \varepsilon) > \ell_n(\theta^*)$.
Soit $\theta_0 \in \arg\min_{[\theta^* - \varepsilon, \theta^* + \varepsilon]} \ell_n(\theta)$.
Alors $\theta_0 \in I$ car c'est un min. local.
D'autre part,

$$\ell_n(\theta^* + \varepsilon) > \ell_n(\theta^*) \geq \ell_n(\theta_0)$$

donc $\theta^* + \varepsilon$ et $\theta^* - \varepsilon$ n'appartiennent pas à I, et ainsi

$$I \subset \,]\theta^* - \varepsilon, \theta^* + \varepsilon[$$

Enfin, $\hat{\theta}_n^{\mathsf{MV}}$ est un min. global de ℓ_n donc $\hat{\theta}_n^{\mathsf{MV}} \in I$, ce qui achève la preuve du lemme.

En vertu de ce qui précède,

$$P(A_n^c) \leq P(\ell_n(\theta^* - \varepsilon) - \ell_n(\theta^*) < 0) + P(\ell_n(\theta^* + \varepsilon) - \ell_n(\theta^*) < 0)$$

Rappelons que

$$\forall \theta \in \Theta, \quad \ell_n(\theta) \xrightarrow[n \to +\infty]{\text{p.s.}} J(\theta)$$

Posons $\delta = J(\theta^* - \varepsilon) - J(\theta^*) > 0$. Alors

$$P(\ell_n(\theta^* - \varepsilon) - \ell_n(\theta^*) < 0) = P(\ell_n(\theta^* - \varepsilon) - J(\theta^* - \varepsilon) + J(\theta^*) - \ell_n(\theta^*) < -\delta) \xrightarrow[n \to +\infty]{} 0$$

■

V.5 Modèles réguliers

On considère $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, avec ici $\boxed{\theta^* \in \Theta \subset \mathbb{R}.}$

Notations. Lorsqu'on écrit $f'(x, \theta)$, $\ell'(x, \theta)$, $\ell''(x, \theta)$... etc., implicitement la dérivation est par rapport à θ.

Définition V.6 Modèle régulier

On dit que le modèle $\{P_\theta : \theta \in \Theta\}$ est régulier s'il vérifie ces quatre hypothèses :

(H1) Le support de P_θ (ensemble des points en lesquels la densité est non nulle) ne dépend pas de θ. Autrement dit :

$$\forall \theta, \theta' \in \Theta, \quad f(x,\theta) > 0 \implies f(x,\theta') > 0$$

(H2) Les fonctions $\ell(x,\cdot)$ et $f(x,\cdot)$ sont C^2 pour μ-presque tous les x.

(H4) $\forall \theta \in \Theta, \quad \underbrace{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(X,\theta)\right)^2\right]}_{\text{information de Fisher}} > 0.$

(H3) *(Non-exigible.)* Pour tout $\theta^* \in \Theta$, il existe un ouvert $U \subset \Theta$ et un $\Lambda : \mathcal{X} \longrightarrow \mathbb{R}$ mesurable tels que

$$\sup_{\theta \in U}\left(|\ell'(x,\theta)| + |\ell''(x,\theta)| + |\ell'(x,\theta)|^2\right) \leq \Lambda(x)$$

et

$$\int \Lambda(x) \sup_{\theta\in U} f(x,\theta) d\mu(x) < +\infty$$

Remarques.

1. Essentiellement ce à quoi sert (H3) c'est à pouvoir faire des inversions dérivée-intégrale dans la suite.
2. Les modèles Bernoulli, Binomiale, Géométrique, Gaussien, exponentiel sont réguliers.
   En revanche, le modèle uniforme $\{\mathcal{U}([0,\theta]) : \theta \in \mathbb{R}\}$ et le modèle de Laplace (P_θ de densité $f(x,\theta) \propto e^{-|x-\theta|}$) ne sont pas réguliers.[19]
3. Cette définition s'étend aux cas multidimensionnels en remplaçant dans (H3) les dérivées par les dérivées partielles et dans (H4) $\mathbb{E}_\theta[\ell'(X,\theta)^2] > 0$ par l'hypothèse que la matrice symétrique $\mathbb{E}_\theta[\nabla_\theta \ell(X,\theta)\nabla_\theta \ell(X,\theta)^\top]$ est définie positive.

[19] Pour le modèle uniforme, c'est parce que le support de $\mathcal{U}([0,\theta])$ dépend de θ. Pour le modèle de Laplace, c'est parce que les densités ne sont pas dérivables en θ donc (H2) non vérifiée.

V.6 Information de Fisher

Soit $\{P_\theta : \theta \in \Theta\}$ un modèle **régulier**.

Définition V.7 Score, information de Fisher

On appelle score du modèle $\{P_\theta : \theta \in \Theta\}$ la fonction

$$\begin{array}{rccl} s_\theta : & \mathcal{X} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \ell'(x,\theta) \end{array}$$

(On note souvent $s(x,\theta)$ plutôt que $s_\theta(x)$.)
On appelle information de Fisher du modèle $\{P_\theta : \theta \in \Theta\}$ la fonction :

$$I(\theta) = \mathrm{Var}_\theta(s(X,\theta))$$

Remarque. Si $\Theta \subset \mathbb{R}^p$, on définit $I(\theta)$ comme la matrice de covariance de $\nabla_\theta \ell(X,\theta)$.

Propriété V.8

Si $\{P_\theta : \theta \in \Theta\}$ est régulier, alors le score est centré :

$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta[s(X,\theta)] = 0$$

Preuve. Pour tout $\theta \in \Theta$:

$$\int_{\mathcal{X}} f(x,\theta) d\mu(x) = 1$$

donc

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(x,\theta) d\mu(x) \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial \theta} f(x,\theta) d\mu(x) \\ &= \int_{\mathcal{X}} \ell'(x,\theta) f(x,\theta) d\mu(x) \\ &= \mathbb{E}_\theta[\ell'(X,\theta)] \end{aligned}$$

■

Conséquence. Dans un modèle régulier,

$$I(\theta) = \mathbb{E}_\theta[s(X,\theta)^2]$$

et l'hypothèse (H4) est équivalente à : pour tout $\theta \in \Theta$, $I(\theta) > 0$.

Propriété V.9

Dans un modèle régulier,

$$I(\theta) = -\mathbb{E}_\theta[\ell''(X,\theta)]$$

Preuve. Comme le score est centré, pour tout $\theta \in \Theta$:

$$\int_{\mathcal{X}} \ell'(x,\theta) f(x,\theta) d\mu(x) = 0$$

puis en dérivant à nouveau par rapport à θ :

$$\int_{\mathcal{X}} \big(\ell''(x,\theta) f(x,\theta) + \ell'(x,\theta) f'(x,\theta)\big) \, d\mu(x) = 0$$

or $f'(x,\theta) = \ell'(x,\theta) f(x,\theta)$ donc :

$$\mathbb{E}_\theta[\ell''(X,\theta)] + \underbrace{\int_{\mathcal{X}} \ell'(x,\theta)^2 f(x,\theta) d\mu(x)}_{=I(\theta)} = 0$$

■

Exemple. Si $P_\theta = \text{Poisson}(\theta)$, $\theta > 0$, alors pour tout $x \in \mathbb{N}$:

$$f(x,\theta) = \frac{\theta^x}{x!} e^{-\theta}$$

On a $\ell(x,\theta) = x \log \theta - \theta - \log(x!)$ donc $\ell'(x,\theta) = \frac{x}{\theta} - 1$ puis

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{X}{\theta} - 1\right)^2\right] = \text{Var}_\theta\left(\frac{X}{\theta}\right) = \frac{1}{\theta^2} \text{Var}_\theta(X) = \frac{1}{\theta}$$

On peut aussi la calculer en utilisant le dernier théorème, puisque $\ell''(X,\theta) = -\frac{X}{\theta^2}$.

