*Université Paris I, Panthéon - Sorbonne*

**Master M.A.E.F.**

# Statistiques

## Plan du cours

1. Quelques rappels la théorie de la mesure.
2. Quelques rappels sur les applications de la théorie de la mesure aux probabilités.
3. Estimation paramétrique.
4. Tests paramétriques.

## Bibliographie

• **Livres pour revoir les bases....**

1. Baillargeon, B. *Probabilités, statistiques et techniques de régression*. SMG.
2. Bercu, B., Pamphile, P. et Azoulay, E. *Probabilités et Applications - Cours Exercices*. Edisciences.
3. Dress, F. *Probabilités et Statistique*. Dunod.
4. Lecoutre, J.-P. *Statistiques et Probabilités*. Dunod.

**Théorie de la mesure et applications aux probabilités**

— Ansel et Ducel, *Exercices corrigés en théorie de la mesure et de l'intégration*, Ellipses.
— Barbe, P. et Ledoux, M., *Probabilités*, Belin.
— Dacunha-Castelle, D. et Duflo, M., *Probabilités et Statistiques (I)*, Masson
— Jacod, J., *Cours d'intégration*, http ://www.proba.jussieu.fr/pageperso/jacod.html.
— Jacod, J., *Cours de Probabilités*, http ://www.proba.jussieu.fr/pageperso/jacod.html.
— Toulouse, P. *Thèmes de probabilités et statistiques*, Masson.

**Statistiques inférentielles**

— Dacunha-Castelle, D. et Duflo, M., *Probabilités et Statistiques (I)*, Masson.
— Fourdrinier, D., *Statistique inférentielle*, Dunod.
— Lecoutre, J.-M. et Tassi, P., *Statistique non paramétrique et robustesse*, Economica.
— Milhaud, X., *Statistique*, Belin.
— Monfort, A., *Cours de statistique mathématique*, Economica.
— Saporta, G., *Probabilités, analyse des données et statistiques*. Technip.
— Tsybakov, A. *Introduction à la statistique non-paramétrique*. Collection : Mathématiques et Applications, Springer.
— van der Vaart, A.W. *Asymptotic statistics* Cambridge series in statistical and probabilistics mathematics, Cambridge University Press.

# Cours de STATISTIQUES 1

## 1 Rappels de probabilités

### 1.1 Quelques rappels d'intégration

#### 1.1.1 Espaces $\mathcal{L}^p$

**Définition 1.** *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré. On appelle espace $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, où $p > 0$, l'ensemble des fonctions $f : \Omega \mapsto \mathbb{R}$, mesurables et telles que $\int |f|^p d\mu < +\infty$.*

**Définition 2.** *Pour $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, où $p > 0$, on note $\| f \|_p = \left( \int |f|^p d\mu \right)^{1/p}$.*

**Propriété 1** (Inégalité de Hölder)**.** *Soit $p > 1$ et $q > 1$ tels que $\frac{1}{p} + \frac{1}{q} = 1$, et soit $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ et $g \in \mathcal{L}^q(\Omega, \mathcal{A}, \mu)$. Alors, $f\,g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ et*

$$\| f\,g \|_1 \leq \| f \|_p \,.\, \| g \|_q \,.$$

**Propriété 2** (Inégalité de Minkowski)**.** *Soit $p > 1$ et soit $f$ et $g \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$. Alors, $f + g \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ et*

$$\| f + g \|_p \leq \| f \|_p + \| g \|_p \,.$$

**Remarque 1.** *Pour $p > 1$, $\| \,.\, \|_p$ définie ainsi sur une semi-norme sur $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$. Pour obtenir une norme, il faut se place dans l'espace $\mathbb{L}^p(\Omega, \mathcal{A}, \mu)$ obtenu en "quotientant" $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ par la relation d'équivalence $f = g$ $\mu$-presque partout (c'est-à-dire que dans $\mathbb{L}^p(\Omega, \mathcal{A}, \mu)$ on dira que $f = g$ lorsque $f = g$ $\mu$-presque partout).*

**Définition 3.** *Pour $f$ et $g \in \mathbb{L}^2(\Omega, \mathcal{A}, \mu)$, on définit le produit scalaire $< f, g >= \int f.g\, d\mu$. On muni ainsi $\mathbb{L}^2(\Omega, \mathcal{A}, \mu)$ d'une structure d'espace de Hilbert. On dira que $f$ est orthogonale à $g$ lorsque $< f, g >= 0$.*

**Conséquence 1.** *Si $A$ est un sous-espace vectoriel fermé de $\mathbb{L}^2(\Omega, \mathcal{A}, \mu)$ (par exemple un sous-espace de dimension finie), alors pour tout $f \in \mathbb{L}^2(\Omega, \mathcal{A}, \mu)$, il existe un unique projeté orthogonal de $f$ sur $A$, noté $f_A$, qui vérifie $f_A = Arg\inf_{g \in A} \| g - f \|_2$.*

#### 1.1.2 Théorèmes fondamentaux

**Théorème 1** (Théorème de convergence monotone (Beppo-Lévi))**.** *Si $(f_n)_n$ est une suite croissante de fonctions mesurables positives convergeant simplement vers $f$ sur $\Omega$, alors :*

$$\lim_{n \to \infty} \left( \int f_n d\mu \right) = \int f \, d\mu = \int \lim_{n \to \infty} f_n d\mu.$$

**Conséquence 2.** *Pour les séries de fonctions mesurables positives, on peut toujours appliquer le Théorème de convergence monotone et donc inverser la somme et l'intégrale.*

**Lemme 1** (Lemme de Fatou)**.** *Soit $(f_n)_n$ est une suite de fonctions mesurables positives alors :*

$$\int \left( \liminf_{n \to \infty} f_n \right) d\mu \leq \liminf_{n \to \infty} \int f_n d\mu.$$

**Exemple 1.** *Appliquer Fatou à $(f_n)$ telle que $f_{2n} = \mathbb{I}_A$ et $f_{2n+1} = \mathbb{I}_B$.*

**Théorème 2** (Théorème de Fubini)**.** *Soit $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2$ et $\mu = \mu_1 \otimes \mu_2$ (mesures $\sigma$ finies), où $(\Omega_1, \mathcal{A}_1, \mu_1)$ et $(\Omega_2, \mathcal{A}_2, \mu_2)$ sont des espaces mesurés. Soit une fonction $f : \Omega \mapsto \mathbb{R}$, $\mathcal{A}$-mesurable et $\mu$-intégrable. alors :*

$$\int_\Omega f d\mu = \int_{\Omega_1} \left( \int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2) \right) d\mu_1(\omega_1) = \int_{\Omega_2} \left( \int_{\Omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1) \right) d\mu_2(\omega_2).$$

**Théorème 3** (Théorème de convergence dominée de Lebesgue)**.** *Soit $(f_n)_n$ est une suite de fonctions de $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ telles que pour tout $n \in \mathbb{N}$, $|f_n| \leq g$ avec $g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$. Si on suppose que $(f_n)$ converge simplement vers $f$ sur $\Omega$ alors :*

$$\lim_{n\to\infty} \int f_n d\mu = \int f \, d\mu.$$

**Extension 1.** *Le Théorème de Lebesgue s'applique également dans le cas où $(f_n)_n$ converge presque partout vers $f$.*

**Exemple 2.** *Convergence d'intégrale dépendant d'un paramètre : par exemple* $\displaystyle\int_0^\infty \frac{f(x)}{1+x^n} dx.$

**Théorème 4** (Inégalité de Jensen)**.** *Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, soit $\phi : \mathbb{R} \mapsto \mathbb{R}$ une fonction convexe et soit $f : \Omega \mapsto \mathbb{R}$ mesurable telle que $\phi(f)$ soit une fonction intégrable par rapport à $P$. Alors :*

$$\phi\left(\int f \, d\mathbb{P}\right) \leq \int \phi(f) \, d\mathbb{P}.$$

**Exemple 3.** *Soit $X$ une v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$. Alors $\phi(\mathbb{E}X) \leq \mathbb{E}(\phi(X))$.*

## 1.2 Espérance de variables aléatoires

**Définition 4.** *Soit $X$ une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Alors si $X \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, on définit l'espérance de $X$ par le nombre $\mathbb{E}X = \int X d\mathbb{P}$. Plus généralement, si $\phi : \mathbb{R} \mapsto \mathbb{R}$ est borélienne et si $\phi(X) \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, on définit l'espérance de $\phi(X)$ par $\mathbb{E}\,\phi(X) = \int \phi(X) d\mathbb{P}$.*

**Propriété 3.** *Si $X$ est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, si $\phi : \mathbb{R} \mapsto \mathbb{R}$ est borélienne telle que $\phi(X) \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, et si $\mathbb{P}_X$ est la mesure de probabilité de $X$ alors :*

$$\mathbb{E}\,\phi(X) = \int_{\mathbb{R}} \phi(x) \, d\mathbb{P}_X(x).$$

**Conséquence 3.**
— *Si $\mathbb{P}_X$ est absolument continue par rapport à la mesure de Lebesgue (donc $X$ est une v.a. dite absolument continue), de densité $f_X$, alors* $\mathbb{E}\,\phi(X) = \displaystyle\int_{\mathbb{R}} \phi(x) f_X(x) dx.$
— *Si $\mathbb{P}_X$ est absolument continue par rapport à la mesure de comptage sur $\mathbb{N}$ (donc $X$ est une v.a. dite discrète), de densité $p_X$, alors* $\mathbb{E}\,\phi(X) = \displaystyle\sum_{k=0}^{\infty} p_X(k)\,\phi(k).$

**Propriété 4.**
1. *Soit $X$ et $Y$ des variables aléatoires telles que $X$ et $Y \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$. Alors pour tout $(a, b) \in \mathbb{R}^2$, $aX + bY \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ et*
$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y.$$
2. *Soit $X$ une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, et soit $A \in \mathcal{A}$. Alors $\mathbb{E}(\mathbb{I}_A(X)) = \mathbb{P}(X \in A)$.*
3. *Soit $X$ et $Y$ des variables aléatoires telles que $X \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ et $Y \in \mathbb{L}^q(\Omega, \mathcal{A}, \mathbb{P})$ avec $\dfrac{1}{p} + \dfrac{1}{q} = 1$ et $p > 1$, $q > 1$. Alors $X.Y \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ et*
$$\mathbb{E}|X.Y| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}.$$
4. *Soit $X$ et $Y$ des variables aléatoires telles que $X$ et $Y \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$, avec $p \geq 1$. Alors $X + Y \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ et*
$$(\mathbb{E}|X + Y|^p)^{1/p} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p}.$$
5. *Soit $X$ une variable aléatoire telle que $X \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ pour $p > 0$. Alors pour tout $0 < r \leq p$, $X \in \mathbb{L}^r(\Omega, \mathcal{A}, \mathbb{P})$ et*
$$(\mathbb{E}|X|^r)^{1/r} \leq (\mathbb{E}|X|^p)^{1/p}.$$

6. *Si $X$ est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, si $\phi : \mathbb{R} \mapsto \mathbb{R}$ est une fonction borélienne convexe telle que $X$ et $\phi(X) \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, alors*

$$\mathbb{E}(\phi(X)) \geq \phi(\mathbb{E}X).$$

**Définition 5.** *Pour $X$ et $Y$ des variables aléatoires telles que $X$ et $Y \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, on définit la covariance de $X$ et $Y$ par*

$$cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right];$$

*On appelle variance de $X$, $var(X) = cov(X, X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.*

**Propriété 5.** *Sur $\mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, $cov(.,.)$ définit un produit scalaire. De plus*

$$|cov(X, Y)|^2 \leq var(X).var(Y).$$

## 1.3 Fonction de répartition et quantiles d'une loi de probabilité

Il y a une correspondance bijective entre la connaissance de $\mathbb{P}_X$ et celle de $F_X = \mathbb{P}_X(]-\infty, x])$. La fonction de répartition permet également de définir les quantiles qui sont essentiels à la construction d'intervalles de confiance et de test.

Soit $\alpha \in [0, 1]$. Des propriétés de la fonction de répartition, on en déduit qu'il existe $x_\alpha \in \mathbb{R}$, tel que :

$$\lim_{x \to x_\alpha} F_X(x) \leq \alpha \leq F_X(x_\alpha). \tag{1}$$

Soit $I_\alpha = \{x_\alpha \in \mathbb{R} \text{ tel que } x_\alpha \text{ vérifie (1)}\}$. On appelle **quantile** (ou fractile, ou percentile en anglais) d'ordre $\alpha$ de la loi $\mathbb{P}_X$, noté $q_\alpha$, le milieu de l'intervalle $I_\alpha$. Evidemment, lorsque $X$ admet une distribution absolument continue par rapport à la mesure de Lebesgue, $q_\alpha = F_X^{-1}(\alpha)$, où $F_X^{-1}$ désigne la fonction réciproque de $F_X$.

Deux cas particuliers sont à connaître :

1/ pour $\alpha = 0.5$, $q_{0.5}$ est appelé la **médiane** de $\mathbb{P}_X$ ;

2/ pour $\alpha = 0.25$ et $\alpha = 0.75$ (respectivement), $q_{0.25}$ et $q_{0.25}$ sont appelés premier et troisième **quartile** (respectivement) de $\mathbb{P}_X$.

3/ pour $\alpha = 0.1, \ldots, 0.9$, on parlera de **décile** de $\mathbb{P}_X$.

## 1.4 Principales lois de probabilités

**Loi uniforme discrète :**

C'est la loi de probabilité discrète à valeurs dans $\{x_1, \ldots, x_n\}$ telle que

$$\mathbb{P}(X = x_i) = \frac{1}{n}.$$

On alors : $\mathbb{E}X = \dfrac{1}{n}(x_1 + \ldots + x_n)$ et $\text{var}(X) = \dfrac{1}{n}(x_1^2 + \ldots + x_n^2) - (\mathbb{E}X)^2$.

**Loi de Bernoulli :**

C'est la loi de probabilité discrète notée $\mathcal{B}(p)$ à valeurs dans $\{0, 1\}$ telle que

$$\mathbb{P}(X = 1) = p \quad \text{et} \quad \mathbb{P}(X = 0) = 1 - p.$$

On alors : $\mathbb{E}X = p$ et $\text{var}(X) = p(1 - p)$.

**Loi binomiale :**

C'est la loi de probabilité discrète notée $\mathcal{B}(n, p)$ à valeurs dans $\{0, 1, \ldots, n\}$ telle que

$$\mathbb{P}(X = k) = C_n^k \cdot p^k \cdot (1 - p)^{n-k} \quad \text{pour } k \in \{0, 1, \ldots, n\}.$$

On alors : $X = X_1 + \cdots + X_n$, où $(X_i)$ est une suite de v.a.i.i.d. de loi $\mathcal{B}(p)$, d'où $\mathbb{E}X = n \cdot p$ et $\mathrm{var}(X) = n \cdot p(1-p)$.

**Loi de Poisson :**

C'est la loi de probabilité discrète notée $\mathcal{P}(\theta)$ à valeurs dans $\mathbb{N}$ telle que

$$\mathbb{P}(X = k) = \frac{\theta^k}{k!} \cdot e^{-\theta} \quad \text{pour } k \in \mathbb{N}.$$

On alors $\mathbb{E}X = \theta$ et $\mathrm{var}(X) = \theta$.

**Loi uniforme sur $[a, b]$ :**

Cette loi est généralement notée $\mathcal{U}([a, b])$, où $-\infty < a < b < \infty$. C'est la loi de probabilité à valeurs dans $[a, b]$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{b-a} \cdot \mathbb{I}_{x \in [a,b]}.$$

On a alors $\mathbb{E}X = \dfrac{b+a}{2}$ et $\mathrm{var}(X) = \dfrac{(b-a)^2}{12}$.

**Loi Gamma :**

Cette loi est généralement notée $\gamma(p, \theta)$, où $p > 0$ et $\theta > 0$. C'est la loi de probabilité à valeurs dans $\mathbb{R}_+$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{\theta^p}{\Gamma(p)} \cdot e^{-\theta \cdot x} \cdot x^{p-1} \cdot \mathbb{I}_{x \in \mathbb{R}_+}.$$

On a alors $\mathbb{E}X = \dfrac{p}{\theta}$ et $\mathrm{var}(X) = \dfrac{p}{\theta^2}$.
Si $X \sim \gamma(p, \theta)$ et $Y \sim \gamma(q, \theta)$ avec $X$ et $Y$ indépendantes et $p > 0$ et $q > 0$, alors $X + Y \sim \gamma(p + q, \theta)$.
Pour $p = 1$, la loi $\gamma(p, \theta)$ est la loi exponentielle $\mathcal{E}(\theta)$.

**Loi Béta :**

Cette loi est généralement notée $\beta(p, \theta)$, où $p > 0$ et $q > 0$. C'est la loi de probabilité à valeurs dans $[0, 1]$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{x^p(1-x)^{q-1}}{B(p,q)} \cdot x^{p-1} \cdot \mathbb{I}_{x \in [0,1]}, \quad \text{où} \quad B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

On a alors $\mathbb{E}X = \dfrac{B(p+1,q)}{B(p,q)}$ et $\mathrm{var}(X) = \dfrac{p \cdot q}{(p+q)^2(p+q+1)}$.
Si $X \sim \gamma(p, \theta)$ et $Y \sim \gamma(q, \theta)$ avec $X$ et $Y$ indépendantes et $p > 0$ et $q > 0$, alors $\dfrac{X}{X+Y} \sim \beta(p, q)$.
Pour $p = 1$, la loi $\gamma(p, \theta)$ est la loi exponentielle $\mathcal{E}(\theta)$.

**Loi normale (ou gaussienne) centrée réduite :**

Cette loi est généralement notée $\mathcal{N}(0, 1)$. C'est la loi de probabilité à valeurs dans $\mathbb{R}$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

On a :

$$\mathbb{E}(X) = 0 \quad \text{et} \quad \mathrm{var}(X) = 1.$$

**Loi normale (ou gaussienne) de moyenne $m$ et de variance $\sigma^2$ :**

Si $Z$ suit la loi $\mathcal{N}(0,1)$, $X = m + \sigma Z$ suit par définition la loi $\mathcal{N}(m, \sigma^2)$, loi normale d'espérance $m$ et de variance $\sigma^2$. La densité de $X$ est donnée par :

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

La figure A.1. représente la densité de la loi normale centrée réduite et celle d'une loi normale non centrée et non réduite. A partir de la loi gaussienne, on peut en déduire les lois suivantes.

**Loi du $\chi^2$ à $n$ degrés de libertés :**

Soit $X_1, \cdots, X_n$, $n$ variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$, alors

$$S = X_1^2 + \cdots + X_n^2$$

suit une loi du $\chi^2$ à $n$ degrés de libertés, loi notée $\chi^2(n)$. Cette loi est à valeurs dans $\mathbb{R}_+$, d'espérance $n$ et de variance $2n$. C'est aussi la loi Gamma $\gamma(n/2, 1/2)$, c'est-à-dire que $X \sim \chi^2(n)$ admet pour densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{2^{n/2} \cdot \Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{x}{2}\right) \cdot \mathbb{I}_{\{x \geq 0\}},$$

où la fonction Gamma est telle que $\Gamma(a) = \displaystyle\int_0^\infty x^{a-1} \cdot e^{-x}$ pour $a \geq 0$. Enfin, si $X$ suit une loi $\chi^2(n)$, par définition on dira que $Y = \sigma^2 \cdot X$ suit une loi $\sigma^2 \cdot \chi^2(n)$. La figure A.2. exhibe trois tracés différents de densité de loi du $\chi^2$. **Loi de Student à $n$ degrés de libertés :**

La loi de Student à $n$ degrés de liberté, notée $T(n)$, est la loi du quotient

$$T = \frac{N}{\sqrt{S/n}}$$

où $N$ suit une loi $\mathcal{N}(0,1)$ et $S$ suit une loi $\chi^2(n)$, $N$ et $S$ étant deux variables aléatoires indépendantes. Il est également possible de déterminer la densité d'une telle loi par rapport à la mesure de Lebesgue, à savoir,

$$f_X(x) = \frac{1}{\sqrt{n} \cdot B(1/2, n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2},$$

où la fonction Beta est telle que $B(a,b) = \dfrac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}$ pour $a > 0$ et $b > 0$. La figure A.3. illustre deux exemples de cette densité, que l'on compare également avec la densité de la loi normale centrée réduite. *Remarque :* Par la loi des grands nombres, plus $n$ est grand, plus $S$ est proche de son espérance qui vaut $n$. Le dénominateur est donc proche de 1. Il s'ensuit que la loi $T(n)$ est d'autant plus proche d'une loi normale que $n$ est grand.

Un des principaux intérêt de la loi de Student réside dans le fait que si $X_1, \cdots, X_n$ sont $n$ variables aléatoires indépendantes de loi $\mathcal{N}(m, \sigma^2)$, si on considère la moyenne et la variance empiriques :

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \quad \text{et} \quad \bar{\sigma}_n^2 = \frac{1}{n-1}\left((X_1 - \bar{X}_n)^2 + \cdots + (X_d - \bar{X}_n)^2\right),$$

alors

$$T = \frac{\sqrt{n} \cdot (\bar{X}_n - m)}{\sqrt{\bar{\sigma}_n^2}}$$

suit une loi de Student à $(n-1)$ degrés de liberté.

**Loi de Fisher à $n_1$ et $n_2$ degrés de liberté :**

Soit $S_1$ et $S_2$ deux variables aléatoires indépendantes de loi respectives $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors par définition :

$$F = \frac{S_1/n_1}{S_2/n_2}$$

suit une loi de Fisher à $n_1$ et $n_2$ degrés de liberté, notée $F(n_1, n_2)$.

*Remarque :* Par les mêmes considérations que précédemment, la loi $F$ est d'autant plus proche de 1 que les degrés de liberté $n_1$ et $n_2$ sont grands.

On a également les propriétés suivantes :

- Si $F$ suit une loi $F(n_1, n_2)$, alors la loi de $\frac{n_1}{n_2} F$ est une loi beta de seconde espèce de paramètres $(n_1/2, n_2/2)$, c'est-à-dire que $F$ est à valeurs dans $\mathbb{R}_+$ et admet la densité par rapport à la mesure de Lebesgue :
$$f_X(x) = \frac{1}{B(n_1/2, n_2/2)} n_1^{n_1/2} \cdot n_2^{n_2/2} \frac{x^{n_1/2-1}}{(n_2 + n_1 \cdot x)^{(n_1+n_2)/2}} \mathbb{I}_{\{x \geq 0\}},$$
la notation $B$ désignant encore la fonction Beta.
- Si $F \sim F(n_1, n_2)$, alors $\mathbb{E}(F) = \frac{n_2}{n_2 - 2}$ lorsque $n_2 > 2$ et $\text{var}(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2}$ lorsque $n_2 > 4$.
- Si $T$ suit une loi de Student $T(n)$, alors $T^2$ suit une loi de Fisher $F(1, n)$.

La figure A.4. donne une idée de la distribution d'une loi de Fisher pour différents choix des paramètres.