V.7 Normalité asymptotique de l'EMV

On considère $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, avec ici $\boxed{\theta^* \in \Theta \subset \mathbb{R}.}$

Théorème V.10

Supposons que :

1. $\hat{\theta}_n^{\mathsf{MV}}$ existe ;
2. $\hat{\theta}_n^{\mathsf{MV}}$ est consistant ;
3. $\{P_\theta : \theta \in \Theta\}$ est identifiable et régulier.

Alors

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{MV}} - \theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0, \frac{1}{I(\theta^*)}\right)$$

Preuve. Comme Θ est ouvert et $\theta \longmapsto \ell(x, \theta)$ est C^1,

$$\hat{\theta}_n^{\mathsf{MV}} \in \underset{\theta\in\Theta}{\arg\min}\, \ell_n(\theta)$$

donc

$$\ell_n'(\hat{\theta}_n^{\mathsf{MV}}) = 0$$

Par le théorème des accroissements finis, il existe $\tilde{\theta}_n \in [\theta^*, \hat{\theta}_n^{\mathsf{MV}}]$ tel que

$$\underbrace{\ell_n'(\hat{\theta}_n^{\mathsf{MV}})}_{=0} - \ell_n'(\theta^*) = \ell_n''(\tilde{\theta}_n)(\hat{\theta}_n^{\mathsf{MV}} - \theta^*)$$

d'où

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{MV}} - \theta^*) = -\frac{\sqrt{n}\ell_n'(\theta^*)}{\ell_n''(\tilde{\theta}_n)} \tag{3}$$

D'autre part, on a

$$-\sqrt{n}\ell_n'(\theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{\Big[\log f(X_i, \theta^*)\Big]'}_{=s(X_i,\theta^*)}$$

On sait que $\mathbb{E}_{\theta^*}[s(X_i, \theta^*)] = 0$. Donc on peut écrire, artificiellement :

$$-\sqrt{n}\ell_n'(\theta^*) = \sqrt{n}\Big(\overline{s(X,\theta^*)} - \mathbb{E}_{\theta^*}[s(X_1, \theta^*)]\Big) \quad \text{où :} \quad \overline{s(X,\theta^*)} = \frac{1}{n}\sum_{i=1}^n s(X_i, \theta^*)$$

Le TCL donne :

$$-\sqrt{n}\ell_n'(\theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \underbrace{\mathrm{Var}_{\theta^*}(s(X_1, \theta^*))}_{=I(\theta^*)}) \tag{4}$$

De plus, on a :

$$\ell_n''(\tilde{\theta}_n) = \ell_n''(\theta^*) + \ell_n''(\tilde{\theta}_n) - \ell_n''(\theta^*)$$

avec, par la LFGN :

$$\ell_n''(\theta^*) = -\frac{1}{n}\sum_{i=1}^n \ell''(X_i, \theta^*) \xrightarrow[n\to+\infty]{\text{p.s.}} -\mathbb{E}_{\theta^*}[\ell''(X_1, \theta^*)] = I(\theta^*) \tag{5}$$

Enfin, $\tilde{\theta}_n \in [\theta^*, \hat{\theta}_n^{\mathsf{MV}}]$ entraîne

$$|\tilde{\theta}_n - \theta^*| \leq |\hat{\theta}_n^{\mathsf{MV}} - \theta^*| \xrightarrow[n\to+\infty]{P} 0$$

donc $\tilde{\theta}_n \xrightarrow[n\to+\infty]{P} \theta^*$.

La dernière étape de la démonstration, qui est admise (mais intuitive) consisterait à montrer que

$$\ell''_n(\tilde{\theta}_n) - \ell''_n(\theta^*) \xrightarrow[n\to+\infty]{P} 0 \tag{6}$$

En combinant (3), (4), (5), (6) et Slutsky :

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{MV}} - \theta^*) = \frac{-\sqrt{n}\ell'_n(\theta^*)}{\ell''_n(\tilde{\theta}_n)} \xrightarrow[n\to+\infty]{\text{loi}} \frac{1}{I(\theta^*)} Z$$

où $Z \sim \mathcal{N}(0, I(\theta^*))$, ce qui achève la preuve. ■

Remarques.

1. Le théorème est valable même si $\hat{\theta}_n^{\mathsf{MV}}$ n'est pas unique.
2. Le théorème peut être étendu au cas où $\Theta \subset \mathbb{R}^d$ de la manière suivante :[20]

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{MV}} - \theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}_d(0, (I(\theta^*))^{-1})$$

3. Le théorème reste vrai si on remplace l'EMV par tout point stationnaire de $\ell_n(\theta)$, i.e. $\hat{\theta}_n$ tel que $\ell'_n(\hat{\theta}_n) = 0$.

[20] La matrice $I(\theta^*)$ est bien inversible car par (H4) $I(\theta^*)$ est définie positive.

Exemple. Soient $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta)$ où $\theta \in \Theta =]0, +\infty[$. Montrer que l'EMV existe, est asymptotiquement normal, et trouver la variance asymptotique.

On a

$$f(x, \theta) = \frac{e^{-x^2/2\theta}}{\sqrt{2\pi\theta}}$$

donc

$$\ell(x, \theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta - \frac{x^2}{\theta}$$

(H1) $\operatorname{supp}(\mathcal{N}(0, \theta)) = \mathbb{R}$, indépendant de θ ;

(H2) $\theta \longmapsto \ell(x, \theta)$ est de classe C^2 pour tout $x \in \mathbb{R}$;

(H4) On a

$$\ell'(x, \theta) = -\frac{1}{2\theta} + \frac{x^2}{2\theta^2}$$

donc

$$\ell''(x, \theta) = \frac{1}{2\theta^2} - \frac{x^2}{\theta^3}$$

donc

$$\mathbb{E}_\theta[\ell'(X_1, \theta)] = \frac{1}{2\theta^2} - \frac{\mathbb{E}_\theta[X_1^2]}{\theta^3} = -\frac{1}{2\theta^2}$$

On peut aussi passer par la variance comme on l'a vu dans une propriété précédente :

$$\begin{aligned}
\operatorname{Var}_\theta(\ell'(X_1, \theta)) &= \operatorname{Var}_\theta\left(\frac{X_1^2}{2\theta^2}\right) \\
&= \frac{1}{4\theta^2}\Big(\underbrace{\mathbb{E}_\theta[X_1^4]}_{=3\theta^2} - \underbrace{\mathbb{E}_\theta[X_1^2]^2}_{=\theta^2}\Big) \\
&= \frac{1}{2\theta^2} > 0
\end{aligned}$$

On a

$$\hat{\theta}_n^{\mathsf{MV}} \in \operatorname*{arg\,min}_{\theta>0} \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2}\log\theta + \frac{X_i^2}{2\theta}\right) = \operatorname*{arg\,min}_{\theta>0} \underbrace{\left(\frac{1}{2}\log\theta + \frac{1}{2\theta}\overline{X^2}\right)}_{\varphi(\theta)}$$

avec

$$\varphi'(\theta) = \frac{\theta - \overline{X^2}}{2\theta^2}$$

de sorte que

$$\varphi'(\theta) > 0 \iff \theta > \overline{X^2}$$

et donc $\hat{\theta}_n^{\mathsf{MV}} = \overline{X^2}$ est l'EMV de θ^*.
Par la LFGN,

$$\overline{X^2} \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbb{E}_{\theta^*}[X_1^2] = \theta^*$$

donc $\hat{\theta}_n^{\mathsf{MV}}$ est consistant.
Toutes les conditions sont remplies pour appliquer le théorème précédent : on a

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{MV}} - \theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, 2(\theta^*)^2)$$

V.8 Interprétation de l'information de Fisher

Introduisons la notion de courbure. Imaginons par exemple qu'on veuille minimiser $\varphi : [0,1] \longrightarrow \mathbb{R}$. Notons $\theta^* = \arg\min_{\theta\in[0,1]} \varphi(\theta)$. Quitte à faire une transformation affine, supposons que $\varphi(\theta^*) = 0$. Soit $\bar{\theta}$ une approximation de θ^* pour laquelle $\varphi(\bar{\theta}) \leq \varepsilon$, avec $\varepsilon \in \mathbb{R}_+^*$ petit. Question : quelle est la distance entre θ^* et $\bar{\theta}$? Réponse : plus la « courbure » de φ est grande, meilleure est l'approximation. Cela se voit intuitivement par le dessin, voir Figure 5.

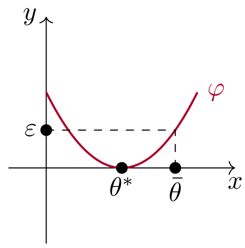

Figure 5 *Courbure de φ : plus $\varphi''(\theta^*)$ est grand, plus la distance entre $\bar{\theta}$ et θ^* est petite, et donc meilleure est l'approximation.*

Définition V.11 DKL

Soient P et Q deux mesures de probabilité sur $(\mathcal{X}, \mathcal{A})$. On appelle divergence de Kullback-Leibler la quantité

$$D_{\mathsf{KL}}(P\|Q) = \begin{cases} \displaystyle\int_{\mathcal{X}} \frac{dP}{dQ}(x) \log \frac{dP}{dQ}(x) dQ(x) & \text{si } P \ll Q \\ +\infty & \text{sinon} \end{cases}$$

Remarques. Dans le cas $P \ll Q$:

— $D_{\mathsf{KL}}(P\|Q) = \mathbb{E}[X \log(X)]$ où X est une variable aléatoire de densité dP/dQ.

— Si $P \ll \mu$ et $Q \ll \mu$,[21] alors

$$D_{\mathsf{KL}}(P\|Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x)$$

où $p = dP/d\mu$ et $q = dQ/d\mu$. En effet, $\displaystyle\frac{dP}{dQ}(x) = \frac{\frac{dP}{d\mu}(x)}{\frac{dQ}{d\mu}(x)} = \frac{p(x)}{q(x)}$.

[21] En fait dès que $P \ll Q$, on peut trouver une mesure μ telle que $P \ll \mu$ et $Q \ll \mu$: prendre $\mu = \frac{P+Q}{2}$.

Propriété V.12

1. $D_{\mathsf{KL}}(P\|Q) \geq 0$
2. $D_{\mathsf{KL}}(P\|Q) = 0$ si et seulement si $P = Q$
3. $D_{\mathsf{KL}}(P\|Q) \neq D_{\mathsf{KL}}(Q\|P)$ en général (donc D_{KL} n'est pas une distance)
4. $(P, Q) \longmapsto D_{\mathsf{KL}}(P\|Q)$ est convexe.

Mais alors, quel foutu lien avec l'EMV ?

Le point 2. de la propriété précédente entraîne que $\theta^* \in \arg\min_{\theta\in\Theta} D_{\mathsf{KL}}(P_{\theta^*}\|P_\theta)$ si le modèle est identifiable. On va alors utiliser la méthode de substitution : si $\theta^* = \Phi(P^*)$ alors $\hat{\theta}_n = \Phi(\hat{P}_n)$ où $\hat{P}_n$ est la mesure empirique. Il serait naturel de définir l'estimateur de DKL minimale :

$$\begin{aligned}\hat{\theta}_n \in\ & \arg\min_{\theta\in\Theta} D_{\mathsf{KL}}(\hat{P}_n \| P_\theta) & (\bigstar)\\ & \arg\min_{\theta\in\Theta}\left(\underbrace{\int_{\mathcal{X}} \hat{p}_n(x)\log \hat{p}_n(x)d\mu(x)}_{\text{indépendant de } \theta} - \int_{\mathcal{X}} \hat{p}_n(x)\log f(x,\theta)d\mu(x)\right)\\ & \arg\min_{\theta\in\Theta}\left(-\int_{\mathcal{X}} \log f(x,\theta)d\hat{P}_n(x)\right)\\ & \arg\min_{\theta\in\Theta}\left(-\frac{1}{n}\sum_{i=1}^{n}\log f(X_i,\theta)\right)\end{aligned}$$

Ainsi cet estimateur coïncide avec l'EMV.

Remarque. $(\bigstar)$ n'est pas toujours bien défini car on peut avoir $D_{\mathsf{KL}}(\hat{P}_n \| P_\theta) = +\infty$ pour tout $\theta \in \Theta$. C'est le cas, en particulier, des modèles à densité. Voici une formalisation propre de $(\bigstar)$:

$$\begin{aligned}\theta^* \in\ & \arg\min_{\theta\in\Theta} D_{\mathsf{KL}}(P_{\theta^*} \| P_\theta)\\ & \arg\min_{\theta\in\Theta}\left(\underbrace{\int_{\mathcal{X}} p^*(x)\log p^*(x)d\mu(x)}_{\text{indépendant de } \theta} - \int_{\mathcal{X}} p^*(x)\log f(x,\theta)d\mu(x)\right)\\ & \underbrace{\arg\min_{\theta\in\Theta}\left(-\int_{\mathcal{X}} \log f(x,\theta)dP^*(x)\right)}_{\widetilde{\Phi}(P^*)}\end{aligned}$$

et on a alors $\hat{\theta}_n^{\mathsf{MV}} = \widetilde{\Phi}(\hat{P}_n)$.