## 1.5 Indépendance

**Définition 6.** *Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé.*

— *Soit $(A_i)_{i \in I}$ une famille dénombrable d'événements de $\mathcal{A}$. On dit que les événements $(A_i)_{i \in I}$ sont indépendants si et seulement si pour tous les sous-ensembles finis $K \subset I$,*

$$\mathbb{P}\left(\bigcap_{i \in K} A_i\right) = \prod_{i \in K} \mathbb{P}(A_i).$$

— *Soit $(\mathcal{A}_i)_{i \in I}$ une famille de sous-tribus de $\mathcal{A}$ (donc pour tout $i \in I$, $\mathcal{A}_i \subset \mathcal{A}$). On dit que les tribus $(\mathcal{A}_i)_{i \in I}$ sont indépendantes si et seulement si pour tous les sous-ensembles finis $K \subset I$, et pour tous les événements $A_k \in \mathcal{A}_k$ avec $k \in K$, les $A_k$ sont indépendants.*

— *Soit $(X_i)_{i \in I}$ des variables aléatoires sur $(\Omega, \mathcal{A})$ à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. On dit que les v.a. $(X_i)_{i \in I}$ sont indépendantes si et seulement si les tribus engendrées $(X_i^{-1}(\mathcal{B}(\mathbb{R})))_{i \in I}$ sont indépendantes.*

**Proposition 1.** *Si $(X_1, \cdots, X_n)$ sont des variables aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$. Alors les $(X_i)$ sont indépendantes si et seulement si $\mathbb{P}_{(X_1, \cdots, X_n)} = \bigotimes_{i=1}^{n} \mathbb{P}_{X_i}$.*

**Proposition 2.** *Si $(X_i)_{i \in I}$ sont des variables aléatoires indépendantes sur $(\Omega, \mathcal{A}, \mathbb{P})$. Alors les $(X_i)$ sont indépendantes si et seulement si pour tout $J \subset I$, $J$ fini, pour toutes fonctions boréliennes $(g_j)_{j \in J}$ telles que $g_j(X_j)$ soit intégrable, alors*

$$\mathbb{E}\left(\prod_{j \in J} g_j(X_j)\right) = \prod_{j \in J} \mathbb{E}(g_j(X_j)).$$

**Corollaire 1.** *$(X_1, \cdots, X_n)$ sont des variables aléatoires indépendantes si et seulement si pour tout $(t_1, \cdots, t_n) \in \mathbb{R}^n$,*

$$\phi_{(X_1, \cdots, X_n)}(t_1, \cdots, t_n) = \prod_{j=1}^{n} \phi_{X_j}(t_j).$$

## 1.6 Vecteurs aléatoires

**Définition 7.** *On dit que $X$ est un vecteur aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, un espace probabilisé, si $X$ est une fonction mesurable de $(\Omega, \mathcal{A})$ dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.*

**Définition 8.** *Soit $X$ un vecteur aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. Alors la loi (ou mesure) de probabilité de $X$, $\mathbb{P}_X$, est définie de façon univoque à partir de la fonction de répartition de $X$, telle que pour $x = (x_1, \cdots, x_d)$,*

$$F_X(x) = \mathbb{P}_X(\prod_{i=1}^{d} ]-\infty, x_i]) = \mathbb{P}(X \in \prod_{i=1}^{d} ]-\infty, x_i]).$$

**Propriété 6.** *Soit $X$ un vecteur aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. On suppose que $X = (X_1, \cdots, X_d)$. Alors les $X_i$ sont des variables aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$, de fonction de répartition*

$$F_{X_i}(x_i) = \lim_{\substack{x_j \to +\infty \\ j \neq i}} F_X(x_1, \cdots, x_i, \cdots, x_d).$$

*Les mesures de probabilités $P_{X_i}$ déterminées de façon univoque à partir des $F_{X_i}$ sont appelées lois marginales de $X$.*

On se place maintenant dans la base canonique orthonormale de $\mathbb{R}^d$. Si $Z$ est un vecteur aléatoire à valeurs sur $\mathbb{R}^d$, on définit $\mathbb{E}(Z)$, le vecteur dont les coordonnées sont les espérances des coordonnées de $Z$. Ainsi, si dans la base canonique de $\mathbb{R}^d$, $Z = (Z_1, \cdots, Z_d)'$,

$$\mathbb{E}(Z) = \mathbb{E}\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}(Z_1) \\ \vdots \\ \mathbb{E}(Z_d) \end{pmatrix}.$$

De la même manière, on définira l'espérance d'une matrice dont les coordonnées sont des variables aléatoires par la matrice dont les coordonnées sont les espérances de chacune de ces variables aléatoires.

Ceci nous permet de définir la matrice de variance-covariance de $Z$ de la manière suivante :

$$\mathrm{var}(Z) = \mathbb{E}\left[(Z - \mathbb{E}(Z)).(Z - \mathbb{E}(Z))'\right]$$

donc si $Z = (Z_1, \cdots, Z_d)'$,

$$\mathrm{var}\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} \mathrm{var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_d) \\ \mathrm{Cov}(Z_1, Z_2) & \mathrm{var}(Z_2) & \cdots & \mathrm{Cov}(Z_2, Z_d) \\ \vdots & \vdots & \cdots & \vdots \\ \mathrm{Cov}(Z_1, Z_d) & \mathrm{Cov}(Z_2, Z_d) & \cdots & \mathrm{var}(Z_d) \end{pmatrix}$$

matrice $(d, d)$ dont les éléments diagonaux sont les variances et les éléments non diagonaux sont les covariances des coordonnées de $Z$ (remarquons que la variance de $Z_1$ est aussi la covariance de $Z_1$ et de $Z_1$).

On vérifie également le résultat suivant : si $C$ est une matrice $(p, d)$ à coordonnées constituées de réels constants et si $Z$ est un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, alors $C \cdot Z$ est un vecteur de taille $p$ de matrice de variance-covariance

$$\mathrm{var}(C \cdot Z) = C \cdot \mathrm{var}(Z) \cdot C'.$$

En particulier, si $p$ vaut 1, alors $C = h'$ où $h$ est un vecteur de taille $d$, et :

$$\mathrm{var}(h' \cdot Z) = h' \cdot \mathrm{var}(Z) \cdot h.$$

Notez que cette dernière quantité est un scalaire. Soit $Y_1, \cdots, Y_d$ des variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$, indépendantes (ce qui, dans le cas gaussien, est équivalent à $\mathrm{cov}(Y_i, Y_j) = 0$ pour $i \neq j$). On considère le vecteur $Y = (Y_1, \cdots, Y_d)'$. En raison de l'indépendance, $Y$ est un vecteur gaussien admettant

une densité $f_Y$ (par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$) qui est le produit des densités de chacune des coordonnées, soit :

$$\begin{aligned} f_Y(y_1,\cdots,y_d) &= f_{Y_1}(y_1)\times f_{Y_2}(y_2)\times\cdots\times f_{Y_d}(y_d) \\ &= \left(2\pi\sigma^2\right)^{-d/2}\exp\left(-\frac{1}{2\sigma^2}(y_1^2+\cdots+y_d^2)\right) \\ &= \left(2\pi\sigma^2\right)^{-d/2}\exp\left(-\frac{\|y\|^2}{2\sigma^2}\right), \end{aligned}$$

avec $y=(y_1,\cdots,y_d)$. On voit donc que la densité de $Y$ ne dépend que de la norme $\|Y\|$ : elle est constante sur toutes les sphères centrées en zéro. Cela implique qu'elle est invariante par rotation ou symétrie orthogonale d'axe passant par 0 : elle est invariante par toutes les isométries de $\mathbb{R}^d$ : on dira que $Y$ suit une loi gaussienne isotrope. Rappelons que les isométries correspondent à des changements de bases orthonormées (BON). En conséquence, on a la première propriété importante :

**Propriété 7.** *Soit $Y$ un vecteur aléatoire de $\mathbb{R}^d$ de loi normale isotrope variance $\sigma^2$, c'est-à-dire que dans une BON les coordonnées de $Y$ vérifient $\mathbb{E}(Y)=0$ et $var(Y)=\sigma^2\cdot Id$. Alors les coordonnées de $Y$ dans toute BON sont encore des lois $\mathcal{N}(0,\sigma^2)$ indépendantes.*

Voici maintenant l'un des résultats (encore appelé Théorème de Cochran) que nous utilisons le plus et nous en donnons donc une démonstration.

**Théorème 5** (Théorème de Cochran)**.** *Soit $E_1$ et $E_2$, deux sous-espaces vectoriels orthogonaux de $E=\mathbb{R}^d$ de dimensions respectives $k_1$ et $k_2$ et soit $Y$ un vecteur aléatoire de $\mathbb{R}^d$ de loi normale centrée isotrope de variance $\sigma^2$. Alors $P_{E_1}(Y)$ et $P_{E_2}(Y)$ sont deux variables aléatoires gaussienne centrées indépendantes et $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2\cdot\chi^2(k_1)$ (resp. $\sigma^2\cdot\chi^2(k_2)$). Ce théorème se généralise naturellement pour $2<m\leq d$ sous-espaces vectoriels orthogonaux $(E_i)_{1\leq i\leq m}$ de $E=\mathbb{R}^d$.*

*Démonstration :* Soit $(e_1,\cdots,e_{k_1})$ et $(e_{k_1+1},\cdots,e_{k_1+k_2})$ deux BON de $E_1$ et $E_2$ (respectivement). L'ensemble de ces deux bases peut être complété en

$$(e_1,\cdots,e_{k_1},e_{k_1+1},\cdots,e_{k_1+k_2},e_{k_1+k_2+1},\cdots,e_d)$$

pour former une BON de $\mathbb{R}^d$ (du fait que $E_1$ et $E_2$ sont orthogonaux).

Soit $(Y_1,\cdots,Y_d)$, les coordonnées de $Y$ dans cette base ; elles sont indépendantes de loi $\mathcal{N}(0,\sigma^2)$ car le changement de base est orthonormal et nous avons vu que la distribution de $Y$ était conservé par transformation isométrique. Comme

$$\begin{aligned} P_{E_1}(Y)=Y_1e_1+\cdots+Y_{k_1}e_{k_1} &\implies \|P_{E_1}(Y)\|^2=\sigma^2\left(\left(\frac{Y_1}{\sigma}\right)^2+\cdots+\left(\frac{Y_{k_1}}{\sigma}\right)^2\right) \\ P_{E_2}(Y)=Y_{k_1+1}e_{k_1+1}+\cdots+Y_{k_1+k_2}e_{k_1+k_2} &\implies \|P_{E_2}(Y)\|^2=\sigma^2\left(\left(\frac{Y_{k_1+1}}{\sigma}\right)^2+\cdots+\left(\frac{Y_{k_1+k_2}}{\sigma}\right)^2\right). \end{aligned}$$

On voit bien ainsi l'indépendance entre les deux projections et le fait que la loi de $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2\cdot\chi^2(k_1)$ (resp. $\sigma^2\cdot\chi^2(k_2)$). ■

On peut définir plus généralement un vecteur gaussien $Y$ à valeurs dans $\mathbb{R}^d$ (non dégénéré), d'espérance $\mu\in\mathbb{R}^d$ et de matrice de variance-covariance $\Sigma$ quelconques (du moment que $\Sigma$ soit une matrice de Toeplitz définie positive). Cela équivaut à définir un vecteur aléatoire de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$,

$$f_Y(y)=\frac{(2\pi)^{-n/2}}{|\Sigma|}\exp\left(-\frac{1}{2}(y-\mu)'.\Sigma^{-1}.(y-\mu)\right),$$

pour $y\in\mathbb{R}^d$, et avec $|\Sigma|$ le déterminant de la matrice $\Sigma$. Remarquons une nouvelle fois que l'espérance et la variance définissent complètement la loi de probabilité d'un vecteur gaussien.

A partir des propriétés générales sur les vecteurs aléatoires, on obtient le fait que :

**Propriété 8.** *Soit $Y$ un vecteur gaussien à valeurs dans $\mathbb{R}^d$ (non dégénéré), d'espérance $\mu \in \mathbb{R}^d$ et de matrice de variance-covariance $\Sigma$. Soit $C$ une matrice réelle de taille $(p, d)$ où $p \in \mathbb{N}^*$. Alors $C \cdot Y$ est un vecteur gaussien tel que :*

$$C \cdot Y \sim \mathcal{N}(C \cdot \mu\,,\, C \cdot \Sigma \cdot C')$$

On en déduit les conséquences suivantes :

- si $Y$ est un vecteur gaussien isotrope de $\mathbb{R}^d$ de variance $\sigma^2$ et $h$ un vecteur de $\mathbb{R}^d$, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de $Y$ tel que :

$$h' \cdot Y \text{ suit la loi } \mathcal{N}(0, \sigma^2 \cdot h' \cdot h) = \mathcal{N}(0, \sigma^2 \cdot \|h\|^2)$$

- si $Y$ est un vecteur gaussien d'espérance $\mu$ et de matrice de variance $\Sigma$ et si $h$ un vecteur de $\mathbb{R}^d$, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de $Y$ et :

$$h' \cdot Y \text{ suit la loi unidimensionnelle } \mathcal{N}(h' \cdot \mu, h' \cdot \Sigma \cdot h)$$

(Pour une présentation plus détaillée des notions sur les vecteurs gaussiens on peut consulter le livre P. Toulouse, 1999, chap.2)

## 1.7 Fonctions caractéristiques

**Définition 9.** *Soit $X$ un vecteur aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. La fonction caractéristique de $X$ est la fonction $\phi_X : \mathbb{R}^d \mapsto \mathbb{C}$ telle que*

$$\phi_X(t) = \mathbb{E}\left[\exp(i < t, X >)\right] = \int_{\mathbb{R}^d} e^{i<t,x>} d\mathbb{P}_X(x),$$

*où $< . >$ désigne le produit scalaire euclidien sur $\mathbb{R}^d$ tel que $< t, x >= \sum_{i=1}^{d} t_i x_i$ pour $t = (t_1, \cdots, t_d)$ et $x = (x_1, \cdots, x_d)$.*

**Remarque 2.** *La fonction caractéristiques existe sur $\mathbb{R}$ et $\phi_X(0) = 1$. $\phi_X$ est aussi la transformée de Fourier de la mesure $\mathbb{P}_X$.*

**Théorème 6.** *Soit $X$ et $Y$ des vecteurs aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$, de lois $\mathbb{P}_X$ et $\mathbb{P}_Y$. Alors $\mathbb{P}_X = \mathbb{P}_Y$ si et seulement si $\phi_X = \phi_Y$.*

**Théorème 7** (Théorème d'inversion)**.** *Si $X$ est un vecteur aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$ et si $\phi_X$ est une fonction intégrable par rapport à la mesure de Lebesgue $\lambda_d$ sur $\mathbb{R}^d$, alors $X$ admet une densité $f_X$ par rapport à $\lambda_d$ telle que pour $x \in \mathbb{R}^d$,*

$$f_X(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i<t,x>} \phi_X(t) dt.$$

**Proposition 3.** *Si $X$ est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ de fonction caractéristiques $\phi_X$. Alors si $\mathbb{E}(|X|^n) < +\infty$ (ou $X \in \mathbb{L}^n(\Omega, \mathcal{A}, \mathbb{P})$), $\phi_X$ est $n$ fois dérivable et $\phi_X^{(n)}(t) = i^n \mathbb{E}(X^n e^{itX})$.*

**Remarque 3.** *Lorsque ces moments existent, on a $i^n \mathbb{E}(X^n) = \phi_X^{(n)}(0)$.*

## 1.8 Convergence de suites de variables aléatoires

**Lemme 2** (Lemme de Borel-Cantelli)**.** *Soit $(A_n)_{n\in\mathbb{N}}$ une suite d'événements sur $(\Omega, \mathcal{A}, \mathbb{P})$.*

1. *Si $\sum \mathbb{P}(A_n) < +\infty$ alors $\mathbb{P}(\limsup A_n) = 0$.*
2. *Si les $(A_n)$ sont indépendants, $\sum \mathbb{P}(A_n) = +\infty$ implique que $P(\limsup A_n) = 1$.*

**Définition 10.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$. On dit que*

— *$(X_n)$ converge en probabilité vers $X$, noté $X_n \xrightarrow[n\to+\infty]{\mathcal{P}} X$, lorsque pour tout $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

— $(X_n)$ *converge dans* $\mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ *vers* $X$, *noté* $X_n \xrightarrow[n\to+\infty]{\mathbb{L}^p} X$, *avec* $p > 0$, *lorsque*

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0.$$

— $(X_n)$ *converge en loi vers* $X$, *noté* $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$, *lorsque, pour toute fonction* $\phi$ *continue bornée,*

$$\lim_{n\to\infty} \mathbb{E}\phi(X_n) = \mathbb{E}\phi(X).$$

— $(X_n)$ *converge presque sûrement vers* $X$, *noté* $X_n \xrightarrow[n\to+\infty]{p.s.} X$, *lorsque* $\exists E \in \Omega$ *avec* $P(E) = 1$ *tel que* $\forall \omega \in E$, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$
$\Leftrightarrow$ *pour tout* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\sup_{m\geq n} |X_m - X| > \varepsilon) = 0.$$

**Propriété 9.** *1. p.s. et* $\mathbb{L}^p \longrightarrow \mathcal{P} \longrightarrow \mathcal{L}$.

*2. pour* $q \geq p$, $\mathbb{L}^q \longrightarrow \mathbb{L}^p$.

*3. La convergence en loi n'entraîne pas la convergence en probabilité. Mais* $(X_n \xrightarrow[n\to+\infty]{\mathcal{P}} C) \Longleftrightarrow (X_n \xrightarrow[n\to+\infty]{\mathcal{L}} C)$ *pour* $C$ *une constante.*

*4. Si* $g$ *est une fonction borélienne continue alors* $(X_n \xrightarrow[n\to+\infty]{\mathcal{P}} X) \Longrightarrow (g(X_n) \xrightarrow[n\to+\infty]{\mathcal{P}} g(X))$.

**Propriété 10.** *1. Si pour tout* $\varepsilon > 0$, $\sum_{n=0}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < +\infty$ *alors* $X_n \xrightarrow[n\to+\infty]{p.s.} X$ *(application du Lemme de Borel-Cantelli).*

*2. Si il existe* $r > 0$ *tel que* $\mathbb{E}(|X_n|^r) < +\infty$ *et* $\sum_{n=0}^{\infty} \mathbb{E}(|X_n - X|^r) < +\infty$ *alors* $X_n \xrightarrow[n\to+\infty]{p.s.} X$.

Le lemme du porte-manteau donne des descriptions équivalentes de la convergence en loi.

**Lemme 3** (Porte-manteau)**.** *Pour une suite de variable aléatoires* $(X_n)_{n\in\mathbb{N}}$ *et un variable aléatoire* $X$ *à valeur dans un même esace métrique* $(E, d)$, *les proriétés suivantes sont équivalentes :*

1. $X_n$ *converge en loi vers* $X$.
2. $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *pour tout ensemble fermé* $F$.
3. $\liminf \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$ *pour tout ensemble ouvert* $G$.
4. $\mathbb{P}(X_n \in B) \to \mathbb{P}(X \in B)$ *pour tout ensemble borélien* $B$ *avec* $\mathbb{P}(X \in \delta B) = 0$, *où* $\delta B = \bar{B} - \mathring{B}$.
5. $\mathbb{E}f(X_n) \to f(X)$ *pour toute fonction bornée et Lipschitz (c'est-à-dire telle qu'il existe* $L$ *avec* $|f(x) - f(y)| \leq Ld(x - y)$ *pour tout* $x, y$.*)*

**Preuve** (1) $\to$ (2). Soit $F$ un fermé de $E$. On note

$$F_k = \left\{ x \in \mathbb{R}^d | d(x, F) \leq \frac{1}{k} \right\}.$$

De plus, pour tout $x \in E$ et $k > 0$, on pose

$$\phi_k(x) = f\left(kd(x, F)\right) = 1 - kd(x, F) \text{ si } x \in F_k, \text{ et } 0 \text{ sinon.}$$

Par définition $\mathbb{I}_F \leq \phi_k \leq \mathbb{I}_{F_k}$, et

$$\limsup_n \mathbb{P}\left(X_n \in F\right) \leq \limsup_n \mathbb{E}\phi_k(X_n).$$

Comme $\phi_k$ est continue et bornée, on aura par (1) :

$$\limsup_n \mathbb{E}\phi_k(X_n) = \lim_n \mathbb{E}\phi_k(X_n) = \mathbb{E}\phi_k(X).$$

Comme $\phi_k \leq \mathbb{I}_{F_k}$, on aura $\mathbb{E}\phi_k(X) \leq P(X \in F_k)$, par conséquent, pour tout $k \geq 1$,

$$\limsup_n \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F_k).$$

Enfin, comme $(F_k)_{k\in\mathbb{N}^*}$ est décroissante et $\cap_{k\geq 1} F_k = \bar{F} = F$, on a

$$\inf_k \mathbb{P}(X \in F_k) = \lim_k \mathbb{P}(X \in F_k) = P(X \in F)$$

et $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$. Montrons maintenant que (2) et (3) sont équivalents, supposons que (2) soit vrai et considérons un ouvert $\mathcal{O}$ de $E$. Alors, $\mathcal{O}^C$ est un fermé, donc

$$\begin{aligned}&\liminf_n \mathbb{P}(X_n \in \mathcal{O}) = \liminf_n (1 - \mathbb{P}(X_n \in \mathcal{O}^c))\\ &= 1 - \limsup_n \mathbb{P}(X_n \in \mathcal{O}^c) \geq 1 - Pro(X \in \mathcal{O}^c) = \mathbb{P}(X \in \mathcal{O}).\end{aligned}$$

La démonstration de (3) implique (2) est identique. Montrons maintenant que (2) et (3) implique (4). Soit $A$ un borélien de $E$ tel que $\mathbb{P}(X \in \delta A) = 0$. Comme

$$\mathring{A} \subset A \subset \bar{A} = \mathring{A} \cup \delta A,$$

ainsi $\mathbb{P}(X \in \bar{A}) = \mathbb{P}(X \in A) = \mathbb{P}(X \in \mathring{A})$. D'après (2),

$$\limsup_n \mathbb{P}(X_n \in A) \leq \limsup_n \mathbb{P}(X_n \in \bar{A}) \leq \mathbb{P}(X \in \bar{A})$$

et d'après (3)

$$\liminf_n \mathbb{P}(X_n \in A) \geq \liminf_n \mathbb{P}(X_n \in \mathring{A}) \leq \mathbb{P}(X \in \mathring{A}).$$

Finalement,

$$\lim_n \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A).$$

Montrons maintenant que (4) entraîne (1), soit $\phi$ une fonction continue et borné, telle que $0 < \phi < 1$, alors :

— Comme $\phi > 0$, on montrera en TD que

$$\mathbb{E}\phi(X) = \int_0^1 \mathbb{P}(\phi(X) > x)dx \text{ et } \mathbb{E}\phi(X_n) = \int_0^1 \mathbb{P}(\phi(X_n) > x)dx$$

— Comme $\phi$ est continue, $\delta\phi^{-1}(]x,\infty[) \subset \phi^{-1}(\{x\})$ donc $\mathbb{P}(X \in \delta\phi^{-1}(]x,\infty[)) = 0$, sauf pour un nombre au plus dénombrable $D_0$ de valeurs de $x$, en effet, dans le cas contraire on aurait $P(X \in E) = \infty$. De même, pour $n \geq 1$, $\mathbb{P}(X_n \in \delta\phi^{-1}(]x,\infty[)) = 0$, sauf pour un nombre au plus dénombrable $D_n$ de valeurs de $x$. L'ensemble $\cup_{n\in\mathbb{N}} D_n$ est dénombrable et donc Lebesgue-négligeable. Par conséquent, par (4), pour presque tout $x \in [0,1]$ :

$$\mathbb{P}(\phi(X_n) > x) \to \mathbb{P}(\phi(X) > x)$$

et par convergence dominée :

$$\mathbb{E}\phi(X_n) \to \mathbb{E}\phi(X).$$

Enfin, pour le cas général, si $\phi$ est une fonction continue bornée $a < \phi < b$, on se ramène au cas précédent en posant $\phi^* = \frac{\phi - a}{b-a}$ et on conclut par linéarité de l'intégrale. Comme toute fonction Lipschitz et bornée est continue et borné, (1) entraîne (5), pour terminer la preuve il suffit de montrer que (5) entraîne (3). Pour tout ensemble ouvert $G$, il existe une suite croissante de fonctions Lipschitz telles que $0 \leq f_m \to \mathbb{I}_G$. Il suffit de prendre la fonction $f_m(x) = \min(md(x, G^c), 1)$ qui est $m$ Lipschitz. Pour tout $m$ fixé, on aura

$$\lim_{n\to\infty} \inf \mathbb{P}(X_n \in G) \geq \lim_{n\to\infty} \inf \mathbb{E}f_m(X_n) = \mathbb{E}f_m(X).$$

Finalement, quand $m \to \infty$, $\mathbb{E}f_m(X) \to \mathbb{P}(X \in G)$ par convergence monotone.