Propriété V.13

Dans un modèle régulier, l'information de Fisher est la courbure de la divergence de Kullback-Leibler :

$$I(\theta) = \frac{d^2}{d\theta^2} D_{\mathsf{KL}}(P^* \| P_\theta)$$

Preuve. On a

$$D_{\mathsf{KL}}(P^* \| P_\theta) = \int_{\mathcal{X}} p^* \log p^* d\mu - \int_{\mathcal{X}} \log f(x,\theta)dP^*(x)$$

donc

$$\begin{aligned}\frac{d^2}{d\theta^2} D_{\mathsf{KL}}(P^* \| P_\theta) &= \frac{d^2}{d\theta^2}\left(-\int_{\mathcal{X}} \log f(x,\theta)dP^*(x)\right)\\ &\overset{\text{(H3)}}{=} -\int_{\mathcal{X}} \frac{d^2}{d\theta^2}\log f(x,\theta)dP^*(x)\\ &= I(\theta)\end{aligned}$$

■

Théorème V.14 Inégalité de Cramer-Rao

Supposons que le modèle $\{P_\theta : \theta \in \Theta\}$ est dominé et régulier.

Si $\bar{\theta}_n$ est un estimateur de θ basé sur $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, alors

$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta[(\bar{\theta}_n - \theta)^2] \geq \frac{(1 + b'(\theta))^2}{nI(\theta)} + b(\theta)^2$$

où $b(\theta) = \mathbb{E}_\theta[\bar{\theta}_n] - \theta$ est le biais de $\bar{\theta}_n$.

En particulier, si $\bar{\theta}_n$ est sans biais, alors $\mathrm{Var}_\theta(\bar{\theta}_n) \geq \dfrac{1}{nI(\theta)}$.

Preuve. Par décomposition biais-variance du risque quadratique, on a $\mathbb{E}_\theta[(\bar{\theta}_n - \theta)^2] = \mathrm{Var}_\theta(\bar{\theta}_n) + b(\theta)^2$. Il suffit donc de montrer que $\mathrm{Var}_\theta(\bar{\theta}_n) \geq \frac{(1+b'(\theta))^2}{nI(\theta)}$.
L'inégalité de Cauchy-Schwarz donne

$$\mathrm{cov}_\theta(s_n(X,\theta), \bar{\theta}_n)^2 \leq \mathrm{Var}_\theta(s_n(X,\theta))\,\mathrm{Var}_\theta(\bar{\theta}_n)$$

où $s_n(X,\theta) = \sum_{i=1}^n \ell'(X_i,\theta)$.
Or le fait que les X_i sont i.i.d. et la définition de $I(\theta)$ entraînent $\mathrm{Var}_\theta(s_n(X,\theta)) = nI(\theta)$. D'où

$$\mathrm{Var}_\theta(\bar{\theta}_n) \geq \frac{1}{nI(\theta)} \mathrm{cov}_\theta(s_n(X,\theta), \bar{\theta}_n)^2$$

D'autre part, la v.a. $s_n(X,\theta)$ est centrée donc

$$\begin{aligned}
\mathrm{cov}_\theta(s_n(X,\theta), \bar{\theta}_n) &= \mathbb{E}_\theta[s_n(X,\theta)\bar{\theta}_n] \\
&= \mathbb{E}_\theta\left[\left(\frac{d}{d\theta}\sum_{i=1}^n \ell(X_i,\theta)\right)\bar{\theta}_n\right]
\end{aligned}$$

Notons $f_n(x,\theta) = f(x_1,\theta)\cdots f(x_n,\theta)$ la densité jointe de $X = (X_1,\dots,X_n)$. Alors

$$\begin{aligned}
\mathrm{cov}_\theta(s_n(X,\theta), \bar{\theta}_n) &= \mathbb{E}_\theta\left[\left(\frac{d}{d\theta}\log f_n(X,\theta)\right)\bar{\theta}_n\right] \\
&= \mathbb{E}_\theta\left[\frac{f_n'(X,\theta)}{f_n(X,\theta)}\bar{\theta}_n(X)\right] \\
&= \int_{\mathcal{X}} \frac{f_n'(x,\theta)}{f_n(x,\theta)}\bar{\theta}_n(x) f_n(x,\theta)d\mu_n(x) \\
&= \frac{d}{d\theta}\int_{\mathcal{X}} f_n(x,\theta)\bar{\theta}_n(x)d\mu_n(x) \\
&= \frac{d}{d\theta}\mathbb{E}_\theta[\bar{\theta}_n(X)] = \frac{d}{d\theta}(\theta + b(\theta)) = 1 + b'(\theta)
\end{aligned}$$

Conclusion :

$$\mathbb{E}_\theta[(\bar{\theta}_n - \theta)^2] = \mathrm{Var}_\theta(\bar{\theta}_n) + b(\theta)^2 \geq \frac{(1+b'(\theta))^2}{nI(\theta)} + b(\theta)^2$$

■

V.9 Efficacité asymptotique

L'objectif de cette section est de déterminer quel est le meilleur estimateur. Dans les comparaisons faites ici, on ne considère que les estimateurs asymptotiquement normaux.[22]

[22] i.e. les estimateurs qui ne le sont pas sont d'ores et déjà considérés comme mauvais. C'est logique car à un TCL et une méthode δ près, on peut souvent se démerder pour trouver un estimateur asymptotiquement normal de la forme $g(\bar{X}_n)$. Donc on se restreint à l'étude des estimateurs potentiellement meilleurs que $g(\bar{X}_n)$.

Définition V.15

Soient $\bar{\theta}_n^{(1)}$ et $\bar{\theta}_n^{(2)}$ deux estimateurs asymptotiquement normaux de θ. On note $v^{(1)}(\theta)$ et $v^{(2)}(\theta)$ leurs variances asymptotiques.

- ▷ On dit que $\bar{\theta}_n^{(1)}$ est asymptotiquement meilleur que $\bar{\theta}_n^{(2)}$ lorsque

$$\begin{cases} \forall \theta \in \Theta, & v^{(1)}(\theta) \leq v^{(2)}(\theta) \\ \exists \theta_0 \in \Theta, & v^{(1)}(\theta_0) < v^{(2)}(\theta_0) \end{cases}$$

- ▷ $\hat{\theta}_n$ est asymptotiquement efficace s'il n'existe pas d'estimateur qui soit asymp. meilleur que $\hat{\theta}_n$.

Théorème V.16

Dans un modèle régulier, un estimateur est asymptotiquement efficace si et seulement si

$$\forall \theta \in \Theta, \quad v(\theta) = \frac{1}{I(\theta)}$$

Conséquence. Si $\hat{\theta}_n^{\mathsf{MV}}$ existe et est consistant et si $\{P_\theta : \theta \in \Theta\}$ est identifiable et régulier, alors $\hat{\theta}_n^{\mathsf{MV}}$ est asymptotiquement efficace.

VI Estimateur de Bayes

VI.1 Définitions

On considère un modèle $\{P_\theta : \theta \in \Theta\}$ **dominé**, $\Theta \subset \mathbb{R}^p$ où $p \in \mathbb{N}^*$, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, $\theta^* \in \Theta$. On note $\overrightarrow{X} = (X_1, \dots, X_n)$ et

$$L_n(\overrightarrow{X}, \theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

la vraisemblance du modèle.

Notation. Soit $h : \mathbb{R}^p \longrightarrow \mathbb{R}_+$ mesurable telle que $0 < \int_{\mathbb{R}^p} h(\theta) d\theta < +\infty$. La phrase « soit π la densité donnée par $\pi(\theta) \propto h(\theta)$ » signifie qu'on définit π comme étant la seule densité proportionnelle à h, i.e. $\pi(\theta) = C \times h(\theta)$ où $C = \frac{1}{\int_{\mathbb{R}^p} h(\theta) d\theta}$.

Définition VI.1 Estimateur bayésien

On dit que $\bar{\theta}_n^{\mathsf{B}}$ est un estimateur bayésien s'il minimise le risque intégré par rapport à une loi *a priori* π_0 :

$$\bar{\theta}_n^{\mathsf{B}} \in \arg\min_{\bar{\theta}} \int_\Theta \mathbb{E}_\theta[\|\bar{\theta} - \theta\|^2] d\pi_0(\theta)$$

i.e. $\bar{\theta}_n^{\mathsf{B}} \in \arg\min_{\bar{\theta}} \int_{\Theta \times \mathcal{X}^n} \|\bar{\theta}_n(\overrightarrow{x}) - \theta\|^2 L_n(\overrightarrow{x}, \theta) d\mu_n(\overrightarrow{x}) d\pi_0(\theta)$, où $d\mu_n(\overrightarrow{x}) = d\mu(x_1) \dots d\mu(x_n)$.

Convention. On supposera toujours que π_0 admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$. Par abus de notation on notera encore π_0 cette densité.

Théorème VI.2

Soit $\pi_n(\theta)$ la densité donnée par

$$\pi_n(\theta) \propto L_n(\overrightarrow{X}, \theta) \pi_0(\theta)$$

Si $\displaystyle\int_\Theta \|\theta\|^2 \pi_n(\theta) d\theta < +\infty$, alors l'estimateur bayésien existe et

$$\bar{\theta}_n^{\mathsf{B}} = \int_\Theta \theta \pi_n(\theta) d\theta$$

Preuve. Pour tout estimateur $\bar{\theta}_n$ de θ,

$$\begin{aligned}
\int_\Theta \mathbb{E}_\theta[\|\bar{\theta}_n - \theta\|^2] d\pi_0(\theta) &= \int_\Theta \mathbb{E}_\theta[\|\bar{\theta}_n - \bar{\theta}_n^{\mathsf{B}} + \bar{\theta}_n^{\mathsf{B}} - \theta\|^2] d\pi_0(\theta) \\
&= \int_\Theta \mathbb{E}_\theta\left[\|\bar{\theta}_n - \bar{\theta}_n^{\mathsf{B}}\|^2 + \|\bar{\theta}_n^{\mathsf{B}} - \theta\|^2 + 2\langle \bar{\theta}_n - \bar{\theta}_n^{\mathsf{B}}, \bar{\theta}_n^{\mathsf{B}} - \theta \rangle\right] d\pi_0(\theta) \\
&\geq \int_\Theta \mathbb{E}_\theta[\|\bar{\theta}_n^{\mathsf{B}} - \theta\|^2] d\pi_0(\theta) + 2 \int_\Theta \mathbb{E}_\theta[\langle \bar{\theta}_n - \bar{\theta}_n^{\mathsf{B}}, \bar{\theta}_n^{\mathsf{B}} - \theta \rangle] d\pi_0(\theta)
\end{aligned}$$

Or

$$\begin{aligned}\int_\Theta \mathbb{E}_\theta[\langle \bar\theta_n - \bar\theta_n^{\mathsf{B}}, \bar\theta_n^{\mathsf{B}} - \theta\rangle] d\pi_0(\theta) &= \int_{\Theta\times\mathcal{X}^n} \langle \bar\theta_n - \bar\theta_n^{\mathsf{B}}, \bar\theta_n^{\mathsf{B}} - \theta\rangle L_n(\overrightarrow{x},\theta)\pi_0(\theta)d\mu_n(\overrightarrow{x})d\theta \\ &= \left\langle \bar\theta_n - \bar\theta_n^{\mathsf{B}},\ \bar\theta_n^{\mathsf{B}} - \int_\Theta \theta\pi_n(\theta)d\theta \right\rangle \\ &= \langle \bar\theta_n - \bar\theta_n^{\mathsf{B}}, 0\rangle \\ &= 0\end{aligned}$$

Ainsi $\bar\theta_n^{\mathsf{B}}$ minimise le risque intégré : c'est un estimateur bayésien. ■

Remarques. π_0 est la loi/densité *a priori*, on la choisit indépendamment des observations $\overrightarrow{X}$. π_n est la loi/densité *a posteriori*.

Interprétation. Si on considère que $L_n(\overrightarrow{X},\theta)$ est la densité conditionnelle[23] de $\overrightarrow{X}$ sachant θ où $\theta \sim \pi_0$, alors $\pi_n(\theta)$ est la densité conditionnelle de θ sachant $\overrightarrow{X}$. En effet, la densité jointe de $(\theta, \overrightarrow{X})$ est $\pi_0(\theta)L_n(\overrightarrow{x},\theta)$, et on définit :

$$\pi_n(\theta) \propto L_n(\overrightarrow{x},\theta)\pi_0(\theta)$$

[23] La densité de Y sachant $X = x$ est donnée par

$$f_{Y|X=x}(y) = \frac{f_{(X,Y)}(x,y)}{f_X(x)}$$

Puisque $f_X(x)$ ne dépend pas de y, $f_{Y|X=x}$ est la seule densité proportionnelle à $y \longmapsto f_{(X,Y)}(x,y)$.
La densité de Y sachant X, quant à elle, est obtenue en remplaçant x par X dans l'expression de $f_{Y|X=x}(y)$. De sorte que $f_{Y|X}(y)$ est une **variable aléatoire**.

Exemple. Soit $\{P_\theta : \theta \in \Theta\} = \{\mathcal{E}(a) : a > 0\}$ où ici histoire de perdre tout le monde on convient que $\mathcal{E}(a)$ est la loi de densité $f(x,a) = ae^{-ax}\mathbb{1}(x \geq 0)$. On peut choisir $\pi_0(a) = e^{-a}\mathbb{1}(a \geq 0)$. Alors,

$$L_n(\overrightarrow{X},a) = a^n \exp\left(-a\sum_{i=1}^n X_i\right) = a^n e^{-an\bar X_n}$$

donc par le théorème,

$$\pi_n(a) \propto a^n e^{-an\bar X_n} e^{-a}\mathbb{1}(a\geq 0) = a^n e^{-a(n\bar X_n+1)}\mathbb{1}(a \geq 0)$$

On constate que π_n est la densité de $\Gamma(n+1, n\bar X_n + 1)$. Donc $\int_0^{+\infty} a\pi_n(a)da$ est l'espérance de la loi $\Gamma(n+1, n\bar X_n+1)$, d'où :

$$\bar a_n^{\mathsf{B}} = \frac{n+1}{n\bar X_n + 1} = \frac{1+\frac{1}{n}}{\bar X_n + \frac{1}{n}}$$