**Conséquence de ce lemme** Une première conséquence de ce lemme est la proposition suivante :

**Proposition 4.** *On note $x = (x_1, \cdots, x_d)^T$ un vecteur de $\mathbb{R}^d$, $F_n$ la fonction de répartition de $X_n$ et $F$ celle de $X$. La suite de vecteurs aléatoires $(X_n)_{n\in\mathbb{N}}$ converge en loi vers $X$ si et seulement si*

$$\lim_n F_n(x) = F(x),$$

*en tout point de continuité de $x \mapsto F(x)$.*

**Preuve** Montrons d'abord que si $X_n$ converge en loi vers $X$ alors $\lim_n F_n(x) = F(x)$ en tout point de continuité de $x \mapsto F(x)$. On a $\lim_n F_n(x) = F(x) \Leftrightarrow \lim_n \mathbb{P}(X_n \in \prod_{i=1}^d ]-\infty, x_i]) = \mathbb{P}(X \in \prod_{i=1}^d ]-\infty, x])$. Comme $\delta \prod_{i=1}^d ]-\infty, x_i] = \{x\}$, on aura $\mathbb{P}(X \in \delta \prod_{i=1}^d ]-\infty, x_i]) = P(X = x) = F(x) - F(x_-)$ qui est nul si et seulement si $F$ est continue à gauche, donc continue en $x$. Réciproquement, l'ensemble de points de discontinuité de $F$, $\mathcal{D}$, est au plus dénombrable, sinon on aurait $P(X \in \mathcal{D}) = \infty$ ce qui est impossible. Ainsi $E = \mathbb{R}^d \setminus \mathcal{D}$ est dense dans $\mathbb{R}^d$ et la mesure de Lebesgue de $\mathcal{D}$ est nulle. Si $x \mapsto f(x)$ est continue sur $\mathbb{R}^d$ et notons $I = \prod_{i=1}^d [a_i, b_i]$ un rectangle compact de $\mathbb{R}^d$ donc les sommets $\{(a_i, b_i), i \in 1, \cdots, d\}$ sont tous dans $E$. Pour tout $\varepsilon > 0$, il existe un rectangle compact $I$ tel que $Pro(X \notin I) < \varepsilon$. Comme toute fonction $f$ continue sur un compact est uniformément continue sur ce compact, il existe une partition finie de rectangles $\{I_j, 1 \leq j \leq J\}$ ($I_j$ éventuellement ouvert aux extrémités) tels que $I = \cup_j I_j$, avec tous les sommet de $I_j$ dans $E$ et telle que $f$ varie d'au plus $\varepsilon$ sur chaque $I_j$. On choisit un point $x_j$ dans chaque $I_j$ et on définit $f_\varepsilon = \sum_j f(x_j) \mathbb{I}_{I_j}$. On aura alors pour tout $x \in I$, $|f_\varepsilon(x) - f(x)| < \varepsilon$ et

$$|\mathbb{E}f(X_n) - \mathbb{E}f_\varepsilon(X_n)| \leq \varepsilon + \mathbb{P}(X_n \notin I),$$
$$|\mathbb{E}f(X) - \mathbb{E}f_\varepsilon(X)| \leq \varepsilon + \mathbb{P}(X \notin I) < 2\varepsilon$$

Si $n$ est suffisamment grand on aura de plus $\mathbb{P}(X_n \notin I) < \varepsilon$ et comme les sommet des $I_j$ sont dans $E$ :

$$\lim_n |\mathbb{E}f_\varepsilon(X_n) - \mathbb{E}f_\varepsilon(X)| \leq \lim_n \sum_j |\mathbb{P}(X_n \in I_j) - \mathbb{P}(X \in I_j)| \, |f(x_j)| = 0,$$

ainsi que

$$|\mathbb{E}f(X_n) - \mathbb{E}f(X)| = |\mathbb{E}f(X_n) - \mathbb{E}f_\varepsilon(X_n) + \mathbb{E}f_\varepsilon(X_n) - \mathbb{E}f_\varepsilon(X) + \mathbb{E}f_\varepsilon(X) - \mathbb{E}f(X)| \leq$$
$$|\mathbb{E}f(X_n) - \mathbb{E}f_\varepsilon(X_n)| + |\mathbb{E}f_\varepsilon(X_n) - \mathbb{E}f_\varepsilon(X)| + |\mathbb{E}f_\varepsilon(X) - \mathbb{E}f(X)|,$$

on aura, pour $n$ suffisamment grand,

$$|\mathbb{E}f(X_n) - \mathbb{E}f(X)| < 5\varepsilon.$$

Comme c'est vrai pour tout $\varepsilon > 0$, on aura bien $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ pour toute fonction continue bornée et $X_n \xrightarrow{\mathcal{L}} X$.

Si la variable $X$ a une fonction de répartition continue alors la convergence de $F_n$ vers $F$ est même uniforme en $x$.

**Lemme 4.** *Si $X_n \xrightarrow{\mathcal{L}} X$ et la fonction de répartition $F$ de $X$ est continue alors en notant $F_n$ la fonction de répartition de $X_n$ :*

$$\sup_x |F_n(x) - F(x)| \to 0.$$

**Preuve** On commence par le cas unidimensionel. Fixons $k \in \mathbb{N}^*$, par la continuité de $F$, il existe $-\infty = x_0 < x_1 < \cdots < x_k = \infty$ avec $F(x_i) = \frac{i}{k}$, pour $x_{i-1} \leq x \leq x_i$ :

$$F_n(x) - F(x) \leq F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + \tfrac{1}{k}$$
$$\geq F_n(x_i) - F_(x_i) \geq F_n(x_{i-1}) - F_(x_i) = F_n(x_{i-1}) - F_(x_{i-1}) - \tfrac{1}{k}$$

Ainsi, $\forall k \in \mathbb{N}^*$,

$$\sup_x |F_n(x) - F(x)| \leq \sup_i |F_n(x_i) - F(x_i)| + \frac{1}{k} \overset{n\to\infty}{\longrightarrow} \frac{1}{k}.$$

Donc, pour tout $k \in \mathbb{N}^*$ $\lim_n \sup_x |F_n(x) - F(x)| \leq \frac{1}{k}$ ce qui montre que $\lim_n \sup_x |F_n(x) - F(x)| = 0$.

Pour le cas multidimensionnel, il suffit de prendre une partition finie de $\mathbb{R}^d$ en hyper-rectangles

$$\left(-\infty = x_0^{(i)} < x_1^{(i)} < \cdots < x_{k^{(i)}}^{(i)} = \infty\right)_{1 \leq i \leq d}$$

telle que pour $x \in \mathbb{R}^d$ il existe deux sommets d'hyper-rectangles $x_i$ et $x_{i-1}$ avec $x_{i-1} \leq x \leq x_i$ tels que $F(x_i) - F(x_{i-1}) = \frac{1}{k}$ et de répéter l'argument précédent .

Une autre conséquence du lemme du porte-manteau est une partie de la proposition suivante :

**Proposition 5** (Lien entre les convergences)**.** *Soit $X_n$, $Y_n$ et $X$ des vecteurs aléatoires, alors :*

1. $X_n \xrightarrow{p.s.} X$ *implique* $X_n \xrightarrow{P} X$.
2. *Si* $X_n \xrightarrow{\mathcal{L}} X$ *et* $d(X_n, Y_n) \xrightarrow{P} 0$, *alors* $Y_n \xrightarrow{\mathcal{L}} X$.
3. $X_n \xrightarrow{P} X$ *implique* $X_n \xrightarrow{\mathcal{L}} X$.
4. $X_n \xrightarrow{P} c$, *où* $c$ *est une constante, si et seulement si* $X_n \xrightarrow{\mathcal{L}} c$.
5. *Si* $X_n \xrightarrow{\mathcal{L}} X$ *et* $Y_n \xrightarrow{P} c$, *où* $c$ *est une constante, alors* $(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, c)$.
6. *Si* $X_n \xrightarrow{P} X$ *et* $Y_n \xrightarrow{P} Y$, *alors* $(X_n, Y_n) \xrightarrow{P} (X, Y)$.

**Preuve**

1. La suite d'ensemble $A_n = \cup_{m \geq n} \{d(X_m, X) > \varepsilon\}$ est décroissante pour tout $\varepsilon > 0$ et décroît vers un ensemble de mesure nulle si $X_n \xrightarrow{p.s.} X$. On aura donc $\mathbb{P}(d(X_n, X) > \varepsilon) \leq P(A_n) \to 0$.
2. Pour toute fonction $f$ à valeur dans $[a, b]$, $L$-Lpischitz et $\varepsilon > 0$,

$$|\mathbb{E}f(X_n) - \mathbb{E}f(Y_n)| \leq \varepsilon \mathbb{E}\mathbb{I}_{d(X_n,Y_n)\leq\varepsilon} + 2L(b-a)\mathbb{E}\mathbb{I}_{d(X_n,Y_n)>\varepsilon}.$$

   Par hypothèse, le second terme de droite de l'inégalité converge vers 0, le premier terme de droite est de taille $\varepsilon$ qui est arbitrairement petit, ainsi $|\mathbb{E}f(X_n) - \mathbb{E}f(Y_n)| \to 0$. Le résultat est donc une conséquence du lemme du porte-manteau.
3. Comme $d(X_n, X) \longrightarrow 0$ et $X \xrightarrow{\mathcal{L}} X$, cette propriété découle de la propriété 2).
4. D'après la propriété précédente, il suffit de montrer que si $X_n \xrightarrow{\mathcal{L}} c$ alors $X_n \xrightarrow{P} c$. Soit une boule ouverte autour de $c$ et de rayon $\varepsilon$ : $B(c, \varepsilon) = \{x, d(x, c) < \varepsilon\}$. On aura $\mathbb{P}(d(X_n, c) \geq \varepsilon) = P(X_n \in B(c, \varepsilon)^c$. Si $X_n \xrightarrow{\mathcal{L}} c$ alors par le lemme du porte-manteau $\limsup_n P(X_n \in B(c, \varepsilon)^c) \leq P(c \in B(c, \varepsilon)^c) = 0$ et $\mathbb{P}(d(X_n, c) \geq \varepsilon) \to 0$.
5. On remarque d'abord que $d(X_n, Y_n), (X_n, c)) = d(Y_n, c) \xrightarrow{P} 0$. D'après 2), il suffit donc de montrer que $(X_n, c) \xrightarrow{\mathcal{L}} (X, c)$, mais pour toute fonction continue bornée $(x, y) \mapsto f(x, y)$, la fonction $x \mapsto f(x, c)$ est continue bornée, ainsi $\mathbb{E}f(X_n, c) \to \mathbb{E}f(X, c)$ si $X_n \xrightarrow{\mathcal{L}} 0$.
6. Cette propriété est la conséquence de $d((x_1, y_1), (x_2, y_2)) \leq d(x_1, x_2) + d(y_1, y_2)$.

Finalement, une autre conséquence du lemme du porte-manteau sera le théorème de l'application continue ci-après.

Les 5 théorèmes suivants sont des classiques de la théorie des probabilités et on renvoit donc au cours de probabilité pour leurs preuves.

**Théorème 8** (Loi faible des Grands Nombres)**.** *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes et identiquement distribuées. Alors si* $\mathbb{E}(|X_i|) < +\infty$,

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} m = \mathbb{E}X_i.$$

**Théorème 9** (Loi forte des Grands Nombres)**.** *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes et identiquement distribuées. Alors si* $\mathbb{E}(|X_i|) < +\infty$,

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \underset{n\to+\infty}{\xrightarrow{p.s.}} m = \mathbb{E}X_i.$$

**Théorème 10** (Théorème de la limite centrale)**.** *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes et identiquement distribuées. Alors si* $\sigma^2 = \mathbb{E}X_i^2 < +\infty$, *et* $m = \mathbb{E}X_i$,

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, 1).$$

**Théorème 11** (Loi forte des Grands Nombres multidimensionnelle)**.** *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de vecteurs aléatoires à valeurs dans* $\mathbb{R}^d$, *indépendants et identiquement distribués. Alors si* $\mathbb{E}(\|X_i\|) < +\infty$ *(pour* $\|.\|$ *une norme sur* $\mathbb{R}^d$*),*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \underset{n\to+\infty}{\xrightarrow{p.s.}} m = \mathbb{E}X_i.$$

**Théorème 12** (Théorème de la limite centrale multidimensionnel). *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, indépendants et identiquement distribués. Alors si $\Sigma$ matrice de covariance de chaque $X_i$ existe, et $m = \mathbb{E}X_i$,*

$$\sqrt{n}\Big(\overline{X}_n - m\Big) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_d(0,\Sigma).$$

**Théorème 13** (Application continue). *Soit $g$ une fonction presque-sûrement continue de $\mathbb{R}^k$ dans $\mathbb{R}^m$ alors :*

1. *Si $X_n \xrightarrow{\mathcal{L}} X$, alors $g(X_n) \xrightarrow{\mathcal{L}} g(X)$.*
2. *Si $X_n \xrightarrow{P} X$, alors $g(X_n) \xrightarrow{P} g(X)$.*
3. *Si $X_n \xrightarrow{p.s.} X$, alors $g(X_n) \xrightarrow{p.s.} g(X)$.*

**Preuve**

1. Il suffit de prouver que pour toute fonction $f$ continue borné $\mathbb{E}f(g(X_n)) \to \mathbb{E}f(g(X))$. Comme $h = f \circ g$ est continue bornée, c'est vrai car $X_n \xrightarrow{\mathcal{L}} X$.
2. Notons $D_g$ l'ensemble des points de discontinuité de $g$, comme $g$ est une fonction presque-sûrement continue on aura $\mathbb{P}(X \in D_g) = 0$. Pour $\varepsilon > 0$, on définit $B_\delta$ comme :

   $$B_\delta = \Big\{x \in \mathbb{R}^k | x \notin D_g : \exists y \in \mathbb{R}^k : \|x-y\| < \delta \text{ et } \|g(x)-g(y)\| > \varepsilon\Big\}$$

   Comme $g$ est presque-sûrement continue, $\lim_{\delta\to 0} B_\delta = \emptyset$. On aura

   $$\mathbb{P}(|g(X_n)-g(X)| > \varepsilon) \le \mathbb{P}(|X_n - X| \ge \delta) + \mathbb{P}(X \in B_\delta) + \mathbb{P}(X \in D_g) \overset{n\to\infty,\delta\to 0}{\longrightarrow} 0.$$

   Ainsi, $\lim_n \mathbb{P}(|g(X_n)-g(X)| > \varepsilon) = 0$ et $g(X_n) \xrightarrow{P} g(X)$.
3. Si $g$ est continue en $X(\omega)$, alors

   $$\lim_n X_n(\omega) = X(\omega) \Rightarrow \lim_n g(X_n(\omega)) = g(X(\omega)).$$

   Ainsi

   $$\begin{aligned}&\mathbb{P}\left(\lim_n g(X_n(\omega)) = g(X(\omega))\right) \ge \mathbb{P}\left(\lim_n g(X_n(\omega)) = g(X(\omega)), X \notin D_g\right)\\ &\ge \mathbb{P}\left(\lim_n X_n(\omega) = X(\omega), X \notin D_g\right) = 1\end{aligned}$$

   et $g(X_n) \xrightarrow{p.s.} g(X)$.

En combinant ce théorème avec la proposition 5 sur les liens entre les convergences , on obtient le lemme de Slutsky :

**Lemme 5** (Slutsky). *Si $X_n$, $Y_n$ sont des vecteurs (ou des matrices) aléatoires, avec des dimensions compatibles avec les propriétés suivantes. Si $X_n \xrightarrow{\mathcal{L}} X$ et $Y_n \xrightarrow{\mathcal{L}} c$, alors :*

1. *$X_n + Y_n \xrightarrow{\mathcal{L}} X + c$.*
2. *$Y_n X_n \xrightarrow{\mathcal{L}} cX$.*
3. *$Y_n^{-1} X_n \xrightarrow{\mathcal{L}} c^{-1}X$, si $c$ est inversible.*

## 1.9 Symboles $o$ et $O$ stochastiques

Une famille de vecteurs $\{X_\alpha, \alpha \in A\}$ est dîte uniformément tendue si pour tout $\varepsilon > 0$, il existe $M$ tel que

$$\sup_{\alpha\in A} \mathbb{P}(\|X_\alpha\| > M) < \varepsilon.$$

Il est pratique d'avoir une notation pour les termes qui convergent vers 0 en probabilité, où qui sont bornés en probabilité. La notation $(X_n)_{n\in\mathbb{N}} = o_P(1)$ signifie que $X_n \xrightarrow{P} 0$ et la notation $(X_n)_{n\in\mathbb{N}} = O_P(1)$ signifie que $(X_n)_{n\in\mathbb{N}}$ est uniformément tendue. Plus généralement, pour une suite aléatoire $(R_n)_{n\in\mathbb{N}}$ on notera :

$$\begin{aligned}&X_n = o_P(R_n) \text{ si } X_n = Y_n R_n \text{ et } Y_n = o_P(1).\\ &X_n = O_P(R_n) \text{ si } X_n = Y_n R_n \text{ et } Y_n = O_P(1).\end{aligned}$$

On montre facilement, en revenant à la définition, les règles de calculs suivantes :

— $o_P(1) + o_P(1) = o_P(1)$
— $o_P(1) + O_P(1) = O_P(1)$
— $O_P(1)o_P(1) = o_P(1)$
— $(1 + o_P(1))^{-1} = O_P(1)$
— $o_P(R_n) = R_n o_P(1)$
— $O_P(R_n) = R_n O_P(1)$
— $o_P(O_P(1)) = o_P(1)$

On aura de plus le lemme suivant :

**Lemme 6** (Comparaison stochastique)**.** *Soit un fonction $R$ définie sur un domaine dans $\mathbb{R}^d$ telle que $R(0) = 0$. Soit $(X_n)_{n\in\mathbb{N}}$ une suite aléatoire dans le domaine de $R$ qui converge en probabilité vers 0. Alors pour tout $p > 0$,*

1. *Si $R(h) = o(\|h\|^p)$, quand $h \to 0$, alors $R(X_n) = o_P(\|X_n\|^p)$.*
2. *Si $R(h) = O(\|h\|^p)$, quand $h \to 0$, alors $R(X_n) = O_P(\|X_n\|^p)$.*

**Preuve** Soit $g(h)$ telle que $g(h) = \frac{R(h)}{\|h\|^p}$ pour $h \neq 0$ et $g(0) = 0$, alors $R(X_n) = g(X_n)\|X_n\|^p$.

1. Par hypothèse $g$ est continue en 0 donc $g(X_n) \xrightarrow{P} 0$ par le théorème de l'application continue.
2. Par hypothèse, il existe $M$ et $\delta > 0$ tel que $|g(h)| \leq M$ quelque soit $h$ avec $\|h\| \leq \delta$, donc $\mathbb{P}(|g(X_n)| > M) \leq \mathbb{P}(\|X_n\| > \delta) \to 0$ et la suite $g(X_n)$ est uniformément tendue.

## 1.10 Delta-méthode

### 1.10.1 Résultat basique

La delta-méthode consiste à utiliser un développement de Taylor pour approximer un vecteur aléatoire de la forme $\phi(T_n)$, quand la fonction $\phi$ est différentiable et $T_n \xrightarrow{P} \theta$, on peut alors déduire la loi limite de $\phi(T_n) - \phi(\theta)$. Une conséquence immédiate du théorème de l'application continue est que si $T_n \xrightarrow{P} \theta$, alors $\phi(T_n) \xrightarrow{P} \phi(\theta)$. De plus si l'application $\phi$ est différentiable, de différentielle $\phi'_\theta$, on aura :

$$\sqrt{n}\left(\phi(T_n) - \phi(\theta)\right) \simeq \phi'_\theta\left(\sqrt{n}(T_n - \theta)\right)$$

et si $\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} T$ pour une variable aléatoire $T$, on s'attend à avoir $\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} \phi'_\theta(T)$. En général, les estimateurs statistiques $T_n$ sont vectoriels : $T_n = (T_{n,1}, \cdots, T_{n,d})$, si on suppose de plus que $\phi$ est une fonction de $\mathbb{R}^d \to \mathbb{R}^m$ définie, au moins, dans un voisinage de $\theta$. On dit que $\phi$ est différentiable en $\theta$ si il existe une application linéaire $\phi'_\theta : \mathbb{R}^d \to \mathbb{R}^m$ telle que :

$$\phi(\theta + h) - \phi(\theta) = \phi'_\theta(h) + o(\|h\|),\ h \to 0.$$

Si $\phi$ est differentiable et $\theta = (\theta_1, \cdots, \theta_d)$, alors on peut représenter cette différentielle par la matrice :

$$\phi'_\theta = \begin{pmatrix} \frac{\partial\phi_1(\theta)}{\partial\theta_1} & \cdots & \frac{\partial\phi_1(\theta)}{\partial\theta_d} \\ \vdots & & \vdots \\ \frac{\partial\phi_m(\theta)}{\partial\theta_1} & \cdots & \frac{\partial\phi_m(\theta)}{\partial\theta_d} \end{pmatrix}$$

Si l'application $\theta \mapsto \phi'(\theta)$ est continue (en identifiant la norme de l'application linéaire à la norme de la matrice qui la représente) on dira alors que $\phi$ est continuement différentiable. Il est préférable de voir la différentielle comme une approximation linéaire $h \mapsto \phi'_\theta(h)$ de la fonction $h \mapsto \phi(\theta + h) - \phi(\theta)$. Si l'image de $\phi$ est réelle, (et sa différentielle est un vecteur ligne) la différentielle de $\phi$ est appelée "gradient". Remarquons que la dérivée d'une fonction réelle $\phi : \mathbb{R} \to \mathbb{R}$, ne correspond pas à la différentielle qui est une application linéaire est non pas un nombre. Ici $\theta \mapsto \phi'_\theta$ est une fonction de $\mathbb{R}$ dans l'ensemble des applications linéaires de $\mathbb{R} \to \mathbb{R}$. Graphiquement, l'approximation affine $h \mapsto \phi(\theta) + \phi'_\theta(h)$ est la tangente de la fonction $\phi$ en $\theta$.

**Théorème 14** (Delta-méthode)**.** *Soit $\phi : \mathcal{D}_\phi \subset \mathbb{R}^d \to \mathbb{R}^m$ une application définie sur un sous-emsemble de $\mathbb{R}^d$ et différentiable en $\theta$. Soit $T_n$ un vecteur aléatoire à valeur dans le domaine de définition $\mathcal{D}_\phi$ de $\phi$. Si $r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$ pour une suite $r_n \to \infty$, alors $r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} \phi'_\theta(T)$. De plus*

$$r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) \xrightarrow{P} 0.$$

**Preuve** Comme $r_n(T_n-\theta) \xrightarrow{\mathcal{L}} T$, elle est uniformément tendue et $T_n-\theta \xrightarrow{P} 0$. Comme $\phi$ est différentiable en $\theta$, $R(h) = \phi(\theta+h) - \phi(\theta) - \phi'_\theta(h)$ satisfait $R(h) = o(\|h\|)$ quand $h \to 0$ et le lemme de comparaison stochastique de la section précédente permet d'en déduire que

$$\phi(T_n) - \phi(\theta) - -\phi'_\theta(T_n-\theta) = R(T_n-\theta) = o_P(\|T_n-\theta\|).$$

En multipliant cette égalité par $r_n$ et en remarquant de $o_P(r_n\|T_n-\theta\|) = o_P(1)$ car $r_n(T_n-\theta)$ est uniformément tendue, on en déduit que

$$r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n-\theta)) \xrightarrow{P} 0.$$

Comme une application linéaire est continue, on en déduit que

$$\phi'_\theta(r_n(T_n-\theta)) \xrightarrow{\mathcal{L}} \phi'_\theta(T).$$

Par la proposition 5 sur le lien entre les différentes convergence, on en déduit que $r_n(\phi(T_n) - \phi(\theta))$, a la même limite $\phi'_\theta(T)$.

Le corollaire ci-après est une conséquence immédiate de ce théorème :

**Corollaire 2.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, indépendants et identiquement distribués, telle que $\Sigma$ matrice de covariance de chaque $X_i$ existe, et $m = \mathbb{E}X_i$. Soit $g : \mathbb{R}^d \to \mathbb{R}^p$ une fonction de classe $\mathcal{C}^1$ sur un voisinage autour de $m$, de matrice Jacobienne $J_g(m)$ en $m$. Alors,*

$$\sqrt{n}\Big(g(\overline{X}_n) - g(m)\Big) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_d(0, J_g(m)\cdot\Sigma\cdot J'_g(m)).$$

**Exemple (Skewness)** Pour un échantillon i.i.d. $(Z_1,\cdots,Z_n)$ on note $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. La Skewness d'une loi de variance $\sigma^2$ est le quotient $\lambda = \frac{\mathbb{E}\left((X-\mathbb{E}(X))^3\right)}{(\sigma^2)^{\frac{3}{2}}}$. L'estimateur empirique de la Skewness est donc

$$S_n = \frac{\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2\right)^{3/2}}$$

Si la loi a un moment d'ordre 3 fini, $S_n$ converge en probabilité vers sa Skewness. La Skewness d'une distribution symmetrique, comme la loi normale, vaut 0, et on peut utiliser la Sknewness empirique pour tester cet aspect de la normalité. Pour $n$ grand, une valeur critique pour tester cette hypothèse peut être déduite de la loi asymptotique de $S_n$. $S_n$ peut être écrite $\phi(\overline{X},\overline{X^2},\overline{X^3})$ avec $\phi(a,b,c) = \frac{c-3ab+2a^3}{(b-a^2)^{3/2}}$. Si on note $\alpha_k = \mathbb{E}(X^k)$, la suite $\sqrt{n}(\overline{X}-\alpha_1, \overline{X^2}-\alpha_2, \overline{X^3}-\alpha_3)$ est de moyenne nulle et asymptotiquement normale dès que $\mathbb{E}(X^6) < \infty$. La valeur $\phi(\alpha_1,\alpha_2,\alpha_3)$ est exactement la Sknewness de la loi. La fonction $\phi$ est différentiable au point $(\alpha_1,\alpha_2,\alpha_3)$ et on peut appliquer la delta-méthode. On peut simplifier les calculs en remarquant que la Skewness ne dépend pas de la moyenne et de la variance de la loi. Ainsi, si on pose $Y_i = \frac{X_i-\alpha_1}{\sigma}$, l'estimateur de la Skewness peut aussi s'écrire $\phi(\overline{Y},\overline{Y^2},\overline{Y^3})$. Si on note $\mu_k = \mathbb{E}(X_i - \mathbb{E}(X_i))^k)$ le moment central de $X_i$, $\lambda = \frac{\mu_3}{\sigma^3}$ et $\kappa = Var(Y^2) - 2$, on peut montrer que

$$\sqrt{n}\begin{pmatrix}\overline{Y}\\ \overline{Y^2}-1\\ \overline{Y^3}-\lambda\end{pmatrix} \xrightarrow{\mathcal{L}} T = \begin{pmatrix}T_1\\T_2\\T_3\end{pmatrix} = \mathcal{N}\left(0, \begin{pmatrix}1 & \lambda & \kappa+3\\ \lambda & \kappa+2 & \frac{\mu_5}{\sigma^5}-\lambda\\ \kappa+3 & \frac{\mu_5}{\sigma^5}-\lambda & \frac{\mu_6}{\sigma^6}-\lambda^2\end{pmatrix}\right).$$

La dérivée de $\phi$ au point $(0,1,\lambda)$ vaudra $(-3,-\frac{3\lambda}{2},1)$. Ainsi $\sqrt{n}(S_n-\lambda) \xrightarrow{\mathcal{L}} (-3T_1 - \frac{3\lambda T_2}{2} + T_3)$. Si la loi de $X$ est normale, alors $\lambda = \mu_5 = 0$, $\kappa = 0$ et $\frac{\mu_6}{\sigma^6} = 15$ et $\sqrt{n}(S_n-\lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0,6)$.

**Exemple : stabilisation de la variance** Si une statistique $T_n$ est telle que $\sqrt{n}(T_n-\theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2(\theta))$, l'intervalle de confiance de niveau $1-2\alpha$ est donné par $[T_n - C_\alpha\frac{\sigma(\theta)}{\sqrt{n}}, T_n + C_\alpha\frac{\sigma(\theta)}{\sqrt{n}}]$, avec $\mathbb{P}(|Z| > C_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$. Ces intervalles de confiance dépendent du paramètre inconnu $\theta$. Une solution est de remplacer $\sigma(\theta))$ par une estimation $\hat{\sigma}(\theta) \xrightarrow{P} \sigma(\theta)$, pour obtenir un intervalle de confiance de niveau asymptotique

$1-2\alpha$. Une autre approche est d'utiliser une transformation pour stabiliser la variance. L'idée est est de choisir une transformation bijective croissante $\phi$ telle que $\phi'_\theta \sigma(\theta) = 1$, ainsi on aura

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

et on en déduira l'intervalle de confiance de niveau $1-2\alpha$ :

$$\left[\phi^{-1}\left(\phi(T_n) - \frac{C_\alpha}{\sqrt{n}}\right), \phi^{-1}\left(\phi(T_n) + \frac{C_\alpha}{\sqrt{n}}\right)\right].$$

Il suffit donc de choisir $\phi(\theta) = \int \frac{1}{\sigma(\theta)} d\theta$.

Par exemple, Si on a un échantillon i.i.d. de loi normale bivariée $(X_1, Y_1), \cdots, (X_n, Y_n)$, avec pour coefficient de corrélation entre $X$ et $Y$ : $\rho$, alors le coefficient de corrélation empirique sera

$$r_n = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

Avec la delta-méthode, il est possible de calculer la loi asymptotique de $\sqrt{n}(r_n - \rho)$. Sous l'hypothèse de normalité, On peut montrer, après de longs calculs ennuyeux, que $\sqrt{n}(r_n - \rho) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1-\rho^2)^2)$. Donc les bornes de l'intervalle de confiance dépendent de la variance inconnue $(1-\rho^2)^2$. La transformation

$$\phi(\rho) = \int \frac{1}{1-\rho^2} d\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho} = \operatorname{arctanh}(\rho)$$

stabilisera la variance. Ainsi, on aura $\sqrt{n}(\operatorname{arctanh}(r_n) - \operatorname{arctanh}(\rho)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ et l'intervalle de confiance de niveau $1-2\alpha$ sera :

$$\left[\tanh\left(\operatorname{arctanh}(T_n) - \frac{C_\alpha}{\sqrt{n}}\right), \tanh\left(\operatorname{arctanh}(T_n) + \frac{C_\alpha}{\sqrt{n}}\right)\right].$$

### 1.10.2 Développement d'ordre supérieur

La delta-méthode est basée sur un developpement de Taylor d'ordre 1, il est possible d'avoir besoin d'ordre plus grand pour pouvoir conclure. C'est la cas si le premier terme du développement est négligeable, c'est-à-dire quand $\phi'_\theta = 0$. On peut alors espérer que le second terme, quadratique, détermine le comportement limite de $\phi(T_n)$.

**Exemple** Si $\overline{X} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ alors $-2n(\cos(\overline{X}) - 1) \xrightarrow{\mathcal{L}} \chi^2(1)$. En effet

$$\cos(\overline{X}) - \cos(0) = (\overline{X} - 0)(-\sin(0)) + \frac{1}{2}(\overline{X} - 0)^2(-\cos(0))^2 + o(|\overline{X}|)$$

et la suite $n\overline{X}^2$ converge en loi vers un $\chi^2(1)$ par le théorème de l'application continue. Par la proposition 5 sur le lien entre les différentes convergences, on en déduit que $-2n(\cos(\overline{X}) - 1)$ a la même limite.

### 1.10.3 Delta méthode uniforme

Il est possible de prouver la normalité asymptotique de $\sqrt{n}(\phi(Tn) - \phi(\theta_n))$ si $\theta_n \to \theta$. Pour cela il suffit que l'application $\phi$ soit continuement dérivable.

**Théorème 15** (Delta-méthode uniforme)**.** *Soit $\phi : \mathcal{D}_\phi \subset \mathbb{R}^d \to \mathbb{R}^m$ une application définie sur un sous-emsemble de $\mathbb{R}^d$ et continuement différentiable en un voisinage de $\theta$. Soit $T_n$ un vecteur aléatoire à valeur dans le domaine de définition $\mathcal{D}_\phi$ de $\phi$. Si $r_n(T_n - \theta_n) \xrightarrow{\mathcal{L}} T$ pour $\theta_n \to \theta$ et une suite $r_n \to \infty$, alors $r_n(\phi(T_n) - \phi(\theta_n)) \xrightarrow{\mathcal{L}} \phi'_\theta(T)$. De plus*

$$r_n(\phi(T_n) - \phi(\theta_n)) - \phi'_\theta(r_n(T_n - \theta_n)) \xrightarrow{P} 0.$$

**Preuve** Il suffit de prouver la convergence ne probabilité

$$r_n(\phi(T_n) - \phi(\theta_n)) - \phi'_\theta(r_n(T_n - \theta_n)) \xrightarrow{P} 0.$$

la proposition 5 sur le lien entre les différentes convergences permettra de conclure. Comme la convergence en probabilité vers 0 d'un vecteur est équivalente à la convergence en probabilité vers 0 de chacune de ses composantes on peut supposé, sans perte de généralité, que $\phi$ est réelle. Pour $0 \leq t \leq 1$ et $h$ fixé, on définit $g_n(t) = \phi(\theta_n + th)$. Si $n$ est suffisamment grand, et $h$ suffisamment petit, $\theta_n$ et $\theta_n + h$ seront tout deux dans le voisinage de $\theta$ où $\phi$ est continuement différentiable. Alors $t \mapsto g_n(t)$ sera continuement différentiable, avec pour dérivée $g'_n(t) = \phi'_{\theta_n+th}(h)$. Le théorème des valeurs intermédiaires assure qu'il existe $0 \leq \xi \leq$, tel que $g_n(1) - g_n(0) = g'_n(\xi)$. Ainsi

$$R_n(h) = \phi(\theta_n + h) - \phi(\theta_n) - \phi'_\theta(h) = \phi_{\theta_n+\xi h}(h) - \phi'_\theta(h).$$

Par la continuité de l'application $\theta \mapsto \phi'_\theta$, pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que, si $\|\zeta - \theta\| < \delta$, on aura $\|\phi'_\zeta - \phi'_\theta\| < \varepsilon\|h\|$ pour tout $h$. Si $n$ est suffisamment grand et $\|h\| < \frac{\delta}{2}$ alors $\|\theta - (\theta_n + h)\| < \delta$ et $\|R_n(h)\| < \varepsilon\|h\|$. Ainsi pour tout $\eta > 0$,

$$\mathbb{P}\left(r_n\|R_n(T_n - \theta_n)\| > \eta\right) \leq \mathbb{P}\left(\|T_n - \theta_n\| \geq \frac{\delta}{2}\right) + \mathbb{P}\left(r_n\|T_n - \theta_n\|\varepsilon > \eta\right).$$

Le premier terme de droite converge vers 0 quand $n \to \infty$ , de plus le deuxième terme est arbitrairement petit si $\varepsilon$ est choisi suffisamment petit et $\mathbb{P}\left(r_n\|R_n(T_n - \theta_n)\| > \eta\right) \to 0$.

## 1.11 Espérance conditionnelle

**Définition 11.** *Soit $Y$ une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$. Si $\mathcal{B}$ est une sous-tribu de $\mathcal{A}$ et si $Y \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$. Alors on note $\mathbb{E}(Y \mid \mathcal{B})$ la projection orthogonale de $Y$ sur $\mathbb{L}^2(\Omega, \mathcal{B}, P)$, appelée espérance conditionnelle de $Y$ sachant $\mathcal{B}$. Ainsi :*

$$\mathbb{E}\left|Y - \mathbb{E}(Y \mid \mathcal{B})\right|^2 = \inf_{Z \in \mathbb{L}^2(\Omega, \mathcal{B}, \mathbb{P})} \left\{\mathbb{E}\left|Y - Z\right|^2\right\}.$$

*Par extension, si $Y \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, on définit l'espérance conditionnelle par rapport à $\mathcal{B}$, comme l'unique (p.s.) variable aléatoire, $\mathcal{B}$-mesurable vérifiant p.s. :*

$$\int_B \mathbb{E}(Y \mid \mathcal{B})\, d\mathbb{P} = \int_B Y\, d\mathbb{P}, \quad \text{pour tout } B \in \mathcal{B}.$$

**Définition 12.** *Par convention, si $X$ un vecteur aléatoire à valeurs dans $\mathbb{R}^n$ sur $(\Omega, \mathcal{A}, \mathbb{P})$ et si $Y$ une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, on note $\mathbb{E}(Y \mid X) = \mathbb{E}(Y \mid X^{-1}(\mathcal{B}(\mathbb{R})))$.*

**Propriété 11.**

1. *Lemme de Doob : Pour $Y \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, et $X$ une v.a. de$(\Omega, \mathcal{A}, \mathbb{P})$, alors p.s. $\mathbb{E}(Y \mid X) = h(X)$, avec $h$ une fonction borélienne.*
2. *Pour $Y_1$ et $Y_2$ deux variables aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$, et $(a, b, c) \in \mathbb{R}^3$, alors*

$$\mathbb{E}(aY_1 + bY_2 + c \mid \mathcal{B}) = a\mathbb{E}(Y_1 \mid \mathcal{B}) + b\mathbb{E}(Y_2 \mid \mathcal{B}) + c.$$

3. *Si $Y_1 \leq Y_2$, alors $\mathbb{E}(Y_1 \mid \mathcal{B}) \leq \mathbb{E}(Y_2 \mid \mathcal{B})$.*
4. *Le Lemme de Fatou, les théorèmes de Beppo-Levi, Lebesgue et Jensen s'appliquent avec l'espérance conditionnelle.*
5. *Si $Y \in \mathbb{L}^2(\Omega, \mathcal{B}, \mathbb{P})$, alors $\mathbb{E}(Y \mid \mathcal{B}) = Y$ ; ainsi $\mathbb{E}(g(X) \mid X) = g(X)$ pour $g$ une fonction mesurable réelle.*
6. *On a $\mathbb{E}(\mathbb{E}(Y \mid \mathcal{B})) = \mathbb{E}Y$.*
7. *Si $Y^{-1}(\mathcal{B}(\mathbb{R}))$ et $\mathcal{B}$ sont indépendantes alors $\mathbb{E}(Y \mid \mathcal{B}) = \mathbb{E}Y$ ; ainsi, si $X$ et $Y$ sont indépendantes, $\mathbb{E}(Y \mid X) = \mathbb{E}Y$.*
8. *Si $(X, Y)$ est un couple de v.a. à valeurs dans $\mathbb{R}^2$ possédant une densité $f_{(X,Y)}$ par rapport à la mesure de Lebesgue, alors si $X$ est intégrable ,*

$$\mathbb{E}(Y \mid X = x) = \frac{\int_{\mathbb{R}} y \cdot f_{(X,Y)}(x, y)\, dy}{\int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy}, \quad \text{pour tout } x \text{ tel que } \int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy > 0.$$

**Proposition 6.** *Si $(Y, X_1, \cdots, X_n)$ est un vecteur gaussien, alors $\mathbb{E}(Y \mid (X_1, \cdots, X_n)) = a_0 + a_1X_1 + \cdots + a_nX_n$, où les $a_i$ sont des réels.*

# 2 Estimation paramétrique

## 2.1 Définitions

Dans toute la suite, on se place sur $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. On considère $(X_n)_{n\mathbb{N}}$ une suite de variable aléatoire, où chaque $X_i$ est définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ et est à valeur dans $\Omega' \subset \mathbb{R}$.

**Définition 13.**
— *On appelle modèle statistique de dimension $n$ un espace $((\Omega')^n, \mathcal{A}'_n, \mu)$, où $\mathcal{A}'_n$ est une tribu sur $(\Omega')^n$ et $\mu$ une mesure de probabilité sur $((\Omega')^n, \mathcal{A}'_n)$.*
— *On appelle échantillon de taille $n$ du modèle statistique $((\Omega')^n, \mathcal{A}'_n, \mu)$ le vecteur aléatoire $(X_1, \dots, X_n)$ distribuée selon la loi $\mu$. Pour $\omega \in \Omega$, $(X_1(\omega), \dots, X_n(\omega))$ vecteur de $\mathbb{R}^n$ est appelé échantillon observé. C'est à partir et sur ce vecteur que le travail statistique s'effectue (en général).*

**Définition 14.** *On appelle :*
— *Modèle statistique paramétrique, une famille de modèle de la forme : $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$.*
— *Modèle statistique semi-paramétrique, une famille de modèle de la forme : $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_{(\theta,f)}, \theta \in \Theta, f \in \mathcal{F})$, où $\Theta \subset \mathbb{R}^p$ et $\mathcal{F}$ n'est pas de dimension finie.*
— *Modèle statistique non-paramétrique, une famille de modèle de la forme : $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_f, f \in \mathcal{F})$, où $\mathcal{F}$ n'est pas de dimension finie.*

**Définition 15.**
— *On dit que le modèle paramétrique : $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, est dominé par une mesure $\mu$ lorsque $\mathbb{P}_\theta$ est absolument continue par rapport à $\mu$ pour tout $\theta \in \Theta$.*
— *On se place dans le cadre d'un modèle paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure $\mu$. Pour $(x_1, \cdots, x_n) \in (\Omega')^n$, la fonction $\theta \in \Theta \mapsto L_\theta(x_1, \cdots, x_n) = \dfrac{d\mathbb{P}_\theta}{d\mu}(x_1, \cdots, x_n)$ est appelée une vraisemblance du modèle statistique.*

**Exemple 4.**
— *Dans le cas où $\mu$ est la mesure de Lebesgue sur $\mathbb{R}^n$, la vraisemblance sera la densité (classique) en $(x_1, \cdots, x_n)$.*
— *Dans le cas où $\mu$ est comptage sur $\mathbb{N}^n$, la vraisemblance sera la probabilité en $(x_1, \cdots, x_n)$.*
— *Attention ! si le support de $\mathbb{P}_\theta$ dépend de $\theta$, la mesure qui domine (ainsi que $\Omega'$ et $\mathcal{A}'_n$) ne peut dépendre de $\theta$ : il ne faut pas oublier de le préciser dans l'expression de la vraisemblance.*

**Définition 16.** *Lorsque l'on dispose d'un échantillon $(X_1, \dots, X_n)$ du modèle statistique $((\Omega')^n, \mathcal{A}'_n, \mu)$, une statistique $\widehat{T}_n$ est une application mesurable de $((\Omega')^n, \mathcal{A}'_n)$ dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, donc un vecteur aléatoire défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeur dans $\mathbb{R}^d$, et telle que :*

$$\widehat{T}_n = h(X_1, \dots, X_n), \qquad \text{où} \quad h : (\Omega')^n \mapsto \mathbb{R}^d \quad \text{est mesurable.}$$

**Exemple 5.** *Estimateur du paramètre d'une loi de Bernoulli.*
*Estimateur de l'espérance et de la variance par la moyenne et la variance empirique.*
*Estimateurs du paramètre $\theta$ d'un $n$-échantillon $(X_1, \cdots, X_n)$ de loi uniforme sur $[0, \theta]$.*
*Test sur la moyenne.*

## 2.2 Statistiques exhaustives

On se place désormais dans le cadre d'une modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure $\mu$.