Extension. La moyenne *a posteriori* est un estimateur bayésien lorsque le risque quadratique est utilisé. Si on remplace le risque quadratique par $\mathbb{E}_\theta[|\bar\theta_n - \theta|]$ ($\Theta \subset \mathbb{R}$), alors $\bar\theta_n^{\mathsf{B}}$ est la médiane de π_n i.e. la médiane *a posteriori*. De manière générale, si on mesure le risque par $\mathbb{E}_\theta[\varphi(\bar\theta_n,\theta)]$, alors

$$\bar\theta_n^{\mathsf{B}}(\overrightarrow{X}) \in \arg\min_{a\in\mathbb{R}^p} \int_\Theta \varphi(a,\theta)\pi_n(\theta,\overrightarrow{X})d\theta$$

i.e.

$$\bar\theta_n^{\mathsf{B}} \in \arg\min_{\bar\theta} \int_\Theta R(\bar\theta,\theta)\pi_0(\theta)d\theta$$

VI.2 Choix de la loi *a priori*

Il n'y a pas de choix universel de π_0. Ce choix est toujours subjectif. Quelques principes utiles :

1. Choisir un π_0 dont le support est tout Θ, i.e. $\pi_0(\theta) > 0$ pour tout $\theta \in \Theta$;

2. Choisir π_0 de telle sorte que le calcul de $\bar{\theta}_n^{\mathsf{B}}$ soit faisable : ou bien par un calcul explicite, ou bien par une approximation utilisant la méthode de Monte-Carlo.

Définition VI.3 Famille conjuguée

Soit $\mathcal{P}_0$ une famille de lois sur Θ. On dit que $\mathcal{P}_0$ est une famille conjuguée pour le modèle $\{P_\theta : \theta \in \Theta\}$ lorsque pour toute loi *a priori* π_0 de $\mathcal{P}_0$, la loi *a posteriori* π_n appartient aussi à $\mathcal{P}_0$.

Exemples.

1. $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\theta^*)$, $\Theta = [0,1]$,
$$f(x,\theta) = \theta^x(1-\theta)^{1-x}, \quad x \in \{0,1\}$$
Soit $\mathcal{P}$ la famille des lois bêta :
$$\mathcal{P} = \left\{\theta \longmapsto \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} : \alpha > 0, \beta > 0\right\}$$
On a $L_n(\overrightarrow{X},\theta) = \theta^{n\bar{X}_n}(1-\theta)^{n-n\bar{X}_n}$. Si $\pi_0 \in \mathcal{P}$ alors
$$\pi_n(\theta) \propto \theta^{n\bar{X}_n+\alpha-1}(1-\theta)^{n-n\bar{X}_n+\beta-1}\mathbb{1}_{[0,1]}(\theta)$$
où $\pi_n(\theta)$ est la densité de la loi $\mathrm{Beta}(n\bar{X}_n+\alpha, n(1-\bar{X}_n)+\beta)$ d'où $\pi_n \in \mathcal{P}$. Ainsi $\mathcal{P}$ est une famille conjuguée pour le modèle de Bernoulli et
$$\hat{\theta}_n^{\mathsf{B}} = \frac{n\bar{X}_n+\alpha}{n+\alpha+\beta} = \frac{\bar{X}_n + \frac{\alpha}{n}}{1+\frac{\alpha+\beta}{n}}$$

2. $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{P}(\theta^*)$, $f(x,\theta) = \frac{\theta^x}{x!}e^{-\theta}$, $x \in \mathbb{N}$, $\Theta = [0,+\infty[$.
Soit $\mathcal{P}_0$ la famille des lois Gamma.
Si on prend $\pi_0 \in \mathcal{P}_0$ alors $\pi_0 \propto \theta^{\alpha-1}e^{-\beta\theta}$ où $\alpha > 0$ et $\beta > 0$.
La densité *a posteriori* $\pi_n(\theta) \propto L_n(\overrightarrow{X},\theta)\pi_0(\theta)$, où
$$L_n(\overrightarrow{X},\theta) = \frac{1}{\prod_{i=1}^n X_i!}\theta^{n\bar{X}_n}e^{-n\theta}$$
donc
$$\pi_n(\theta) \propto \theta^{n\bar{X}_n+\alpha-1}e^{-(n+\beta)\theta}$$
Ainsi $\pi_n \in \mathcal{P}_0$ car c'est la densité de $\Gamma(\alpha_n,\beta_n)$ où $\alpha_n = n\bar{X}_n+\alpha$ et $\beta_n = n+\beta$ d'où
$$\bar{\theta}_n^{\mathsf{B}} = \frac{\alpha_n}{\beta_n}$$

Remarques.

1. $\bar{\theta}_n^{\mathsf{B}} - \bar{X}_n = \frac{\alpha-\beta\bar{X}_n}{n+\beta}$
2. $n(\bar{\theta}_n^{\mathsf{B}} - \bar{X}_n) \xrightarrow[n\to+\infty]{\text{p.s.}} \alpha - \beta\theta^*$
3. $\boxed{\bar{\theta}_n^{\mathsf{B}} - \bar{X}_n = \frac{1}{n}O_P(1).}$ (admis, grosse dinguerie)

Par conséquent,
$$\sqrt{n}(\bar{\theta}_n^{\mathsf{B}} - \theta^*) = \underbrace{\sqrt{n}(\bar{X}_n - \theta^*)}_{\xrightarrow[n\to+\infty]{\text{loi}}\mathcal{N}(0,\theta^*)} + \underbrace{\sqrt{n}(\bar{\theta}_n^{\mathsf{B}} - \bar{\theta}_n)}_{\xrightarrow[n\to+\infty]{\text{p.s.}}0}$$
donc par Slutsky, $\sqrt{n}(\bar{\theta}_n^{\mathsf{B}} - \theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0,\theta^*) = \mathcal{N}(0, 1/I(\theta^*))$. Ainsi $\bar{\theta}_n^{\mathsf{B}}$ est asymptotiquement efficace.

VI.3 Propriétés de $\bar{\theta}_n^{\mathsf{B}}$

Théorème VI.4

Dans un modèle régulier, si π_0 a pour support Θ et si $\bar{\theta}_n^{\mathsf{B}}$ est consistant, alors

$$\sqrt{n}(\bar{\theta}_n^{\mathsf{B}} - \theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0, \frac{1}{I(\theta^*)}\right)$$

Par conséquent, $\bar{\theta}_n^{\mathsf{B}}$ est asymptotiquement efficace.