**Exemple 6.** *1. Soit le modèle statistique paramétrique $\Big([0, \infty[^n, \mathcal{B}([0, \infty[^n), \mathcal{U}([0, \theta])^{\otimes n}, \theta \in ]0, +\infty[\Big)$. On dispose donc d'un $n$-échantillon $(X_1, \cdots, X_n)$ de v.a.i.i.d. suivant une loi uniforme sur $[0, \theta]$. Si on considère $\max\{X_1 \dots, X_n\}$ cela semble suffire pour posséder toute l'information sur $\theta$ que contenait $(X_1, \dots, X_n)$ : on a donc résumé l'"information" sur $\theta$ contenait $(X_1, \dots, X_n)$, un vecteur de taille $n$, par une statistique de taille 1.*
*2. De même, si on considère le modèle statistique paramétrique $\Big(\{0, 1\}^n, \mathcal{P}(\{0, 1\}^n), \mathcal{B}(p)^{\otimes n}, p \in [0, 1]\Big)$ (on dispose donc d'un $n$-échantillon $(X_1, \cdots, X_n)$ de v.a.i.i.d. suivant une loi de Bernoulli de paramètre $p$) alors la statistique $X_1 + \cdots + X_n$ contient toute l'"information" sur $p$ contenue dans l'échantillon $(X_1, \dots, X_n)$.*

*Comment exprimer formellement ce fait qu'une statistique puisse résumer à elle seule toute l'information sur le paramètre ?*

**Définition 17.** *Soit $\widehat{T}$ une statistique du modèle statistique paramétrique dominé à valeurs dans $\mathbb{R}^d$. On dit que $\widehat{T}$ est une statistique exhaustive si pour toute statistique $S$ intégrable (donc dans $\mathbb{L}^1((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta)$) alors $\mathbb{E}_\theta(S \mid \widehat{T})$ ne dépend ($\mathbb{P}_\theta$-presque sûrement) pas de $\theta$.*

**Théorème 16** (Théorème de factorisation de Neyman)**.** *Soit $(X_1, \cdots, X_n)$ un $n$-échantillon et soit $\widehat{T}$ une statistique du modèle statistique paramétrique dominé avec $\widehat{T}$ à valeurs dans $\mathbb{R}^d$, où $d \in \mathbb{N}^*$. La statistique $\widehat{T}$ est exhaustive si et seulement s'il existe une fonction $h : \mathbb{R}^n \to \mathbb{R}_+$ et une fonction $g_\theta(.) : \mathbb{R}^d \mapsto \mathbb{R}_+$, telle que l'on puisse écrire pour tout $(x_1, \ldots, x_n) \in (\Omega')^n$ :*

$$L_\theta(x_1, \ldots, x_n) = g_\theta(\widehat{T}(x_1, \ldots, x_n)) \cdot h(x_1, \ldots, x_n) \quad \textit{pour tout } \theta \in \Theta.$$

**Lemme 7.** *Soit le modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$. Alors ce modèle est dominé si et seulement si il existe une sous-famille dénombrable $(\mathbb{P}_{\theta_i})_{i\in\mathbb{N}}$ telle que pour tout $A \in \mathcal{A}$, $\forall i \in \mathbb{N}$, $\mathbb{P}_{\theta_i}(A) = 0$ entraîne $\forall \theta$, $\mathbb{P}_\theta(A) = 0$. Toute mesure de probabilité de la forme $\mathbb{P}^* = \sum_{i\in\mathbb{N}} a_i \cdot \mathbb{P}_{\theta_i}$ avec $c_i > 0$ pour tout $i \in \mathbb{N}$ et $\sum_{i\in\mathbb{N}} c_i = 1$ domine le modèle.*

*Démonstration du lemme :* $\Longleftarrow$ Il est bien clair que si une telle mesure $P^*$ existe, le modèle est dominé.
$\Longrightarrow$ Montrons maintenant que si le modèle est dominé par une mesure $\mu$ alors la famille $(P_{\theta_i})_{i\in\mathbb{N}}$ existe. En premier lieu, si $\mu$ est une mesure non finie mais $\sigma$-finie (par exemple la mesure de Lebesgue), alors $\mu_P$ définie par $\mu_P(A) = \displaystyle\sum_{i=1}^{\infty} \frac{1}{2^i} \frac{\mu(A \cap A_i)}{\mu(A_i)}$ pour tout $A \in \mathcal{A}$, est une mesure de probabilité équivalente à $\mu$ (avec $(A_i)_{i\in\mathbb{N}^*}$ une partition de $(\Omega')^n$ telle que $0 < \mu(A_i) < \infty$ pour tout $i \in \mathbb{N}^*$). On travaille donc désormais avec $\mu_P$.
Pour $\theta \in \Theta$, soit $B_\theta$ le sous-ensemble de $(\Omega')^n \subset \mathbb{R}^n$ qui est le support de la densité de $\mathbb{P}_\theta$ par rapport à $\mu$. Soit

$$\mathcal{C} = \Big\{ \bigcup_{i\in I} B_{\theta_i},\ I \subset \mathbb{N}, \theta_i \in \Theta \Big\},$$

l'ensemble de toutes les unions dénombrables d'ensembles $B_\theta$. On note $M = \sup_{C\in\mathcal{C}} \mu_P(C)$. Soit $(C_n)_{n\in\mathbb{N}}$ une suite d'ensembles de $\mathcal{C}$ telle que la suite $(\mu_P(C_n))_n$ converge vers $M$ (une telle suite existe forcément sinon $M$ ne serait pas le supremum). Remarquons que chaque $C_i$ étant une union dénombrable de $B_{\theta_k}$, alors une suite $(\theta_n)$ de $\theta$ suffit pour engendrer la suite $(C_n)_{n\in\mathbb{N}}$. Si on pose :

$$D = \bigcup_{n\in\mathbb{N}} C_n = \bigcup_{k\in\mathbb{N}} B_{\theta_k},$$

alors $M = \mu_P(D)$ et pour tout $\theta \in \Theta$, $B_\theta \cup D \in \mathcal{C}$ et :

$$\mu_P(B_\theta \cup D) \leq M \leq \mu_P(B_\theta \cup D) = \mu_P(B_\theta \cap D^c) + \mu_P(D)$$

Donc pour tout $\theta \in \Theta$, $\mu_P(B_\theta \cap D^c) = 0$ soit $\forall \theta \in \Theta$, $\mathbb{P}_\theta(B_\theta \cap D^c) = 0$ puisque $\mathbb{P}_\theta << \mu_P$. En conséquence, pour tout $A \in \mathcal{A}'_n$, $A \subset B_\theta \cup B_\theta^c = (\Omega')^n$, soit :

$$\mathbb{P}_\theta(A \cap D^c) = 0, \quad \text{car par définition des } B_\theta,\ \mathbb{P}_\theta(B_\theta^c) = 0.$$

Si on suppose maintenant que $A \in \mathcal{A}'_n$ est tel que $\mathbb{P}_{\theta_k}(A) = 0$, avec la suite $(\theta_k)$ précédemment définie, alors $\mu_P(A \cap B_{\theta_k}) = 0$ par définition des $B_\theta$ et donc $\mu_P(A \cap D) = 0$ (par la propriété de $\sigma$-additivité d'une mesure). Comme $\mathbb{P}_\theta << \mu_P$, on en déduit que $\forall \theta \in \Theta$, $\mathbb{P}_\theta(A \cap D) = 0$ et donc $\mathbb{P}_\theta(A) = \mathbb{P}_\theta(A \cap D) + = \mathbb{P}_\theta(A \cap D^c) = 0$. Ainsi, $\mathbb{P}^*$ domine bien $\mathbb{P}_\theta$ pour tout $\theta \in \Theta$. ■

*Démonstration du Théorème de factorisation de Neyman :* Soit $\mathbb{P}^* = \sum_{i\in\mathbb{N}} a_i \cdot \mathbb{P}_{\theta_i}$ une mesure de probabilité dominante construite comme dans le lemme.
$\Longleftarrow$ Si $g_\theta(\widehat{T}(x)) \cdot h(x)$ avec $x \in (\Omega')^n$ est la densité de $\mathbb{P}_\theta$ par rapport à $\mu$, alors $\sum_{i\in\mathbb{N}} a_i \cdot g_{\theta_i}(\widehat{T}(x)) \cdot h(x) = g_*(\widehat{T}(x)) \cdot h(x)$ est une densité de $P^*$ par rapport à $\mu$. Alors, comme $g_*(\widehat{T}(x)) \cdot h(x) > 0$ $P^*$-p.s., donc $\mathbb{P}_\theta$-p.s.,

pour toute variable aléatoire $S$ intégrable, pour tout $\theta \in \Theta$ :

$$\begin{aligned}
\mathbb{E}_\theta(S \cdot \mathbb{I}_B) &= \int_B S \, d\mathbb{P}_\theta, \quad \text{pour tout } B \in \sigma(\widehat{T}), \quad \text{tribu engendrée par } \widehat{T} \\
&= \int_B S(x) \cdot g_\theta(\widehat{T}(x)) \cdot h(x) \, d\mu(x) \\
&= \int_B S(x) \cdot \frac{g_\theta(\widehat{T}(x)) \cdot h(x)}{g_*(\widehat{T}(x)) \cdot h(x)} \, d\mathbb{P}^*(x) \\
&= \mathbb{E}_*\Big(\mathbb{I}_B \cdot \frac{g_\theta(\widehat{T})}{g_*(\widehat{T})} \cdot S\Big) \\
&= \mathbb{E}_*\Big(\mathbb{I}_B \cdot \frac{g_\theta(\widehat{T})}{g_*(\widehat{T})} \cdot \mathbb{E}_*(S \mid \widehat{T})\Big) \quad \text{(d'après la définition de l'espérance conditionnelle)} \\
&= \mathbb{E}_\theta\Big(\mathbb{I}_B \cdot \mathbb{E}_*(S \mid \widehat{T})\Big).
\end{aligned}$$

En conséquence, d'après la définition de l'espérance conditionnelle dans $\mathbb{L}^1((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta)$, on a $\mathbb{P}_\theta$-p.s., $\mathbb{E}_*(S \mid \widehat{T}) = \mathbb{E}_\theta(S \mid \widehat{T})$ : la statistique $\widehat{T}$ est bien exhaustive.
$\Longrightarrow$ On suppose que $\widehat{T}$ est une statistique exhaustive pour le modèle. Donc pour toute statistique intégrable $S$, $\forall \theta$, $\mathbb{E}_*(S \mid \widehat{T}) = \mathbb{E}_\theta(S \mid \widehat{T})$. En conséquence, si on note $\phi(x, \theta) = \dfrac{d\mathbb{P}_\theta}{d\mathbb{P}_*}(x)$ la densité de $\mathbb{P}_\theta$ par rapport à $\mathbb{P}_*$,

$$\begin{aligned}
\mathbb{E}_\theta(S) &= \mathbb{E}_\theta\big(\mathbb{E}_*(S \mid \widehat{T})\big), \quad \text{(car } \widehat{T} \text{ est exhaustive et d'après les propriétés de l'espérance conditionnelle)} \\
&= \mathbb{E}_*\big(\phi(X, \theta) \cdot \mathbb{E}_\mu(S \mid \widehat{T})\big), \quad \text{où } X \sim \mathbb{P}_* \\
&= \mathbb{E}_*\Big[\mathbb{E}_*\Big(\phi(X, \theta) \cdot \mathbb{E}_*(S \mid \widehat{T}) \mid \widehat{T}\Big)\Big], \quad \text{(d'après les propriétés de l'espérance conditionnelle)} \\
&= \mathbb{E}_*\Big[\mathbb{E}_*\Big(\phi(X, \theta) \mid \widehat{T}\Big) \cdot \mathbb{E}_*\Big(S \mid \widehat{T}\Big)\Big], \quad \text{(car } \mathbb{E}_*\Big(S \mid \widehat{T}\Big) \text{ est une fonction de } \widehat{T}) \\
&= \mathbb{E}_*\Big[\mathbb{E}_*\Big(S \cdot \mathbb{E}_*\big(\phi(X, \theta) \mid \widehat{T}\big) \mid \widehat{T}\Big)\Big] \\
&= \mathbb{E}_*\Big[S \cdot \mathbb{E}_*\big(\phi(X, \theta) \mid \widehat{T}\big)\Big]
\end{aligned}$$

Ainsi, la variable aléatoire $\mathbb{E}_*\big(\phi(X, \theta) \mid \widehat{T}\big)$, qui est une fonction de $\widehat{T}$ (qui est elle-même une fonction sur $(\Omega')^n$), est la densité de $\mathbb{P}_\theta$ par rapport à $\mathbb{P}_*$. Par suite, la vraisemblance, qui est la densité de $\mathbb{P}_\theta$ par rapport à $\mu$, s'écrit :

$$L_\theta(x_1, \ldots, x_n) = \frac{d\mathbb{P}_\theta}{d\mu}(x_1, \ldots, x_n) = \frac{d\mathbb{P}_\theta}{d\mathbb{P}_*}(x_1, \ldots, x_n) \cdot \frac{d\mathbb{P}_*}{d\mu}(x_1, \ldots, x_n) = \mathbb{E}_*\big(\phi(X, \theta) \mid \widehat{T}\big) \cdot h(x_1, \ldots, x_n),$$

avec $h$ une fonction mesurable. ■

**Exemple 7.** *Différentes statistiques exhaustives pour les modèles paramétriques de loi uniforme, de loi de Bernoulli, de loi gaussienne...*

**Propriété 12.** *On se place dans le cadre d'un modèle paramétrique dominé.*

1. *La statistique $\widehat{T} = (X_1, \ldots, X_n)$ est exhaustive.*
2. *Si $\widehat{T}$ est une statistique exhaustive et s'il existe une fonction borélienne $h$ telle qu'une autre statistique $\widehat{U}$ vérifie $\widehat{T} = h(\widehat{U})$, alors $\widehat{U}$ est également exhaustive.*

On vient de voir que l'on peut toujours trouver une statistique exhaustive (l'échantillon lui-même par exemple). Comme on aurait plutôt tendance à vouloir le "maximum d'information" dans une statistique exhaustive, lorsque le paramètre est dans $\mathbb{R}^d$, on aimerait savoir quelle dimension minimale peut avoir cette statistique. En particulier, si $d = 1$, peut-on toujours trouver une statistique exhaustive de taille 1 ? L'exemple suivant montre que ce n'est pas toujours le cas :

**Exemple 8.** *Soit le modèle statistique* $([0,\infty[^n, \mathcal{B}([0,\infty[^n), (\mathbb{P}_\theta)^{\otimes n}, \theta \in \mathbb{R}_+)$*, où la densité de* $\mathbb{P}_\theta$ *par rapport à la mesure de Lebesgue est :* $f_\theta(x) = \theta(e^{\theta^2}-1)\cdot e^{-\theta\cdot x}\cdot \mathbb{I}_{x\in[0,\theta]}\cdot$*. Alors les statistiques* $\widehat{T}_1 = \max(X_1,\ldots,X_n)$ *et* $\widehat{T}_2 = X_1+\ldots+X_n$ *ne sont pas chacune exhaustive alors que* $\widehat{T} = (\widehat{T}_1, \widehat{T}_2)$ *est exhaustive. On pourra même montrer que cette statistique est de taille minimale...*

Qu'elle serait une sorte d'opposée de la notion de statistique exhaustive ? Ce devrait être une statistique ne dépendant pas du paramètre, soit :

**Définition 18.** *Une statistique* $\widehat{T}$ *d'un modèle paramétrique est dite libre si sa loi ne dépend pas du paramètre.*

Peut-on rajouter une autre caractérisation des statistiques exhaustives pour pouvoir atteindre une forme d'optimalité pour ces statistiques, qui serait qu'aucune fonctionnelle non constante de la statistique ne peut être libre. Cela peut également se traduire de la façon suivante :

**Définition 19.** *Une statistique exhaustive* $\widehat{T}$ *du modèle statistique paramétrique dominé avec* $\widehat{T}$ *à valeur dans* $\mathbb{R}^d$ *est dite complète si pour toute fonction borélienne* $h : \mathbb{R}^d \to \mathbb{R}$ *telle que* $h(\widehat{T})$ *soit intégrable, alors :*

$$\forall\theta\in\Theta,\quad \mathbb{E}_\theta\big(h(\widehat{T})\big) = 0 \quad\Longrightarrow\quad h(\widehat{T}) = 0.$$

**Propriété 13.** *Soit un modèle statistique paramétrique dominé.*

1. *si* $\widehat{T}$ *est une statistique exhaustive complète alors pour toute fonction borélienne* $h$ *bijective* $h(\widehat{T})$ *est une statistique exhaustive complète.*
2. *(Théorème de Basu) si* $\widehat{T}$ *est une statistique exhaustive complète alors* $\widehat{T}$ *est indépendante de toute statistique libre sur le modèle.*

*Démonstration de la propriété :* 3. Théorème de Basu. Soit $\widehat{S}$ une statistique libre pour le modèle et soit $f$ une fonction telle que $\mathbb{E}_\theta(f(\widehat{S}))$ existe. Comme $\widehat{S}$ est libre, on peut noter $e(f) = \mathbb{E}_\theta(f(\widehat{S}))$ une application linéaire ne dépendant pas de $\theta$. Par suite, la statistique $\mathbb{E}_\mu(f(\widehat{S}) \mid \widehat{T}) - e(f)$ est une fonction de $\widehat{T}$ mesurable telle que $\mathbb{E}_\theta\Big(\mathbb{E}_\mu(f(\widehat{S}) \mid \widehat{T}) - e(f)\Big) = 0$ pour tout $\theta\in\Theta$. Comme on a supposé que $\widehat{T}$ est exhaustive complète, alors $\mathbb{E}_\mu(f(\widehat{S}) \mid \widehat{T}) = e(f)$ presque-sûrement : les statistiques $\widehat{S}$ et $\widehat{T}$ sont indépendantes. ■

**Définition 20.** *On suppose un modèle paramétrique* $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta\in\Theta\subset\mathbb{R}^p)$ *dominé par une mesure* $\mu$*. Si, pour tout* $(x_1,\cdots,x_n)\in(\Omega')^n$ *et* $\theta\in\Theta$*, la vraisemblance de ce modèle par rapport à* $\mu$ *peut s'écrire sous la forme :*

$$L_\theta(x_1,\cdots,x_n) = \exp\left(\beta(\theta) + b(x_1,\cdots,x_n) + \sum_{j=1}^{p} a_j(x_1,\cdots,x_n)\cdot\alpha_j(\theta)\right), \tag{2}$$

*avec les fonctions* $a_j : (\Omega')^n \to \mathbb{R}$*,* $b : (\Omega')^n \to \mathbb{R}$*,* $\alpha_j : \Theta\subset\mathbb{R}^p \to \mathbb{R}$*, et* $\beta : \Theta\to\mathbb{R}$*, alors on dit que le modèle est exponentiel (ou qu'il appartient à la famille exponentielle).*

**Exemple 9.** *Appartiennent à la famille exponentielle les lois :*
- *Loi discrètes : Lois de Bernoulli, binomiales, de Poisson,...*
- *Loi "continues" : Lois normales, exponentielles, gamma, du chi-deux,...*

**Remarque 4.** *Si* $(X_1,\cdots,X_n)$ *est un n-échantillon d'un modèle exponentiel (avec* $\theta$ *fixé) alors l'ensemble des valeurs prises par* $(X_1,\cdots,X_n)$ *ne dépend pas du paramètre* $\theta$*.*

**Propriété 14.** *Soit un modèle exponentiel. Si pour tout* $\theta\in\Theta$ *on note* $\alpha(\theta) = (\alpha_1(\theta),\ldots,\alpha_p(\theta))$ *et si l'ensemble* $\alpha(\Theta)$ *est d'intérieur non vide, alors* $\widehat{T}(x_1,\cdots,x_n) = (a_1(x_1,\cdots,x_n),\ldots,a_p(x_1,\cdots,x_n))$ *est une statistique exhaustive et complète.*

*Démonstration de la propriété :* Soit $g : \mathbb{R}^p \to \mathbb{R}$ telle que $\mathbb{E}_\theta(g(\widehat{T})) = 0$. Or, $\forall\theta\in\Theta$,

$$\mathbb{E}_\theta(g(\widehat{T})) = \int_{(\Omega')^n} g(\widehat{T}(x))\cdot\exp\big(\beta(\theta) + b(x) + <\widehat{T}(x),\alpha(\theta)>\big)\,d\mu(x),$$

où $< ., . >$ désigne le produit scalaire. En considérant la mesure $\nu$ de densité $\exp(b(x))$ par rapport à $\mu$, on obtient :

$$\begin{aligned} \mathbb{E}_\theta(g(\widehat{T})) = 0 \quad &\Longrightarrow \quad \int_{(\Omega')^n} g(\widehat{T}(x)) \cdot \exp(< \widehat{T}(x), \alpha(\theta) >) \, d\nu(x) = 0 \\ &\Longrightarrow \quad \int_{\widehat{T}((\Omega')^n)} g(y) \cdot \exp(< y, \alpha(\theta) >) \, d\nu_{\widehat{T}}(y) = 0 \end{aligned}$$

pour tout $\theta \in \Theta$, en ayant noté $\nu_{\widehat{T}}$ la mesure image de $\nu$ par $\widehat{T}$ et avec $\widehat{T}((\Omega')^n) \in \mathbb{R}^p$. Si on note $g^+$ et $g^-$ les parties positives et négatives de $g$ (donc $g = g^+ - g^-$), et $\pi^+$ et $\pi^-$ les mesures de densités $g^+$ et $g^-$ par rapport à $\nu_{\widehat{T}}$, alors, pour tout $\theta \in \Theta$ :

$$\int_{\widehat{T}((\Omega')^n)} \exp(< y, \alpha(\theta) >) \, d\pi^+(y) = \int_{\widehat{T}((\Omega')^n)} \exp(< y, \alpha(\theta) >) \, d\pi^-(y).$$

En conséquence sur $\Theta$, donc sur une partie d'intérieure non vide, les mesures $\pi^+$ et $\pi^-$ ont des transformées de Laplace égales : ces deux mesures sont donc égales et donc $g^+ = g^- \ \nu_{\widehat{T}}$-presque partout (ce qui revient à $g = 0$). A partir des expressions des différentes mesures, on montre que $g = 0$, $\widehat{T}(\mathbb{P}_\theta)$-presque partout. ■

## 2.3 Information de Fisher

Pour mesurer l'information fournit par un modèle paramétrique dominé (ou une statistique sur ce modèle) au sujet d'un paramètre, une idée naturelle serait de mesurer comment varie localement la mesure de probabilité, ou encore sa vraisemblance. Les fluctuations moyennes de cette vraisemblance serait donc un bon indicateur : pour ce faire on considérera, lorsqu'il existe $\text{grad}_\theta(L_\theta(X_1, \cdots, X_n))$, et on s'intéressera à la matrice de covariance de $\text{grad}_\theta(L_\theta(X_1, \cdots, X_n))$, dont on peut montrer qu'elle ne dépend pas du choix de la mesure dominante choisie. Précisons d'abord la notion de modèle régulier qui nous permettra de définir cette quantité d'information.

**Définition 21.** *Dans le cadre d'un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure $\mu$, on dira que ce modèle est régulier lorsque :*

1. *$\Theta$ est un ouvert de $\mathbb{R}^d$ ;*
2. *la vraisemblance $L_\theta(.)$ vérifie $\forall (x_1, \ldots, x_n) \in (\Omega')^n$, $\forall \theta \in \Theta$, $L_\theta(x_1, \ldots, x_n) > 0$ ;*
3. *$\forall (x_1, \ldots, x_n) \in (\Omega')^n$, la fonction $\theta \in \Theta \mapsto \log(L_\theta(.))$ est différentiable sur $\Theta$ par rapport à $\theta$, et son gradient appartient à $\mathbb{L}^2((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta)$ $\forall \theta \in \Theta$ ;*
4. *$\forall \theta \in \Theta$, pour toute fonction $h : \mathbb{R}^n \to \mathbb{R}$ appartenant à $\mathbb{L}^1((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta)$, alors :*

$$\frac{\partial}{\partial \theta} \int_{(\Omega')^n} h(x) \cdot L_\theta(x) \, d\mu(x) = \int_{(\Omega')^n} h(x) \cdot \frac{\partial}{\partial \theta} L_\theta(x) \, d\mu(x). \tag{3}$$

**Conséquence 4.** *Pour un modèle régulier, $\mathbb{E}_\theta\Big(grad_\theta(\log L_\theta(.))\Big) = 0$.*

*Démonstration* : On a $\mathbb{E}_\mu(L_\theta(.)) = 1$ donc $\mathbb{E}_\mu(\text{grad}_\theta L_\theta(.)) = 0$. Par conséquent, $\mathbb{E}_\mu\Big(\dfrac{\text{grad}_\theta(L_\theta(.))}{L_\theta(.)}\Big) = 0$, soit $\mathbb{E}_\theta\big(\text{grad}_\theta(\log L_\theta(.))\big) = 0$. ■

**Définition 22.** *Pour un modèle statistique paramétrique dominé régulier, on appelle information de Fisher, la matrice :*

$$I_n(\theta) = \left[\mathbb{E}_\theta\left(\frac{\partial(\log L_\theta(X_1, \cdots, X_N))}{\partial \theta_i} \times \frac{\partial(\log L_\theta(X_1, \cdots, X_N))}{\partial \theta_j}\right)\right]_{1 \le i,j \le p}.$$

**Propriété 15.** *Pour un modèle statistique paramétrique dominé régulier, et si $\forall (x_1, \ldots, x_n) \in (\Omega')^n$, la fonction $\theta \in \Theta \mapsto \log(L_\theta(.))$ est $\mathcal{C}^2(\Theta)$, alors :*

$$I_n(\theta) = -\left[\mathbb{E}_\theta\left(\frac{\partial^2(\log L_\theta(X_1, \cdots, X_N))}{\partial \theta_i \cdot \partial \theta_j}\right)\right]_{1 \le i,j \le p}.$$

**Définition 23.** *L'information de Fisher $I_n^{\widehat{T}}(\theta)$ associée à une statistique $\widehat{T}$, si elle existe, est la matrice de Fisher de la vraisemblance de $\widehat{T}$ (déterminée à partir de la vraisemblance de $\widehat{T}$).*

**Propriété 16.** *Pour un modèle régulier, $\widehat{T}$ est une statistique libre si et seulement si $I_n^{\widehat{T}}(\theta) = 0$.*

<u>*Démonstration :*</u> $\Longrightarrow$ Si $\widehat{T}$ est libre alors sa loi ne dépend pas de $\theta$ donc le gradient du logarithme de sa vraisemblance est nul ; l'information de Fisher associée à $\widehat{T}$ est nulle.
$\Longleftarrow$ Si $I_n^{\widehat{T}}(\theta) = 0$, donc la statistique $\mathrm{grad}_\theta(\log L_\theta^{\widehat{T}}(\widehat{T}))$ est centrée et de matrice de covariance nulle. Ainsi, pour tout $\theta \in \Theta$, il existe un ensemble $N_\theta$ de mesure 1 pour la mesure de probabilité associée à $\widehat{T}$ (donc, d'après la première hypothèse d'un modèle régulier, tel que $\mu(N_\theta) = 1$)) et tel que pour tout $t \in N_\theta$, $\mathrm{grad}_\theta(\log L_\theta^{\widehat{T}}(t))) = 0$. Pour montrer que $\mathrm{grad}_\theta(\log L_\theta^{\widehat{T}}(t))) = 0$ est bien une variable aléatoire nulle $\mu$-p.s., et donc que $\log L_\theta^{\widehat{T}}(.)$ est une fonction constante en $\theta$, il nous faut montrer que finalement les $\theta$ ne dépendent pas de $\theta$. Soit $\Theta^{(d)} = \{\theta_i^{(d)}\}_{i\in\mathbb{N}}$ un sous-ensemble dénombrable de $\Theta$, dense dans $\Theta$. Comme $\Theta^{(d)}$ est dénombrable, il est clair que $N = \bigcup_{i\in\mathbb{N}} N_{\theta_i^{(d)}}$ est tel que $\mu(N) = 1$. De plus, pour tout $\theta \in \Theta$, il existe une sous-suite $(\theta_{\phi(n)}^{(d)})_n$ de $\Theta^{(d)}$ convergeant vers $\theta$ et telle que pour tout $t \in N$, pour tout $n \in \mathbb{N}$, $\mathrm{grad}_{\theta_{\phi(n)}^{(d)}}(\log L_{\theta_{\phi(n)}^{(d)}}^{\widehat{T}}(t))) = 0$. Comme une telle fonction de $\theta_{\phi(n)}^{(d)}$ est continue, cette propriété passe à la limite, et donc pour tout $t \in N$, $\forall\theta \in \Theta$, $\mathrm{grad}_\theta(\log L_\theta^{\widehat{T}}(t))) = 0$. Comme $N$ ne dépend pas de $\theta$, alors la fonction $\theta \to \log L_\theta^{\widehat{T}}(.)$ est une constante ne dépendant pas de $\theta$, $\mu$-p.s. : la statistique $\widehat{T}$ est bien libre. ■

**Propriété 17.** *Pour un modèle régulier, si $\widehat{T}$ est une statistique exhaustive : $I_n^{\widehat{T}}(\theta) = I_n(\theta)$ pour tout $\theta \in \Theta$.*

<u>*Démonstration :*</u> Comme $\widehat{T}$ est une statistique exhaustive, on peut écrire d'après la démonstration du Théorème de factorisation de Neyman que pour tout $(x_1, \ldots, x_n) \in (\Omega')^n$ et tout $\theta \in \Theta$ :

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}^*}(x_1, \ldots, x_n) = g_\theta(\widehat{T}(x_1, \ldots, x_n)).$$

On peut réécrire cela pour la densité de $\widehat{T}$ sous la forme : $\dfrac{d\mathbb{P}_\theta^{\widehat{T}}}{d\mathbb{P}^{*\widehat{T}}}(t) = g_\theta(t))$, pour tout $t \in \widehat{T}((\Omega')^n)$ et tout $\theta \in \Theta$. En conséquence, pour tout $\theta \in \Theta$,

$$\begin{aligned}
I(\theta) &= \left[\mathbb{E}_\theta\left(\frac{\partial(\log L_\theta(X_1, \cdots, X_N))}{\partial\theta_i} \times \frac{\partial(\log L_\theta(X_1, \cdots, X_N))}{\partial\theta_j}\right)\right]_{1\le i,j\le p} \\
&= \left[\int_{(\Omega')^n}\left(\frac{\partial(\log L_\theta(x))}{\partial\theta_i} \times \frac{\partial(\log L_\theta(x))}{\partial\theta_j}\right) d\mathbb{P}_\theta(x)\right]_{1\le i,j\le p} \\
&= \left[\int_{(\Omega')^n}\left(\frac{\partial(\log g_\theta(\widehat{T}(x)))}{\partial\theta_i} \times \frac{\partial(\log g_\theta(\widehat{T}(x)))}{\partial\theta_j}\right) g_\theta(\widehat{T}(x)) d\mathbb{P}^*(x)\right]_{1\le i,j\le p} \quad \text{car } \log L_\theta(x) = \log g_\theta(\widehat{T}(x)) + \log h(x) \\
&= \left[\int_{\widehat{T}(\Omega')^n}\left(\frac{\partial(\log g_\theta(t))}{\partial\theta_i} \times \frac{\partial(\log g_\theta(t))}{\partial\theta_j}\right) g_\theta(t)) d\mathbb{P}^{*\widehat{T}}(x)\right]_{1\le i,j\le p} \quad \text{d'après le théorème du transport} \\
&= \left[\int_{\widehat{T}(\Omega')^n}\left(\frac{\partial(\log g_\theta(t))}{\partial\theta_i} \times \frac{\partial(\log g_\theta(t))}{\partial\theta_j}\right) d\mathbb{P}_\theta(t)\right]_{1\le i,j\le p} \\
&= I_n^{\widehat{T}}(\theta).
\end{aligned}$$

■

**Remarque 5.** *En rajoutant certaines hypothèses de continuité sur la vraisemblance de $\widehat{T}$, on peut montrer que la réciproque est également vraie, et donc que $I_n^{\widehat{T}}(\theta) = 0$ si et seulement si la statistique $\widehat{T}$ est exhaustive.*

Ainsi, on retrouve à l'aide de la notion d'information de Fisher les "intuitions" qui nous avaient guidées dans la section précédentes. Voyons maintenant les applications de la notion d'exhaustivité à l'estimation paramétrique.

## 2.4 Application à l'estimation paramétrique

On se place dans le cadre d'un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure $\mu$. Par ailleurs, on suppose que $\Theta$ est un ouvert.

**Définition 24.** — *Soit $g : \Theta \to \Theta'$, où $\Theta' \subset \mathbb{R}^{p'}$ avec $p' \in \mathbb{N}^*$, une fonction mesurable. On appelle estimateur de la fonction $g$ du paramètre, donc de $g(\theta)$, une statistique $\widehat{T}$ à valeurs dans $\mathbb{R}^{p'}$. En particulier, un estimateur du paramètre $\theta$ est une statistique à valeurs dans $\mathbb{R}^p$. Une estimation de $g(\theta)$ est une réalisation de $\widehat{T}$.*

— *On appelle biais d'un estimateur $\widehat{T}$ de $g(\theta)$ le vecteur constant de $\mathbb{R}^{p'}$, $B(\theta) = \mathbb{E}_\theta(\widehat{T}) - g(\theta)$. On dira que l'estimateur est sans biais si $B(\theta) = 0$ pour tout $\theta \in \Theta$.*

— *On appelle risque quadratique de l'estimateur $\widehat{T}$ de $g(\theta)$ le réel positif $R(\theta) = \mathbb{E}_\theta(\|\widehat{T}) - g(\theta)\|^2)$, où $\|.\|$ désigne usuellement la norme euclidienne (mais peut être une autre fonctionnelle positive et convexe). Si l'estimateur est sans biais alors, $R(\theta) = Trace(cov(\widehat{T}))$.*

Pour pouvoir parler du comportement asymptotique d'une statistique, on va devoir se placer dans un "gros" modèle, dans lequel un échantillon est une suite de v.a. En quelque sorte, ce gros modèle pourra s'écrire $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$ (la dimension du paramètre reste constante). Pour un $n$ fixé, une statistique $\widehat{T}_n$ sera d'abord une projection du "gros" modèle sur le modèle de taille $n$, puis une statistique "normale". On devra donc parler d'une suite d'estimateurs $(\widehat{T}_n)_n$

**Définition 25.** *Pour un modèle statistique paramétrique $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, et pour $(\widehat{T}_n)_n$ une suite d'estimateurs de $g(\theta)$ :*

— *Si $\lim_{n\to\infty} B_n(\theta) = 0$, on dit que l'estimateur est asymptotiquement sans biais.*

— *On dit que $(\widehat{T}_n)_n$ est convergent lorsque $\widehat{T}_n \xrightarrow[n\to+\infty]{\mathcal{P}} g(\theta)$.*

— *S'il existe $(a_n)$ une suite de réels positifs tels que $a_n(\widehat{T}_n - g(\theta)) \xrightarrow[n\to+\infty]{\mathcal{L}} Z_\theta$, où $Z_\theta$ est une loi centrée non nulle (ne dépendent pas de $n$), on dit $(\widehat{T}_n)_n$ converge vers $g(\theta)$ à la vitesse $a_n$.*

A priori, être sans biais n'est pas un bon critère pour garantir une certaine optimalité de la convergence d'un estimateur. On préférera plutôt discriminer entre de potentiels estimateurs à l'aide d'un critère portant sur le risque quadratique ou sur la matrice de variance-covariance. Cependant, il n'existe pas de résultats généraux pour trouver un "meilleur" estimateur en ce sens. Pour en obtenir, on devra se limiter à une certaine classe d'estimateurs, celle des estimateurs sans biais.

**Définition 26.** *Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, et soit $\widehat{T}$ un estimateur sans biais de $g(\theta)$. On dit que $\widehat{T}$ est de variance uniformément minimum parmi les estimateurs sans biais de $g(\theta)$ lorsque pour tout estimateur sans biais de $g(\theta)$, on a $\forall \theta \in \theta$, $cov(\widehat{T}) \leq cov(\widehat{S})$ (au sens où $cov(\widehat{T}) - cov(\widehat{S})$ est une matrice positive).*

**Propriété 18.** *Si $\widehat{T}$ est un estimateur de variance uniformément minimum parmi les estimateurs sans biais, alors il est unique $\mathbb{P}_\theta$-p.s.*

<u>*Démonstration*</u> *:* Soit $\widehat{S}$ un autre estimateur que l'on suppose également de variance uniformément minimum parmi les estimateurs sans biais. Montrons d'abord que $E_\theta((\widehat{T} - \widehat{S}) \cdot {}^t\widehat{T}) = 0$. En effet, si $\alpha \in \mathbb{R}$, comme $\widehat{T}$ est de variance minimum, en utilisant des inégalités sur les matrices symétriques :

$$\begin{aligned} \operatorname{cov}(\widehat{T}) &\leq \operatorname{cov}(\widehat{T} + \alpha(\widehat{T} - \widehat{S})) \\ &\leq \operatorname{cov}(\widehat{T}) + \alpha^2 \operatorname{cov}(\widehat{T} - \widehat{S}) + 2\alpha \cdot \mathbb{E}_\theta(\widehat{T} \cdot {}^t\widehat{S}) \\ \Longrightarrow \quad 0 &\leq \alpha \cdot \big(\alpha \cdot \operatorname{cov}(\widehat{S}) + 2\mathbb{E}_\theta(\widehat{T} \cdot {}^t(\widehat{T} - \widehat{S}))\big) \quad \text{pour tout } \alpha \in \mathbb{R}. \end{aligned}$$

Comme $\operatorname{cov}(\widehat{T} - \widehat{S})$ est une matrice positive, la seule possibilité pour avoir la dernière inégalité est que : $\mathbb{E}_\theta(\widehat{T} \cdot {}^t(\widehat{T} - \widehat{S})) = 0$. Par suite, comme $\operatorname{cov}(\widehat{T} - \widehat{S}) = \mathbb{E}_\theta\big((\widehat{T} - \widehat{S}) \cdot {}^t(\widehat{T} - \widehat{S})\big) = \mathbb{E}_\theta\big(\widehat{T} \cdot {}^t(\widehat{T} - \widehat{S})\big) - \mathbb{E}_\theta\big(\widehat{S} \cdot {}^t(\widehat{T} - \widehat{S})\big)$, et que l'on a supposé $\widehat{T}$ et $\widehat{S}$ de variance minimum, $\operatorname{cov}(\widehat{T} - \widehat{S}) = 0$. Donc $\widehat{T} = \widehat{S}$ sur un ensemble de $\mathbb{P}_\theta$-mesure égale à 1. ■

**Théorème 17** (Rao-Blackwell)**.** *Si $\widehat{T}$ est un estimateur sans biais de $g(\theta)$ et si $\widehat{S}$ est une statistique exhaustive, alors $\widehat{R} = \mathbb{E}_\theta(\widehat{T} \mid \widehat{S})$, qui ne dépend pas de $\theta$ car $\widehat{S}$ est exhaustive, est un estimateur sans biais de $g(\theta)$ de matrice de covariance inférieure ou égale à celle de $\widehat{T}$.*

*Démonstration :* il est clair que $\mathbb{E}_\theta(\widehat{R}) = \mathbb{E}_\theta(\widehat{T}) = g(\theta)$. De plus, pour tout $u \in \mathbb{R}^{p'}$ (avec $g : \mathbb{R}^p \to \mathbb{R}^{p'}$),

$$\begin{aligned}
\mathrm{cov}({}^t u \cdot \widehat{T}) &= \mathbb{E}_\theta\Big[\big({}^t u \cdot (\widehat{T} - g(\theta))\big)^2\Big] \\
&= \mathbb{E}_\theta\Big(\mathbb{E}_\theta\Big[\big({}^t u \cdot (\widehat{T} - g(\theta))\big)^2\Big] \mid \widehat{S}\Big) \\
&\geq \mathbb{E}_\theta\Big(\mathbb{E}_\theta\Big[{}^t u \cdot (\widehat{T} - g(\theta)) \mid \widehat{S}\Big]^2\Big) \quad \text{d'après l'inégalité de Jensen,} \\
&\geq \mathrm{cov}({}^t u \cdot \widehat{R}).
\end{aligned}$$

Cela revient bien à écrire que $\mathrm{cov}(\widehat{T}) \geq \mathrm{cov}(\widehat{R})$. ■

**Théorème 18** (Lehmann-Scheffé)**.** *Si $\widehat{T}$ est un estimateur sans biais de $g(\theta)$ et si $\widehat{S}$ est une statistique exhaustive et complète, alors l'unique estimateur de $g(\theta)$ sans biais uniformément de variance minimale est $\widehat{R} = \mathbb{E}_\theta(\widehat{T} \mid \widehat{S})$ (c'est-à-dire que $\widehat{R}$ est une fonction de $\widehat{S}$).*

*Démonstration :* Soit $\widehat{T}'$ un autre estimateur sans biais de $g(\theta)$. Si $\widehat{R}' = \mathbb{E}_\theta(\widehat{T}' \mid \widehat{S})$, on sait que $\mathrm{cov}(\widehat{T}') \geq \mathrm{cov}(\widehat{R}')$ d'après le Théorème de Rao-Blackwell. Or $\mathbb{E}_\theta(\widehat{R} - \widehat{R}') = 0$ pour tout $\theta \in \Theta$ car les deux estimateurs sont sans biais. De plus comme $\widehat{R}$ et $\widehat{R}'$ sont des fonctions de $\widehat{S}$, $\widehat{R} - \widehat{R}'$ l'est aussi, et du fait que $\widehat{S}$ est une statistique exhaustive et complète, alors pour tout $\theta \in \Theta$, $\widehat{R} = \widehat{R}'$, $\mathbb{P}_\theta$-p.s. Par conséquent, pour tout $\theta \in \Theta$, $\mathrm{cov}(\widehat{R}') = \mathrm{cov}(\widehat{R})$ et donc $\mathrm{cov}(\widehat{R}) \leq \mathrm{cov}(\widehat{T}')$ : $\widehat{R}$ est bien l'estimateur sans biais de variance uniformément minimale. ■

Retenons donc de tout ceci que l'estimateur sans biais de $g(\theta)$ et de variance uniformément minimale est une unique fonction d'une statistique exhaustive et complète, lorsqu'une telle statistique existe. On aimerait maintenant connaître un peu mieux la covariance d'un tel estimateur.

**Théorème 19** (Inégalité de Cramer-Rao)**.** *Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'^{\otimes n}, \mathbb{P}_\theta, \theta \in \Theta)$ dominé et régulier, et soit $\widehat{T}$ un estimateur sans biais de $g(\theta)$, tel que $\mathbb{E}_\theta\|\widehat{T}\|^2 < +\infty$. Si on suppose que l'information de Fisher est une matrice définie positive, alors, en notant $\dfrac{\partial g}{\partial \theta}(\theta)$ la matrice jacobienne de $g$, pour tout $\theta \in \Theta$ :*

$$cov(\widehat{T}) \geq \frac{\partial g}{\partial \theta}(\theta) \cdot (I_n(\theta))^{-1} \cdot {}^t\frac{\partial g}{\partial \theta}(\theta) \quad (\textit{au sens des matrices symétriques}).$$

*En particulier, si $\widehat{T}$ est un estimateur sans biais de $\theta$, alors :*

$$cov(\widehat{T}) \geq (I_n(\theta))^{-1} \quad (\textit{au sens des matrices symétriques}).$$

*Démonstration :* Soit $Z_\theta(x) = \mathrm{grad}(\log L_\theta(x))$ où $x \in (\Omega')^n$ suit $\mathbb{P}_\theta$. On sait que comme le modèle est régulier, $\mathbb{E}_\theta(Z_\theta) = 0$ pour tout $\theta \in \Theta$ et donc :

$$\mathrm{cov}(Z_\theta) = I(\theta) \quad \text{pour tout } \theta \in \Theta.$$

De plus, $\widehat{T}$ est un estimateur sans biais de $g(\theta)$ donc pour tout $\theta \in \Theta$ :

$$\begin{aligned}
\mathbb{E}_\theta(\widehat{T}) = g(\theta) \quad &\Longrightarrow \quad \int_{(\Omega')^n} \widehat{T}(x) \cdot \frac{\partial L_\theta}{\partial \theta}(x)\, d\mu(x) = \frac{\partial g}{\partial \theta}(\theta) \quad \text{(en dérivant)} \\
&\Longrightarrow \quad \int_{(\Omega')^n} \widehat{T}(x) \cdot \frac{\partial L_\theta}{\partial \theta}(x) \cdot (L_\theta(x))^{-1}\, d\mathbb{P}_\theta(x) = \frac{\partial g}{\partial \theta}(\theta) \\
&\Longrightarrow \quad \mathbb{E}_\theta(\widehat{T} \cdot {}^t Z_\theta) = \frac{\partial g}{\partial \theta}(\theta).
\end{aligned}$$

Ainsi, d'après ce qui précède,

$$\begin{aligned}\mathrm{cov}_\theta(\widehat{T} - \frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta) &= \mathrm{cov}_\theta(\widehat{T}) - 2\frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial\theta}(\theta) + \frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial\theta}(\theta)\\ &= \mathrm{cov}_\theta(\widehat{T}) - \frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial\theta}(\theta).\end{aligned}$$

En conséquence, comme $\mathrm{cov}_\theta(\widehat{T} - \frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta)$ est une matrice positive, l'inégalité de Cramer-Rao est prouvée. ■

**Corollaire 3.** *Deux cas particuliers méritent attention :*

— *Si le modèle est de la forme $((\Omega')^n, \mathcal{A}'^{\otimes n}_n, (f_\theta\cdot d\mu)^{\otimes n}, \theta\in\Theta)$, alors $I_n(\theta) = n\cdot I_1(\theta)$, où $I_1(\theta)$ est la matrice d'information de Fisher d'une seule variable aléatoire $X$ distribuée suivant $f_\theta\cdot d\mu$ et l'Inégalité de Cramer-Rao devient donc :*

$$cov(\widehat{T}) \geq \frac{1}{n}\cdot\left(\frac{\partial g}{\partial\theta}(\theta)\cdot(I_1(\theta))^{-1}\cdot {}^t\frac{\partial g}{\partial\theta}(\theta)\right) \quad (\textit{au sens des matrices symétriques}).$$

*On voit donc que pour un échantillon de variables indépendantes et identiquement distribuées, si la vraisemblance est régulière, alors la vitesse de convergence de tout estimateur sans biais est au mieux en $\sqrt{n}$.*

— *Si le modèle n'est pas régulier, mais que sous la probabilité $\mathbb{P}_\theta$, la matrice d'information de Fisher existe et est inversible, et surtout si la propriété (3) est vérifiée, alors l'Inégalité de Cramer-Rao est vérifiée.* **Cela exclut cependant les modèles dont le support de $\mathbb{P}_\theta$ dépend de $\theta$,** *comme par exemple le simple modèle de v.a.i.i.d. de loi $\mathcal{U}(]0,\theta[)$, avec $\theta>0$.*

**Définition 27.** *Si un estimateur sans biais atteint (respectivement asymptotiquement) la borne de Cramer-Rao (qui ne dépend pas de l'estimateur), on dit qu'il est (resp. asymptotiquement) efficace.*

**Remarque 6.** *Un estimateur peut être sans biais, de variance minimale, mais ne pas atteindre la borne de Cramer-Rao, donc ne pas être efficace. De la même manière, il peut exister des estimateurs biaisés atteignant la borne de Cramer-Rao.*

Nous allons voir que les modèles exponentiels jouent un rôle central pour l'estimation paramétrique puisque sous certaines conditions ils sont les seuls pour lesquels on aura une estimation sans biais efficace.