Le théorème est admis. En gros, voilà l'idée de la preuve. On a $\pi_n(\theta) \propto e^{-n\ell_n(\theta)+\log\pi_0(\theta)} \propto e^{-n(\ell_n(\theta)-\ell_n(\hat{\theta}_n^{\mathsf{MV}})+\log\pi_0(\theta)}$. Ensuite on fait un DL2 de $\ell_n(\theta)$ au voisinage de $\hat{\theta}_n^{\mathsf{MV}}$ où, en utilisant le fait que $\ell_n'(\hat{\theta}_n^{\mathsf{MV}}) = 0$ (car $\hat{\theta}_n^{\mathsf{MV}}$ maximise la log-vraisemblance), on obtient $\ell_n(\theta) - \ell_n(\hat{\theta}_n^{\mathsf{MV}}) \approx \frac{1}{2}(\theta - \hat{\theta}_n^{\mathsf{MV}})^\top \ell_n''(\hat{\theta}_n^{\mathsf{MV}})(\theta - \hat{\theta}_n^{\mathsf{MV}})$. De plus $\ell_n''(\hat{\theta}_n^{\mathsf{MV}}) \approx \ell_n''(\theta^*) = -\frac{1}{n}\sum_{i=1}^n \ell''(X_i, \theta^*) \xrightarrow[n\to+\infty]{\text{p.s.}} -\mathbb{E}_{\theta^*}[\ell''(X,\theta)] = I(\theta^*)$. Bilan : $\ell_n(\theta) - \ell_n(\hat{\theta}_n^{\mathsf{MV}}) \approx \frac{1}{2}(\theta - \hat{\theta}_n^{\mathsf{MV}})^\top I(\theta^*)(\theta - \hat{\theta}_n^{\mathsf{MV}})$. Et donc $\pi_n(\theta)$ est proportionnelle à la densité de $\mathcal{N}(\hat{\theta}_n^{\mathsf{MV}}, (nI(\theta^*))^{-1})$. Et donc, intuitivement, $\int_\Theta \theta\tilde{\pi}_n(\theta)d\theta \approx \hat{\theta}_n^{\mathsf{MV}}$ par conséquent $n\mathbb{E}_{\theta^*}[(\bar{\theta}_n^{\mathsf{B}} - \theta^*)(\bar{\theta}_n^{\mathsf{B}} - \theta^*)^\top] \approx I(\theta^*)^{-1}$. Logique, quoi !

VI.4 Calcul approché de $\bar{\theta}_n^{\mathsf{B}}$ par la méthode de Monte-Carlo

Ici, $\Theta = \mathbb{R}$, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$ et

$$\pi_n(\theta) \propto e^{-n\ell_n(\theta)+\log\pi_0(\theta)}, \qquad \bar{\theta}_n^{\mathsf{B}} = \int_\Theta \theta\pi_n(\theta)d\theta$$

Définition VI.5 Algorithme de Monte-Carlo

L'algorithme de Monte-Carlo est le suivant :

1. Choisir $N \in \mathbb{N}^*$ (penser cet entier comme grand)
2. Générer $\theta_1, \theta_2, \dots, \theta_N \overset{\text{i.i.d.}}{\sim} \pi_n$
3. On calcule :

$$\tilde{\theta}_n^{\mathsf{B}} = \frac{1}{N}\sum_{i=1}^N \theta_i$$

qui par la LFGN converge p.s. vers $\mathbb{E}_{\theta\sim\pi_n}[\theta] = \bar{\theta}_n^{\mathsf{B}}$ lorsque $N \to +\infty$.

L'étape 2 est en pratique celle qui est la plus difficile à réaliser. Pour cela, on peut utiliser :

— Inversion de la f.d.r. : $\theta_i = F_n^{-1}(U_i)$ où $U_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$ et F_n est la fonction de répartition de π_n.
Problème : inapplicable dans le cas multivarié.

— Acceptation-rejet : voir TD.
Problème : très coûteux en temps.

— On fait de l'approximation : $\theta_i \sim \nu_n$ tel que $d_{\mathsf{VT}}(\nu_n, \pi_n) < \varepsilon = 10^{-3}$, où $d_{\mathsf{VT}}(\nu_n, \pi_n) = \sup_{A\in\mathcal{B}(\mathbb{R})} |\nu_n(A) - \pi_n(A)| = \frac{1}{2}\int_\Theta |\nu_n(x) - \pi_n(x)|dx$.[24]

— Méthode de Langevin :

[24] d_{VT} représente en gros la *distance infinie* sur l'espace des mesures de probas sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$: à quel point ν_n diffère *au pire* de π_n.

Définition VI.6 Méthode de Langevin

On souhaite générer θ dans $\mathbb{R}^d$ tel que la loi de θ est à peu près la loi associée à $\pi_0(\theta) \propto e^{-f(\theta)}$ où $f : \mathbb{R}^d \longrightarrow \mathbb{R}$.

- ▷ Initialisation : $\theta_0 \in \mathbb{R}^d$.
- ▷ Hérédité : $\theta_{k+1} = \theta_k - h\nabla f(\theta_k) + \sqrt{2h}\xi_k$ où $h < 0$, $\xi_0, \xi_1, \ldots$ sont i.i.d. de loi $\mathcal{N}(0, I_d)$.

La propriété suivante est issue de papiers de recherche vieux de 5-6 ans grand max :

Propriété VI.7

Si f est m-fortement convexe ($m > 0$) et ∇f est M-lipschitzien, alors pour tout $h \leq 1/M$,

$$\text{distance}(\text{loi de } \theta_k, \pi_n) \leq (1 - mh)^K \text{distance}(\text{loi de } \theta_0, \pi_n) + \sqrt{\frac{M}{m} dh}$$

On n'introduit pas précisément toutes les définitions mais en gros : imaginons qu'on arrive effectivement à définir une distance sur l'espace des lois. Approximer π_n à ε près ($\varepsilon \in \mathbb{R}_+^*$), c'est trouver une loi π satisfaisant $\text{distance}(\pi, \pi_n) \leq \varepsilon$.[25] Ce qui rend la propriété ci-dessus utile, c'est qu'en prenant h et K tels que

$$\begin{cases} \sqrt{\frac{M}{m} dh} \leq \frac{\varepsilon}{2} \\ (1 - mh)^K \text{distance}(\text{loi de } \theta_0, \pi_n) \leq \frac{\varepsilon}{2} \end{cases}$$

i.e.

$$\begin{cases} h \leq \frac{m\varepsilon^2}{4dM} \\ K \geq \ln\left(\frac{\varepsilon}{2 \text{ distance}(\text{loi de } \theta_0, \pi_n)}\right) \Big/ \ln(1 - mh) \end{cases}$$

alors π_n est une approximation à ε près de la loi de θ_k.

[25] Par analogie avec le cas réel : lorsqu'on veut approcher la solution x_0 d'une équation d'inconnue réelle, on dit que x est une approximation de x_0 à ε près lorsque $|x_0 - x| \leq \varepsilon$.

VII Ensembles de confiance

Hors programme du partiel.

VII.1 Définition et exemple

On considère un modèle $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$ où $p \in \mathbb{N}^*$, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta^*}$, $\theta^* \in \Theta$.

L'objectif est d'estimer θ^* par un ensemble, au lieu de l'estimer par un singleton – ce qui a été fait jusqu'à présent. L'avantage est que cela permet de quantifier l'incertitude de l'estimation.

Définition VII.1

Soit $\alpha \in [0,1]$ petit (5%, 1%...). On dit que l'ensemble aléatoire $E \subset \mathbb{R}^p$ est un ensemble de confiance de niveau $1-\alpha$ pour θ^* si

(i) E est construit en se basant sur les observations uniquement, de sorte que son expression n'utilise pas θ^*.

(ii) $\forall \theta \in \Theta, \quad P_\theta(E \ni \theta) \geq 1-\alpha$.

Lorsqu'on a égalité dans l'inégalité ci-dessus pour tout θ, on dira que E est de niveau exact $1-\alpha$.
Enfin si $E = E_n = E(X_1, \dots, X_n)$ vérifie

$$\lim_{n \to +\infty} P_\theta(E_n \ni \theta) \geq 1-\alpha$$

alors on dira que $(E_n)_{n\in\mathbb{N}}$ est un ensemble de confiance de niveau asymptotique $1-\alpha$.