**Théorème 20.** *Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'^{\otimes n}_n, \mathbb{P}_\theta, \theta\in\Theta)$, avec $\Theta\subset\mathbb{R}^p$, dominé et régulier. Soit $g:\mathbb{R}^d\to\mathbb{R}^d$ de classe $\mathcal{C}^1$ sur $\Theta$ telle que la matrice carrée de taille $p$, $\frac{\partial g}{\partial\theta}(\theta)$ soit de rang $p$ pour tout $\theta\in\Theta$. Alors $\widehat{T} = {}^t(\widehat{T}_1,\dots,\widehat{T}_d)$ est un estimateur sans biais de $g(\theta)$ atteignant la borne de Cramer-Rao si et seulement si le modèle est exponentiel et plus précisément s'il existe des fonctions $a:(\Omega')^n\to\mathbb{R}$, $\beta:\Theta\to\mathbb{R}$ et $\alpha_j:\Theta\to\mathbb{R}$ $(1\leq j\leq p)$, telles que pour tout $\theta\in\Theta$, $g(\theta) = -\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\leq i,j\leq p}\cdot\frac{\partial\beta}{\partial\theta}(\theta)$ et*

$$L_\theta(x_1,\cdots,x_n) = \exp\left(\beta(\theta) + b(x_1,\cdots,x_n) + \sum_{j=1}^d T_j(x_1,\cdots,x_n)\cdot\alpha_j(\theta)\right).$$

*Démonstration :* $\Longleftarrow$ On suppose donc le modèle exponentiel décrit dans le théorème. Si on dérive par rapport à $\theta$ un tel modèle, on obtient que pour $\mu$-presque tout $x\in(\Omega')^n$ :

$$\frac{\partial}{\partial\theta}(\log L_\theta(x)) = \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\leq i,j\leq p}\cdot\widehat{T} + \frac{\partial\beta}{\partial\theta}(\theta), \quad \text{pour tout } \theta\in\Theta. \tag{4}$$

En conséquence, comme $I(\theta) = \mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}(\log L_\theta(.))\right)\cdot {}^t\left(\frac{\partial}{\partial\theta}(\log L_\theta(.))\right)\right)$, on en déduit que :

$$I(\theta) = \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)\cdot\mathrm{cov}_\theta(\widehat{T})\cdot {}^t\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\leq i,j\leq p} \implies \mathrm{cov}_\theta(\widehat{T}) = \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\leq i,j\leq p}\cdot I(\theta)\cdot {}^t\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\leq i,j\leq p}$$

Par ailleurs, comme $\widehat{T}$ est un estimateur sans biais de $g(\theta)$ d'après la preuve de l'Inégalité de Cramer-Rao,

$$\mathbb{E}_\theta\left(\widehat{T}(.)\cdot {}^t\Big(\frac{\partial}{\partial\theta}(\log L_\theta(.))\Big)\right)=\frac{\partial g}{\partial\theta}(\theta)$$

et en utilisant (4) que l'on multiplie par $\Big(\frac{\partial}{\partial\theta}(\log L_\theta(.))\Big)$, on obtient :

$$\mathbb{E}_\theta\left(\Big(\frac{\partial}{\partial\theta}(\log L_\theta(.))\Big)\cdot {}^t\Big(\frac{\partial}{\partial\theta}(\log L_\theta(.))\Big)\right)=\mathbb{E}_\theta\left(\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p}\cdot\widehat{T}\cdot {}^t\Big(\frac{\partial}{\partial\theta}(\log L_\theta(.))\Big)\right)+\mathbb{E}_\theta\left(\frac{\partial\beta}{\partial\theta}(\theta)\cdot {}^t\Big(\frac{\partial}{\partial\theta}(\log L_\theta(.))\Big)\right),$$

et donc $I(\theta)=\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p}\cdot\frac{\partial g}{\partial\theta}(\theta)$. A l'aide de cette égalité, et en reprenant le calcul précédent, on en arrive à ce que :

$$\mathrm{cov}_\theta(\widehat{T})=\frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial\theta}(\theta),$$

donc $\widehat{T}$ atteint bien la borne de Cramer-Rao. De plus, grâce à (4),

$$\begin{aligned}\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}(\log L_\theta(x))\right) &= \mathbb{E}_\theta\left(\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p}\cdot\widehat{T}+\frac{\partial\beta}{\partial\theta}(\theta)\right)\\ \text{soit}\qquad 0 &= \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p}\cdot g(\theta)+\frac{\partial\beta}{\partial\theta}(\theta)\\ \text{et donc}\qquad g(\theta) &= -\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\le i,j\le p}\cdot\frac{\partial\beta}{\partial\theta}(\theta).\end{aligned}$$

$\Longrightarrow$ D'après la preuve de l'Inégalité de Cramer-Rao, si $\widehat{T}$ est un estimateur sans biais de $g(\theta)$ atteignant la borne de Cramer-Rao, alors

$$\mathrm{cov}_\theta(\widehat{T}-\frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta)=0.$$

Ainsi, pour tout $\theta\in\Theta$, il existe un ensemble $N_\theta\subset(\Omega')^n$ tel que $\mathbb{P}_\theta(N_\theta)=1$ et tel que pour tout $x\in N_\theta$, $\widehat{T}(x)-g(\theta)=\frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta(x)$. Par le même procédé que celui de la preuve de la nullité de l'information de Fisher pour une statistique libre, on peut déterminer un ensemble $N$ ne dépendant pas de $\theta$, tel que cette propriété soit également vraie, avec $\mu(N)=1$, ce qui revient à écrire que $\forall x\in N$,

$$I(\theta)\cdot\left(\frac{\partial g}{\partial\theta}(\theta)\right)^{-1}\cdot\Big(\widehat{T}(x)-g(\theta)\Big)=\frac{\partial}{\partial\theta}(\log L_\theta(x)),\quad\text{pour tout }\theta\in\Theta.$$

Alors en intégrant par rapport à $\theta$, et en notant $\begin{cases}\alpha(\theta)\text{ le vecteur colonne "intégrant" } I(\theta)\cdot\left(\frac{\partial g}{\partial\theta}(\theta)\right)^{-1}\\ \beta(\theta)\text{ la fonction "intégrant" } -I(\theta)\cdot\left(\frac{\partial g}{\partial\theta}(\theta)\right)^{-1}\cdot g(\theta)\\ b(x)\text{ une fonction ne dépendant pas de }\theta\end{cases}$

on a $\log L_\theta(x)=\alpha(\theta)\cdot\widehat{T}(x)+\beta(\theta)+b(x)$, d'où l'écriture de la vraisemblance sous forme d'un modèle exponentiel, et on retrouve l'expression de $g(\theta)$ par le même raisonnement que plus haut. ■

**Corollaire 4.** *A l'inverse, si l'on dispose d'un modèle exponentiel régulier (2), alors il n'existe qu'une seule fonction (à une transformation affine près) du paramètre pouvant être estimer efficacement, il s'agit de* $g(\theta)=-\frac{1}{n}\cdot\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\le i,j\le p}\cdot\frac{\partial\beta}{\partial\theta}(\theta)$ *(noter que cette fonction semble dépendre de* $n$ *; dans le cas de v.a.i.i.d. ce n'est pas le cas). L'estimateur est alors :* $\widehat{T}=\frac{1}{n}\cdot(a_1(X_1,\dots,X_n),\dots,a_p(X_1,\dots,X_n))$ *et sa matrice de covariance minimale est donnée par sa borne de Cramer-Rao, soit :*

$$cov_\theta(\widehat{T})=\frac{1}{n}\cdot\frac{\partial g}{\partial\theta}(\theta)\cdot\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\le i,j\le d}.$$

## 2.5 Estimateur des moments

La méthode des moments consiste à utiliser la convergence des moments empiriques vers les moment théorique pour identifier le paramètre. Dans le cas des familles exponentielles l'estimateur des moments coïncide avec l'estimateur du maximum de vraisemblance étudié à la section suivante.

### 2.5.1 Méthode des moments

Soit $X_1, \cdots, X_n$ un échantillon de loi $\mathbb{P}_\theta$, pour $\theta \in \Theta$. La méthode des moments consiste à estimer $\theta$ en résolvant un système d'équations

$$\frac{1}{n}\sum_{i=1}^{n} f_j(X_i) = \mathbb{E}_\theta f_j(X),\ j = 1, \cdots, k,$$

pour des fonctions $f_1, \cdots, f_k$ données. Souvent $f_j(x) = x^j$. Notons $f = (f_1, \cdots, f_k)$ et soit $e : \Theta \to \mathbb{R}^k$ le vecteur des espérances $e(\theta) = \mathbb{P}_\theta f$. L'estimateur des moments $\hat{\theta}_n$ résoud le système d'équations

$$\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) = e(\theta) := \mathbb{P}_\theta f.$$

Il est nécessaire que le vecteur $\mathbb{P}_n f$ soit dans l'image de la fonction $e$. Si $e$ est bijective l'estimateur des moments est uniquement déterminé par $\hat{\theta}_n = e^{-1}(\mathbb{P}_n f)$ et si $e^{-1}$ est différentiable, par la delta-méthode, on aura, en notant $\theta_0$ le vrai paramètre :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, {e'_{\theta_0}}^{-1} \mathbb{P}_{\theta_0} f f^T \left({e'_{\theta_0}}^{-1}\right)^T\right).$$

Pour établir ce résultat on commence par établir deux lemmes, qui impliquent le théorème d'inversion locale. En effet, si $f$ est une fonction d'une variable réelle définie sur un intervalle $I \subset \mathbb{R}$, si $f$ est dérivable sur $I$ et si sa dérivée ne s'annule pas, alors $f$ est strictement monotone, donc bijective de $I$ sur $f(I)$ et

$$(f^{-1})' = \frac{1}{f'(f^{-1}(x))}.$$

Le théorème d'inversion locale est la généralisation de cette propriété aux fonctions de plusieurs variables.

**Lemme 8.** *Soit $\Theta \subset \mathbb{R}^k$ et soit $e$ une bijection de $\Theta$ dans $\mathbb{R}^k$ différentiable en $\theta_0$ avec une différentielle inversible. Alors, si la fonction inverse $e^{-1}$ définie sur l'image de $e$ est continue en $e(\theta_0)$, elle sera aussi différentiable en $e(\theta_0)$.*

**Preuve** Notons $\eta = e(\theta_0)$ et $\Delta h = e^{-1}(\eta + h) - e^{-1}(\eta)$. Comme $e^{-1}$ est continue en $\eta$ on a $\Delta h \to 0$ quand $h \to 0$. Ainsi,quand $h \to 0$ :

$$\begin{aligned} &\eta + h = e e^{-1}(\eta + h) = e(\Delta h + \theta_0) = e(\theta_0) + e'_{\theta_0}(\Delta h) + o(\|\Delta h\|) \\ &\Leftrightarrow e'_{\theta_0}(\Delta h) = h + o(\|\Delta h\|). \end{aligned}$$

quand $h \to 0$. Ainsi, par continuité de l'inverse de $e'_{\theta_0}$, cela implique que $\Delta h = {e'_{\theta_0}}^{-1}(h) + o(\|\Delta h\|)$. En particulier $(\|\Delta h\|(1 + o(1)) \leq \|{e'_{\theta_0}}^{-1}(h)\| = O(\|h\|)$. Finalement, comme $e'_{\theta_0}(\Delta h) = h + o(\|\Delta h\|)$, on obtient que $\Delta h = {e'_{\theta_0}}^{-1}(h) + o(\|h\|)$.

**Lemme 9.** *Soit $e : \Theta \to \mathbb{R}^k$, différentiable en un voisinage de $\theta_0$, et continuement différentiable en $\theta_0$ avec une différentielle inversible. Alors, $e$ l'image par $e$ d'un voisinage $U$ de $\theta_0$, suffisamment petit dans un ouvert $V$ et $e^{-1} : V \to U$ est bien définie et continue.*

**Preuve** Par hypothèse, $e'_\theta \to A^{-1} := e'_{\theta_0}$ quand $\theta \to \theta_0$. Ainsi $\|I - A e'_\theta\| \leq \frac{1}{2}$ for tout $\theta$ dans un voisinage $U$ de $\theta_0$, suffisamment petit. Soit un point $\eta_1 = e(\theta_1)$ de $V = e(U)$, il existe alors $\varepsilon > 0$ tel que $\overline{B}(\theta_1, \varepsilon) \subset U$. Fixons alors un point $\eta$ avec $\|\eta - \eta_1\| = \delta := \frac{1}{2}\|A\|^{-1}\varepsilon$, on va montrer que $\eta = e(\theta)$ pour $\theta \in \overline{B}(\theta_1, \varepsilon)$, ainsi tout point $\eta \in B(\eta, \varepsilon)$ aura un antécédent dans $\overline{B}(\theta_1, \varepsilon)$. Si $e$ est bijective sur $U$ cet antécédent sera unique ainsi $V$ sera ouvert et $e^{-1}$ sera continue en $\eta_1$.

Soit $\phi(\theta) = \theta + A(\eta - e(\theta))$, comme la norme de la dérivée $\phi'_\theta = I - Ae'_\theta$ est bornée par $\frac{1}{2}$ $\forall \theta \in U$. De plus, si $\|\theta - \theta_1\| \leq \varepsilon$ :

$$\|\phi(\theta) - \theta_1\| \leq \|\phi(\theta) - \phi(\theta_1)\| + \|\phi(\theta_1) - \theta_1\| \leq \frac{1}{2}\|\theta - \theta_1\| + \|A\|\|\eta - \eta_1\| < \varepsilon.$$

Ainsi, $\phi$ est une application de $\overline{B}(\theta_1, \varepsilon)$ dans $\overline{B}(\theta_1, \varepsilon)$. Comme $\phi$ est une contraction, elle a un point fixe $\theta \in \overline{B}(\theta_1, \varepsilon) : \theta = \phi(\theta)$ et par définition de $\phi$ $e(\theta) = \eta$. Tout autre point $\tilde{\theta}$ avec $e(\tilde{\theta}) = \eta$ sera aussi un point fixe de $\phi$ Dans ce cas la différence $\tilde{\theta} - \theta = \phi(\tilde{\theta}) - \phi(\theta)$ aura une norme majorée par $\frac{1}{2}\|\tilde{\theta} - \theta\|$ ce qui n'est possible que si $\tilde{\theta} = \theta$, ainsi $e$ est bijective sur $U$.

On peut maintenant établir le théorème des propriétés asymptotiques de l'estimateur des moments.

**Théorème 21.** *Supposons que $e(\theta) = \mathbb{P}_\theta f$ est bijective sur ensemble ouvert $\Theta \subset \mathbb{R}^k$ et continuement différentiable en $\theta_0$, avec une dérivée $e'_{\theta_0}$ inversible. On suppose de plus que $\mathbb{P}_{\theta_0}\|f\|^2 < \infty$. Alors, l'estimateur des moments $\hat{\theta}_n$ existe avec probabilité tendant vers 1 et satisfait :*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, {e'_{\theta_0}}^{-1}\mathbb{P}ff^T\left({e'_{\theta_0}}^{-1}\right)^T\right).$$

**Preuve** Par le théorème d'inversion locale, il existe un ouvert $U$ de $\theta_0$ et un ouvert $V$ de $\mathbb{P}_{\theta_0}f$ tel que $e : U \to V$ est une bijection différentiable, d'inverse différentiable $e^{-1} : V \to U$. L'estimateur des moments $\hat{\theta}_n = e^{-1}(\mathbb{P}_n f)$ existe dès que $\mathbb{P}_n f \in V$, ce qui arrive avec probabilité qui tend vers 1 par la loi des grands nombres. La convergence en loi est alors une application immédiate de la delta-méthode.

**Famille exponentielle** L'estimateur des moments pour une famille exponentielle se confond avec l'estimateur du maximum de vraisemblance. On étudiera plus en détail les propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans un cadre plus général, mais ici les preuves sont particulièrement faciles. On rappelle que la densité d'une famille exponentielle s'écrit sous la forme

$$L_\theta(x_1, \cdots, x_n) = \exp\left(\beta(\theta) + b(x_1, \cdots, x_n) + \sum_{j=1}^{k} a_j(x_1, \cdots, x_n) \cdot \alpha_j(\theta)\right) := c(\theta)h(x)e^{\alpha(\theta)^T a(x)},$$

avec $x = (x_1, \cdots, x_n)$, $c(\theta) = \exp(\beta(\theta))$, $h(x) = \exp(b(x))$, $\alpha(\theta)^T = (\alpha_1(\theta), \cdots, \alpha_k(\theta))$ et $a(x)^T = (a_1(x), \cdots, a_k(x))$.

Supposons dans un premier temps que $\alpha(\theta) = \theta$, alors en remarquant que la vraisemblance est la composée de fonctions analytiques, on aura :

**Lemme 10.** *Soit $\Theta \subset \left\{\theta \in \mathbb{R}^k : c(\theta)^{-1} := \int h(x)e^{\theta^T a(x)}d\mu(x) < \infty\right\}$, alors la fonction : $\theta \mapsto \int h(x)e^{\theta^T a(x)}d\mu(x)$ est analytique sur l'ensemble $\{\theta \in \mathbb{C}^k,\ Re(\theta) \in \mathring{\Theta}\}$. Ses dérivées peuvent être calculées en différenciant sous le signe intégrale :*

$$\frac{\partial^p \int h(x)e^{\theta^T a(x)}d\mu(x)}{\partial\theta_1^{i_1}\cdots\theta_k^{i_k}} = \int h(x)a_1(x)^{i_1}\cdots a_k(x)^{i_k}d\mu(x),$$

*pour tout entier $p > 0$ et $i_1 + \cdots + i_k = p$.*

Ce lemme implique que la log-vraisemblance $l_\theta(x) = \log L_\theta(x)$ peut être différenciée elle aussi une infinité de fois par rapport à $\theta$. Le vecteur des dérivées partielles (la fonction score) satisfait :

$$\dot{l}_\theta(x) = \frac{\dot{c}}{c}(\theta) + a(x) = a(x) - \mathbb{E}_\theta a(X),$$

en remarquant que $\int L_\theta d\mu = 1$ et que

$$0 = \frac{\partial}{\partial\theta_i}\int L_\theta d\mu = \int \frac{\partial c(\theta)}{\partial\theta_i}h(x)e^{\theta^T a(x)}d\mu(x) + \int c(\theta)h(x)a_i(x)e^{\theta^T a(x)}d\mu(x).$$

et qu'ainsi $\frac{\dot{c}}{c}(\theta) = E_\theta a(X)$. Il s'ensuit que les équations du maximum de vraisemblance $\sum_{i=1}^n \dot{l}_\theta(X_i) = 0$ se réduisent à

$$\frac{1}{n}\sum_{i=1}^{n} a(X_i) = \mathbb{E}_\theta a(X)$$

et dans le cas des modèles de la famille exponentielle, l'estimateur du maximum de vraisemblance est un estimateur des moments. Ses propriétés asymptotiques dépendent de la fonction $e(\theta) = E_\theta a(X)$. Si on différencie $E_\theta(X)$ sous la racine :

$$\begin{aligned} e'_\theta &= \left( \int \frac{\partial c(\theta)}{\partial \theta_i} h(x) a_j(x) e^{\theta^T a(x)} d\mu(x) + \int c(\theta) h(x) a_i(x) a_j(x) e^{\theta^T a(x)} d\mu(x) \right)_{1 \leq i,j \leq k} \\ &= \left( \mathbb{E}_\theta(a_i(X) a_j(X)) - \mathbb{E}_\theta(a_i(X)) \mathbb{E}_\theta(a_j(X)) \right)_{1 \leq i,j \leq k} \\ &= Cov_\theta(a(X)) \end{aligned}$$

Si la famille exponentielle est correctement paramétrisé alors l'intérieur de $\Theta$ est non vide, aucune combinaison linéaire $\sum_{j=1}^k \lambda_j a_j(X)$ n'est constante avec probabilité 1 et $Cov_\theta(a(X))$ sera définie positive. Ainsi $e(\theta)$ sera une fonction bijective et il existera au plus une solution à l'équation des moments. D'aprés l'expression de $\dot{l}(\theta)$, la matrice $-ne'_\theta$ est la matrice hessienne de $\sum_{i=1}^n l_\theta(X_i)$ et la solution de l'équation des moments sera le maximum de la vraisemblance. On aura de plus :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, {e'_\theta}^{-1} Cov_{\theta_0} a(X) \left({e'_\theta}^{-1}\right)^T\right) = \mathcal{N}\left(0, (Cov_{\theta_0} a(X))^{-1}\right).$$

On peut maintenant généraliser à $\alpha(\theta)$, une bijection continuement différentiable et expliciter la matrice de variance-covariance asymptotique :

**Théorème 22.** *Soit $\Theta \subset \mathbb{R}^k$ un ouvert d'intérieur non vide et $\alpha : \Theta \to \mathbb{R}^k$ une bijection continuement différentiable, avec une dérivée inversible. Soit la famille exponentielle régulière qui correspond à ce $\alpha$, alors les équations du maximum de vraisemblance ont une unique solution $\hat{\theta}_n$ avec une probabilité qui tend vers* 1 *et en notant $I_{\theta_0}$ la matrice d'information de Fisher, on aura :*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I_{\theta_0}^{-1}\right)$$

**Preuve** Par le théorème d'inversion locale, l'image de $\alpha$ est ouverte et l'application inverse $\alpha^{-1}$ est différentiable sur cet image. La delta-méthode assure donc la normalité asymptotique de l'estimateur. Il suffit donc de calculer la matrice de variance-covariance asymptotique qui est égale à :

$$\alpha'(\theta_0)^{-1} \left(Cov_{\theta_0} a(X)\right)^{-1} \left(\alpha'(\theta_0)^{-1}\right)^T.$$

La fonction score du modèle est égale à

$$\dot{l}_\theta(x) = \frac{\dot{c}}{c}(\theta) + \left(\alpha'(\theta)\right)^T a(x).$$

Cette fonction de score a une moyenne nulle, et peut être réécrite :

$$\dot{l}_\theta(x) = \left(\alpha'(\theta)\right)^T \left(a(x) - \frac{\dot{c}}{c}(\theta)\right)$$

et la matrice d'information de Fisher vaudra $I_\theta = \left(\alpha'(\theta)\right)^T Cov_\theta a(X) \alpha'(\theta)$. Ainsi

$$I_{\theta_0}^{-1} = \alpha'(\theta_0)^{-1} \left(Cov_{\theta_0} a(X)\right)^{-1} \left(\alpha'(\theta_0)^{-1}\right)^T.$$

## 2.6 Estimateur du maximum de vraisemblance

Nous allons voir une méthode permettant d'obtenir aisément et dans la plupart des cas un estimateur possédant de très bonnes qualités... Par la suite on se place une nouvelle fois dans le cadre d'un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, avec $\Theta \subset \mathbb{R}^p$, dominé.

**Définition 28.** *Pour $(x_1, \ldots, x_n) \in (\Omega')^n$, soit $\theta \in \Theta \mapsto L_\theta(x_1, \ldots, x_n)$ la vraisemblance du modèle. On appelle estimateur du maximum de vraisemblance une statistique $\widehat{\theta}_n$ telle que pour $(X_1, \ldots, X_n)$ un n-échantillon quelconque du modèle :*

$$L_{\widehat{\theta}_n}(X_1, \cdots, X_n) = \sup_{\theta \in \Theta} L_\theta(X_1, \cdots, X_n).$$

**Remarque 7.** *Il n'y a pas de garantie de l'unicité d'un tel estimateur. Une méthode pour l'obtenir (mais pas toujours) est de rechercher un extremum local de $L_\theta$ sur $\Theta$, ce qui pourra être fait en annulant les dérivées partielles de $L_\theta$ par rapport à $\theta_i$. De même, il est clair que l'estimateur du maximum de vraisemblance pourra être également obtenu en maximisant le logarithme de la vraisemblance, appelé encore la log-vraisemblance. Enfin, si l'on désire estimer $g(\theta)$ avec $g$ une fonction bijective, alors $g(\widehat{\theta})$ sera l'estimateur du maximum de vraisemblance de $g(\theta)$.*

**Propriété 19.** *S'il existe une statistique exhaustive $\widehat{T}$ pour le modèle, alors $\widehat{\theta}$ est une fonction mesurable de $\widehat{T}$ pour tout $\theta \in \Theta$.*

*Démonstration :* Si $\widehat{T}$ est exhaustive, d'après le théorème de factorisation, la vraisemblance du modèle par rapport à la mesure dominante $P^*$ est $g_\theta(\widehat{T}(x_1,\dots,x_n))$ pour tout $\theta \in \Theta$ et $\mathbb{P}_\theta$-presque tout $(x_,\dots,x_n) \in (\Omega')^n$, ce qui revient à $P^*$-presque tout $(x_,\dots,x_n) \in (\Omega')^n$ par la même démonstration que celle de la nullité de l'information de Fisher d'une statistique libre. Ainsi, prendre l'argument maximal de $\theta \to L_\theta$ revient à prendre l'argument maximal de $\theta \to g_\theta(\widehat{T}(x_1,\dots,x_n))$, et $\widehat{\theta}$ sera donc une fonction de $\widehat{T}$. ■

**Propriété 20.** *On suppose que le modèle est régulier. Si on suppose qu'il existe un estimateur sans biais efficace de $\theta$ alors c'est l'estimateur du maximum de vraisemblance de $\theta$.*

*Démonstration :* D'après ce qui précède, si le modèle est régulier et que $\widehat{T}$ est un estimateur sans biais efficace de $\theta$ , alors le modèle est exponentiel et l'égalité (4) a encore lieu, soit pour tout $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}(\log L_\theta(x)) = \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p} \cdot \widehat{T} + \frac{\partial\beta}{\partial\theta}(\theta) \implies \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p} \cdot \mathbb{E}_\theta(\widehat{T}) + \frac{\partial\beta}{\partial\theta}(\theta) = 0.$$

Comme $\widehat{T}$ est un estimateur sans biais de $\theta$, on a donc $\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p} \cdot \theta + \frac{\partial\beta}{\partial\theta}(\theta) = 0$, pour tout $\theta \in \Theta$,

ce qui s'applique également à $\widehat{\theta}$ et donc :

$$\left(\frac{\partial\alpha_j}{\partial\theta_i}(\widehat{\theta})\right)_{1\le i,j\le p} \cdot \widehat{\theta} + \frac{\partial\beta}{\partial\theta}(\widehat{\theta}) = 0.$$

Mais d'après sa définition, le modèle étant régulier $\widehat{\theta}$ minimise la log-vraisemblance et annule donc sa dérivée, ce qui implique que :

$$\left(\frac{\partial\alpha_j}{\partial\theta_i}(\widehat{\theta})\right)_{1\le i,j\le p} \cdot \widehat{T} + \frac{\partial\beta}{\partial\theta}(\widehat{\theta}) = 0.$$

En conséquence, obtient :

$$\left(\frac{\partial\alpha_j}{\partial\theta_i}(\widehat{\theta})\right)_{1\le i,j\le p} \cdot \left(\widehat{T} - \widehat{\theta}\right) = 0 \implies \widehat{T} = \widehat{\theta},$$

car la matrice des dérivées des $\alpha_j$ est supposée de rang $d$. Enfin, l'unicité de $\widehat{\theta}$ est liée à l'écriture du modèle exponentiel. ■

Nous allons nous intéresser maintenant au comportement asymptotique de l'estimateur du maximum de vraisemblance (lorsqu'il existe), donc quand la taille $n$ de l'échantillon tend vers l'infini. Il est clair que pour chaque $n$ l'expression de l'estimateur est différente et, surtout, le modèle statistique change. Pour palier à cela, on se placera dans un "gros" modèle, $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$ (la dimension du paramètre reste constante) dans lequel un échantillon est une suite de v.a. Par ailleurs, on supposera désormais que **tout échantillon de ce modèle est constitué de v.a.i.i.d.**, et que $d\mathbb{P}_\theta^{\mathbb{N}} = (f_\theta \cdot d\mu)^{\otimes\mathbb{N}}$, le modèle étant dominé par la mesure $\mu$, et $f_\theta$ étant la densité de chaque $X_i$ par rapport à $\mu$.