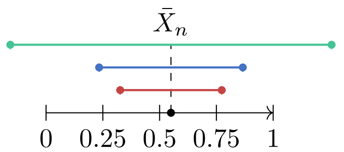

Figure 6 *Intervalles de confiance pour θ^* dans le modèle de Bernoulli ($n=50$) au niveau $1-\alpha$ pour $\alpha = 90\%$ (rouge), $\alpha = 95\%$ (bleu) et $\alpha = 99\%$ (vert). On voit le tradeoff entre avoir une région de confiance relativement restreinte mais un faible niveau de confiance, VS avoir un niveau de confiance élevé mais une région de confiance absurdement grande.*

Remarques.

1. Le point (i) revient à dire que $\{\omega : E(\omega) \ni \theta\}$ appartient à $\sigma(X_1, \dots, X_n)$ pour tout $\theta \in \Theta$.
2. On dit « ensemble de confiance asymptotique » par abus de langage même si techniquement $(E_n)_{n\in\mathbb{N}}$ est une suite d'ensembles de confiance.

Exemple. Si $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\theta^*)$ avec $\theta^* \in [0,1]$ inconnu, alors

$$I = \left[\bar{X}_n - \frac{1}{2\sqrt{n\alpha}}, \bar{X}_n + \frac{1}{2\sqrt{n\alpha}}\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour θ^*.

En effet, par Tchebychev,

$$P_\theta\left(|\bar{X}_n - \theta| > \frac{1}{2\sqrt{n\alpha}}\right) \leq \frac{\mathbb{E}_\theta[(\bar{X}_n - \theta)^2]}{\left(\frac{1}{2\sqrt{n\alpha}}\right)^2} = 4n\alpha \operatorname{Var}_\theta(\bar{X}_n) \leq \alpha$$

i.e. $P_\theta(I \ni \theta) \geq 1-\alpha$.

Remarques.

1. Dans un modèle paramétrique régulier, la taille d'un intervalle de confiance tend vers 0 à la vitesse $\frac{1}{\sqrt{n}}$ lorsque $n \to +\infty$.

2. Dans l'exemple précédent, l'intervalle I est de longueur de I est $\frac{1}{\sqrt{n\alpha}}$. Cette longueur est sous-optimale dans le cas de variables aléatoires bornées. Cette sous-optimalité vient de l'inégalité de Tchebychev. Elle peut être améliorée en utilisant à la place l'inégalité de Hoeffding ou l'inégalité de Bernstein, ce qui permettrait de remplacer $\frac{1}{\sqrt{\alpha}}$ par $\sqrt{\log(1/\alpha)}$

Propriété VII.2 Inégalité de Hoeffding

Soit (X_n) une suite de variables aléatoires indépendantes. Soient (a_n) et (b_n) deux suites de nombres réels telles que, pour tout entier n, on a $P(a_n \leq X_n \leq b_n) = 1$. Alors, notant $S_n = X_1 + \cdots + X_n$, pour tout $t > 0$ et tout $n \in \mathbb{N}$, on a

$$P\left(\left|S_n - E(S_n)\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right)$$

Propriété VII.3 Inégalité de Bernstein

Soit $a < b$ deux réels, soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes et vérifiant $a < X_i < b$ pour tout $i = 1, \ldots, n$. Alors on a

$$P\left(\left|\sum_{i=1}^n (X_i - E(X_i))\right| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{n(b-a)^2}\right)$$

VII.2 Méthode de la fonction pivotale

Supposons qu'on a à notre disposition un estimateur consistant $\hat{\theta}_n$ de θ^* et une fonction $G : \mathbb{R}^p \times \Theta \longrightarrow \mathbb{R}^p$ telle que $G(\hat{\theta}_n, \theta^*)$ est un vecteur aléatoire dont la loi ne dépend pas de θ^*. On dit que G est une *fonction pivotale.*

Supposons pour le moment que $p = 1$. Pour tout $\beta \in [0, 1]$, on note q_β le quantile d'ordre β de $G(\hat{\theta}_n, \theta^*)$. Comme G est pivotale, q_β ne dépend pas de θ^*.

Propriété VII.4

L'ensemble

$$E = \left\{\theta \in \Theta : G(\hat{\theta}_n, \theta) \in [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]\right\}$$

est un ensemble de confiance de niveau $1 - \alpha$

Preuve. $P_\theta(E \ni \theta) = P_\theta(G(\hat{\theta}_n, \theta) \in [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]) \geq 1 - \alpha$ par définition des quantiles. ■

Extension au cas $p > 1$: remplacer $G(\hat{\theta}_n, \theta^*)$ par $\|G(\hat{\theta}_n, \theta^*)\|$.[26]

[26] On a la liberté du choix de la norme.

VII.3 Méthode basée sur le théorème de Slutsky

Supposons que le modèle est régulier et que l'information de Fisher est une fonction continue.

Propriété VII.5

L'intervalle

$$I_n = \left[\hat{\theta}_n - \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{nI(\hat{\theta}_n)}},\ \hat{\theta}_n + \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{nI(\hat{\theta}_n)}}\right]$$

où q^N_β est le β-quantile de $\mathcal{N}(0,1)$, est de niveau de confiance asymptotique $1-\alpha$.

VII.4 Stabilisation de la variance

Supposons qu'on a un estimateur $\hat{\theta}_n$ de θ^* asymptotiquement normal :

$$\forall \theta \in \Theta^*, \quad \sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, v(\theta^*))$$

Par la méthode delta, on sait que pour toute fonction $g : \mathbb{R} \longrightarrow \mathbb{R}$ de classe C^1,

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta^*)) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, g'(\theta^*)^2 v(\theta^*))$$

Choisissons g telle que pour tout x de Θ, $g'(x)^2 = 1/v(x)$. Posons

$$E_n = \left\{\theta \in \Theta : g(\theta) \in \left[g(\hat{\theta}_n) - \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{n}},\ g(\hat{\theta}_n) + \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right]\right\}$$

Alors $P_\theta(E_n \ni \theta) \xrightarrow[n\to+\infty]{} 1-\alpha$.

Exemple. Si $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta^*)$, $\theta^* \in]0, +\infty[$. Nous avons déjà montré que $\hat{\theta}_n^{\mathsf{MV}} = \bar{X}_n$. Comme $\mathbb{E}_\theta[X_1] = \text{Var}_\theta(X_1) = \theta$, on obtient

$$\sqrt{n}(\hat{\theta}_n^{\mathsf{MV}} - \theta) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0, \theta)$$

g doit être solution de $g'(x)^2 = 1/x$. Il suffit de prendre par exemple $g(x) = 2\sqrt{x}$. Par la méthode présentée précédemment, l'ensemble

$$E_n = \left\{\theta : 2\sqrt{\theta} \in \left[2\sqrt{\bar{X}_n} - \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{n}},\ 2\sqrt{\bar{X}_n} + \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right]\right\}$$

est de niveau asymptotique $1-\alpha$. Et on constate que

$$E_n = \left[\left(\sqrt{\bar{X}_n} - \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)^2, \left(\sqrt{\bar{X}_n} + \frac{q^N_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)^2\right]$$