**Théorème 23** (Convergence de l'estimateur du maximum de vraisemblance)**.** *On suppose le modèle paramétrique $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^d$ dominé par une mesure $\mu$ et régulier. On suppose en plus que le modèle est identifiable (au sens où $f_{\theta_1} = f_{\theta_2}$, $\mu$-presque partout, entraîne $\theta_1 = \theta_2$). Alors si la suite $(X_n)_{n\in\mathbb{N}}$ est issue du modèle avec pour paramètre $\theta_0 \subset \Theta$,*

$$\widehat{\theta}_n \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \theta_0 \quad \textit{pour la mesure } (f_{\theta_0} \cdot d\mu)^{\otimes\mathbb{N}}.$$

*Démonstration :* En premier lieu, pour $n$ fixé, il est clair que pour tout $\theta \in \Theta$ :

$$\log(L_\theta(x_1,\dots,x_n)) - \log(L_{\theta_0}(x_1,\dots,x_n)) = \sum_{i=1}^{n} \log\left(\frac{f_\theta(x_i)}{f_{\theta_0}(x_i)}\right).$$

Par ailleurs, pour tout $i \in \mathbb{N}$, les $X_i$ ont tous la même loi et pour $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right] &\leq \log\left(\mathbb{E}_{\theta_0}\left[\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right]\right) \quad \text{(Inégalité de Jensen pour la fonction } -\log) \\
&\leq \log\left(\mathbb{E}_\mu\left[f_\theta(X_i)\right]\right) \\
&\leq 0.
\end{aligned}$$

En fait, du fait que la fonction $-\log$ est strictement convexe, la borne 0 ne peut être atteinte que si $f_\theta = f_{\theta_0}$. Ainsi, avec la contrainte d'un modèle identifiable, dès que $\theta \neq \theta_0$, alors :

$$\mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right] < 0.$$

On peut appliquer la loi forte des grands nombres pour les variables aléatoires $\left(\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right)_{i\in\mathbb{N}}$ (qui sont bien i.i.d. et $\mathbb{L}^1$ car le modèle est régulier), et ainsi :

$$\begin{aligned}
\frac{1}{n}\left(\log(L_\theta(X_1,\dots,X_n)) - \log(L_{\theta_0}(X_1,\dots,X_n))\right) &= \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right) \\
&\xrightarrow[n\to+\infty]{p.s.} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right] < 0,
\end{aligned}$$

la convergence presque sûre ayant lieu pour la mesure $(f_{\theta_0}\cdot d\mu)^{\otimes\mathbb{N}}$. Considérons maintenant pour tout $\varepsilon > 0$ une famille dénombrable $(\theta_i^{(\varepsilon)})_{i\in I}$ dense sur la sphère de centre $\theta_0$ et de rayon $\varepsilon$. Du fait du caractère dénombrable de cette famille, pour tout $\varepsilon > 0$, il existe $n_\varepsilon$ tel que pour tout $n \geq n_\varepsilon$, pour tout $i \in I$ :

$$\log(L_{\theta_i^{(\varepsilon)}}(X_1,\dots,X_n)) < \log(L_{\theta_0}(X_1,\dots,X_n)) \quad \text{p.s. pour la mesure } (f_{\theta_0}\cdot d\mu)^{\otimes\mathbb{N}}.$$

Comme le modèle est régulier, pour tout $n \in \mathbb{N}^*$, la log-vraisemblance de $X_1,\dots,X_n$ est continue sur $\Theta$. De plus pour tout $n$ elle atteint son unique maximum en $\theta_0$. En conséquence, pour $n \geq n_\varepsilon$, $\widehat{\theta}_n$ sera à l'intérieur de la boule de centre $\theta_0$ et de rayon $\varepsilon$ (toujours p.s. pour la mesure $(f_{\theta_0}\cdot d\mu)^{\otimes\mathbb{N}}$). Le raisonnement étant vrai pour tout $\varepsilon > 0$, le théorème s'en déduit. ■

Le théorème suivant, est un cas particulier du théorème 26 démontré dans la section suivante.

**Théorème 24** (Normalité asymptotique de l'estimateur du maximum de vraisemblance)**.** *On suppose le modèle paramétrique $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, (f_\theta\cdot d\mu)^{\otimes\mathbb{N}}, \theta\in\Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure $\mu$ et régulier. On suppose en plus que le modèle est identifiable et que, pour tout $x \in \mathbb{R}^d$, la fonction $\theta \in \Theta \mapsto L_\theta(x)$ est de classe $\mathcal{C}^2(\Theta)$. Alors si la suite $(X_n)_{n\in\mathbb{N}}$ est issue du modèle avec pour paramètre $\theta_0 \subset \Theta$ :*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_d(0, I_1^{-1}(\theta_0)),$$

*où $I_1(\theta)$ est la matrice de Fisher de taille $p$ (supposée inversible) pour la variable $X_1$.*

**Remarque 8.** *Sous ces hypothèses, l'estimateur du maximum de vraisemblance est asymptotiquement sans biais et efficace. Cependant, à n fixé, il peut avoir un biais et ne pas être un estimateur efficace.*

## 2.7 M-estimateur

Il s'agit ici de généraliser la méthode de l'estimateur du maximum de vraisemblance.

### 2.7.1 Introduction

Supposons que nous voulons estimé un paramètre $\theta$ relié à la loi de probabilité des observations $(X_1, \cdots, X_n)$. La méthode pour trouver un tel estimateur est de maximiser (ou bien minimiser) une fonction critère de la forme

$$\theta \mapsto M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i)$$

Ici $m_\theta : \mathbb{R}^p \to \overline{\mathbb{R}}$ sont des fonctions connues. Un paramètre $\theta$ miximisant (ou minimisant) $M_n(\theta)$ est appelé un M-estimateur. Souvent, on cherche cette valeur annulant une différentielle, ainsi le nom M-estimateur est aussi utiliser pour des équations du type

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi_\theta(X_i) = 0 \tag{5}$$

avec $\phi_\theta$ des applications vectorielles connues. L'équation (5) n'est pas obligée de correspondre à un maximum ou un minimum, il serait plus adapté d'appeler alors l'estimateur correspondant un Z-estimateur (pour zéro) mais le nom M-estimateur est plus répandu. Par moment on ne dispose pas de solution exacte au problème de minimisation, mais une minimisation approximative peut suffir si celle-ci est quand même suffisamment précise. On va introduire quelques notations pour la suite :

— On note $\mathbb{P}$ la loi marginale des observations i.i.d. $X_1, \cdots, X_n$.
— On écrit $\mathbb{P}f := \int f d\mathbb{P} = \mathbb{E}f(X)$.
— On écrit $\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ la moyenne empirique.
— $\mathbb{P}_n$ est donc la loi empirique qui met une masse $\frac{1}{n}$ à toute observations $X_i$.

Avec ces notations, on aura $M_n = \mathbb{P}_n m_\theta$ et $\Psi_n(\theta) = \mathbb{P}_n \psi_\theta$. On va aussi noter

$$\mathbb{G}f := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - Pf)$$

le processus empirique en $f$.

**Exemple : estimateur du maximum de vraisemblance** Admetttons que $(X_1, \cdots, X_n)$ sont i.i.d. et leur loi à pour densité $f_{\theta^0}$. L'estimateur du maximum de vraisemblance $\hat{\theta}_n$ maximise :

$$\theta \mapsto M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i).$$

En machine learning, on minimise plus souvent une fonctionnelle, on remarquera donc que l'estimateur de maximum de vraisemblance $\hat{\theta}_n$ minimise aussi :

$$\theta \mapsto M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\log f_\theta(X_i).$$

Si le modèle est régulier, on cherche l'estimateur du maximum de vraisemblance en tant que solution de l'annulation de la dérivée de la log-vraisemblance :

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{f}_\theta}{f_\theta}(X_i) = 0$$

Cependant, si le modèle n'est pas régulier, la formulation du M-estimateur reste valide, par exemple si $X_1, \cdots, X_n$ sont i.i.d., $X_i \sim U_{[0,\theta]}$, alors on peut maximiser

$$\theta \mapsto M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(\mathbb{I}_{[0,\theta]}(x_i) - \log\theta), \text{ avec } \log(0) = -\infty.$$

**Exemple : Estimateur de location** Soit $X_1, \cdots, X_n$ i.i.d. réelles, on veut estimer une "location", terme vague qui peut désigner la moyenne ou bien la médiane et bien d'autre quantité. La moyenne et la médiane sont toutes deux estimables avec des Z-estimateur car ils vérifient les équations :

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

pour la moyenne et

$$\sum_{i=1}^{n} sign(X_i - \theta) = 0, \text{ où } sign(x) = -1 \text{ si } x < 0, sign(x) = 0 \text{ si } x = 0 \text{ et } sign(x) = 1 \text{ si } x > 0$$

pour la médiane. Les deux sont solutions d'une équation du type :

$$\sum_{i=1}^{n}\psi(X_i - \theta) = 0,$$

si toutes les observations $X_i$ sont translaté par une constante $\alpha$, alors l'estimateur $\hat{\theta} + \alpha$ résoud l'équation $\sum_{i=1}^{n}\psi(X_i + \alpha - \theta) = 0$, si $\theta$ résoud l'équation initiale. Un choix populaire pour la fonction $\psi$ est l'estimateur de Huber

$$\psi(x) = \begin{cases} -k \text{ si } x \leq -k \\ x \text{ si } |x| \leq k \\ k \text{ si } x \geq k \end{cases}$$

cet estimateur réalise un compromis entre la moyenne (proche pour $k$ grand) et la médiane (proche pour $k$ petit).

Un autre exemple sont les quantiles $\theta(p)$ tels que à peu près $pn$ observations soient plus petites que $\theta(p)$ et $(1-p)n$ plus grandes que $\theta(p)$. On peut définir l'estimateur quantile $p$, tout paramètre $\hat{\theta}(p)$ qui résoud les inégalités :

$$-1 < \sum_{i=1}^{n} p\mathbb{I}_{]\theta(p),\infty]}(X_i) - (1-p)\mathbb{I}_{]-\infty,\theta(p)[}(X_i) < 1.$$

En effet on aura pour un bon $\theta(p)$, $\sum_{i=1}^{n} p\mathbb{I}_{]\theta(p),\infty]}(X_i) \simeq \sum_{i=1}^{n}(1-p)\mathbb{I}_{]-\infty,\theta(p)[}(X_i) \simeq np(1-p)$. C'est approximativement un Z-estimateur avec $\phi(x) = p$, si $x > 0$, 0 si $x = 0$ et $\phi(x) = 1 - p$ si $x < 0$. Sauf dans des cas exeptionnels, on ne peut pas annulé exactement cette somme. Si il n'y a pas d'observations ex-aequo et $p > 0$, alors tous les sauts sont d'amplitude strictement plus petite que 1. Si il y a des observations ex-aequo, il peut être nécessaire d'augmenter les constantes $-1$ et 1 pour assurer l'existence de solutions.

Tous ces estimateurs peuvent être vu comme des M-estimateurs qui minimise $\sum_{i=1}^{n} m(X_i - \theta)$, en "intégrant" les équations. Avec $m(x) = x^2$ pour la moyenne, $m(x) = |x|$ pour la médiane, $m(x) = x^2\mathbb{I}_{[-k,k]}(x) + (2k|x| - k^2)\mathbb{I}_{]-\infty,-k[\cup]k,\infty[}(x)$ pour l'estimateur d'Huber et $m(x) = px^+ + (1-p)x^-$ pour l'estimateur des quantiles.

**Exemple : estimateur des moindres carrés** Admetttons que $((X_1, Y_1), \cdots, (X_n, Y_n))$ sont i.i.d. , $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$ et vérifient l'equations de régression linéaire :

$$Y = \theta^T X + \varepsilon$$

où $\varepsilon$ est une variable aléatoire centrée, indépendante de $X$ et de carré intégrable. L'estimateur des moindres carrés $\hat{\theta}_n$ de $\theta$ est alors celui qui minimise :

$$\theta \mapsto M_n(\theta) = \sum_{i=1}^{n}(Y_i - \theta^T X_i)^2 = \sum_{i=1}^{n} m(Y_i - \theta^T X_i),$$

avec $m(x) = x^2$. On peut aussi généraliser cet estimateur en utilisant les autres fonctions $m$ : $m(x) = |x|$, $m(x) = x^2\mathbb{I}_{[-k,k]}(x) + (2k|x| - k^2)\mathbb{I}_{]-\infty,-k[\cup]k,\infty[}(x)$ et $m(x) = px^+ + (1-p)x^-$.

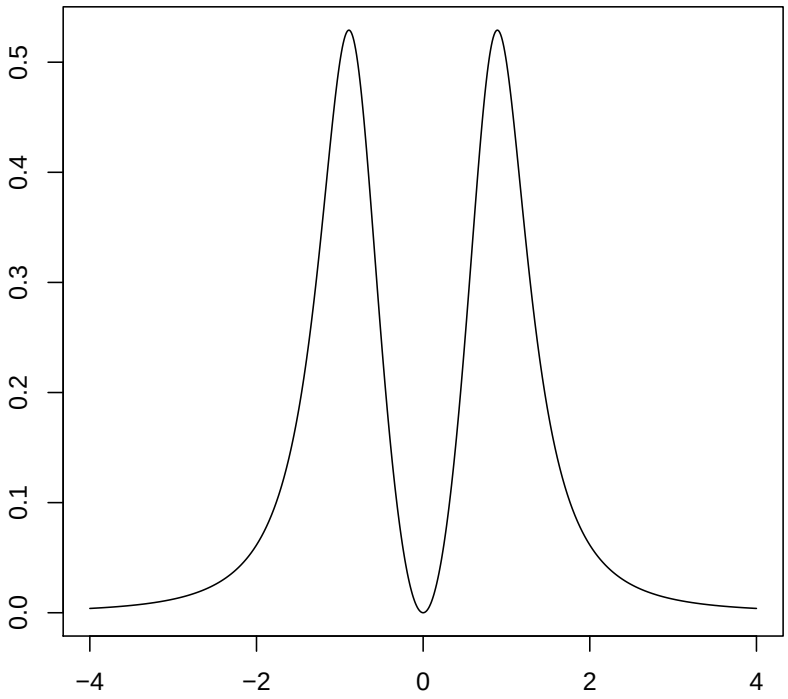


FIGURE 1 – Minimum mal séparé

### 2.7.2 Consistance

Il est important que l'estimateur converge vers la vrai valeur du paramètre $\theta_0$ lorsque le nombre d'observations $n$ converge vers l'infini. Si c'est le cas l'estimateur est dit asymptotiquement consistant. Par exemple la moyenne empirique $\bar{X}$ est asymptotiquement consistant pour la moyenne de la population $\theta_0 = E(X)$ si $E(X)$ existe. On suppose que l'ensemble des paramètres possibles $\Theta$ est un espace métrique, avec métrique $d$ et on veut prouver que On veut donc prouver que :

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{P} 0.$$

On suppose que le M-estimateur $\hat{\theta}_n$ minimise la fonction $M_n(\theta)$. Clairement, le comportement asymptotique de $\hat{\theta}_n$ dépend de du comportement asymptotique de la fonction $M_n(\theta)$. Pour une bonne normalisation il existera une fonction critère déterministe $M(\theta)$ telle que :

$$\forall \theta \ : \ M_n(\theta) \xrightarrow{P} M(\theta)$$

Par exemple, si $M_n(\theta)$ est une moyenne empirique $\mathbb{P}_n m_\theta$, alors la loi des grands nombres donnera $M(\theta) = Pm_\theta$ dès que cet espérance existe. Il semble raisonnable que le minimisateur $\hat{\theta}_n$ de $M_n(\theta)$ converge sous des conditions raisonnable vers $\theta_0$ le minimisateur de $M(\theta)$. Il faut quand même faire attention à ce que le problème de minimisation de $M(\theta)$ soit bien "séparé" et éviter ce genre de situation. La figure 1 donne un exemple d'une fonction qui a bien un minimum égal à 0, mais la droite d'équation $y = 0$ est aussi une asymptote. Il faut aussi faire attention que la valeur $\hat{\theta}_n$ dépend de la fonction $\theta \mapsto M_n(\theta)$ et une forme de convergence "fonctionnelle" de $M_n$ vers $M$ est nécessaire. Il y a plusieurs possibilité, la convergence uniforme en $\theta$ est une forme très forte, mais vérifiée par de nombreux modèles usuels.

On peut aussi facilement généraliser les M-estimateurs aux quasi-M-estimateurs c'est-à-dire $\hat{\theta}_n$ qui vérifient :

$$M_n(\hat{\theta}_n) \geq \sup_{\theta} M_n(\theta) - o_P(1).$$

On aura alors le théorème :

**Théorème 25.** *Soit $\Theta$ l'espace des paramètres possibles et $M_n$, $M$ des fonctions telles que pour tout $\varepsilon > 0$ :*

$$\begin{array}{l} \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p.s.} 0, \\ \inf_{\theta \,:\, \|\theta-\theta_0\|\geq\varepsilon} M(\theta) > M(\theta_0) \end{array}$$

*alors pour toute suite d'estimateur $\hat{\theta}_n$ telle que*

$$M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$$

*On aura*

$$\hat{\theta}_n \xrightarrow{P} \theta_0$$

**Preuve** Par la convergence uniforme et l'hypothèse $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$, on aura $M_n(\hat{\theta}_n) \geq M(\theta_0) - o_P(1)$, ainsi

$$M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \xrightarrow{P} 0.$$

De plus, par l'hypothèse $\inf_{\theta \,:\, \|\theta-\theta_0\|\geq\varepsilon} M(\theta) > M(\theta_0)$, pour tout $\varepsilon > 0$, il existe $\eta > 0$, tel que $M(\theta) < M(\theta_0) - \eta$, si $d(\theta, \theta_0) \geq \varepsilon$. Ainsi

$$\mathbb{P}(d(\theta_n, \theta_0) \geq \varepsilon) \leq \mathbb{P}(M(\theta_n) < M(\theta_0) - \eta) \xrightarrow{P} 0.$$

On peut aussi généraliser ce théorème aux Z-estimateur en notant qu'un zéro de $\Psi_n(\theta)$ minimise $\|\Psi_n(\theta)\|$.

**Conditions suffisantes pour le théorème** Pour un modèle régulier les hypothèses du théorème sont faciles à vérifiées. Pour obtenir l'hypothèse de la loi uniforme des grands nombres :

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p.s.} 0$$

une condition suffisante est que l'ensemble des paramètres possibles $\Theta$ soit compact, que la fonction $\theta \mapsto m_\theta(x)$ soit continue pour tout $x$ et qu'il existe une fonction $h$ intégrable qui domine $|m_\theta(x)|$ pour tout $\theta \in \Theta$.

**Preuve** Pour une boule ouverte $B$ de $\Theta$, notons $m^B = \sup_{\theta\in B} m_\theta$ et $m_B = \inf_{\theta\in B} m_\theta$. Par le théorème de convergence dominée $E[m^B - m_B] \to 0$ lorsque le diamètre de la boule tend vers 0. Pour $\varepsilon > 0$, soit $B^1, \cdots, B^k$ un recouvrement fini de $\Theta$ tel que $E[m^{B^i}(X) - m_{B^i}(X)] < \varepsilon$. Pour tout $\theta \in B^i$, on aura :

$$\begin{array}{l} M_n(\theta) - M(\theta) \leq \frac{1}{n} m^{B^i}(X_i) - E[m_{B^i}(X)] \leq \frac{1}{n} m^{B^i}(X_i) - E[m^{B^i}(X)] + \varepsilon \\ M_n(\theta) - M(\theta) \geq \frac{1}{n} m_{B^i}(X_i) - E[m^{B^i}(X)] \geq \frac{1}{n} m_{B^i}(X_i) - E[m_{B^i}(X)] - \varepsilon \end{array}$$

Ainsi,

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \leq \sup_{i\in\{1,\cdots,n\}} \max\left(\left|\frac{1}{n} m^{B^i}(X_i) - E[m^{B^i}(X)]\right|, \left|\frac{1}{n} m_{B^i}(X_i) - E[m_{B^i}(X)]\right|\right) + \varepsilon$$

Ainsi, presque-sûrement, $\forall \varepsilon > 0, \lim_{n\to\infty} \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| < \varepsilon$. ■

Pour obtenir la condition :

$$\inf_{\theta \,:\, \|\theta-\theta_0\|\geq\varepsilon} M(\theta) > M(\theta_0)$$

Il suffit que l'ensemble paramètres possibles $\Theta$ soit compact, que la fonction limite $M(\theta)$ soient une fonction continue, minimale en $\theta_0$ et que pour tout $\theta \neq \theta_0$, $M(\theta) \neq M(\theta_0)$. Cela sera le cas, si $M(\theta)$ est l'espérance de l'opposé de la log-vraisemblance, c'est-à-dire la distance de Kullback à une constante près :

$$M(\theta) = \int f_{\theta_0} \log\left(\frac{f_{\theta_0}}{f_\theta}\right)$$

### 2.7.3 Normalité asymptotique

Soit $X_1, \cdots, X_n$ un échantillon i.i.d. et soit une fonction critère de la forme :

$$\Psi_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \phi_\theta(X_i) = \mathbb{P}_n\psi_\theta,\ \Psi_\theta = \mathbb{P}\psi_\theta.$$

Supposons que l'estimateur $\hat{\theta}_n$ soit un zéro de $\Psi_n$ et converge en probabilité vers un zéro $\theta_0$ de $\Psi$. Comme $\hat{\theta}_n \xrightarrow{P} \theta_0$, on peut utiliser une développement limité autour de $\theta_0$. Supposons, pour simplifier, que $\theta$ soit unidimensionnel, alors pour tout $h$ fixé :

$$0 = \Psi_n(\theta_0 + h) = \Psi_n(\theta_0) + h\dot{\Psi}_n(\theta_0) + o(|h|)$$

En utilisant le lemme (6) des comparaisons stochastiques, on aura donc pour la suite aléatoire $\hat{\theta}_n$ :

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\dot{\Psi}_n(\theta_0) + o_P(|(\hat{\theta}_n - \theta_0)|).$$

Ainsi,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + o_P(|\hat{\theta}_n - \theta_0|)}.$$

Si $\mathbb{P}\psi_{\theta_0}^2$ est fini, alors le numérateur $-\sqrt{n}\Psi_n(\theta_0) = -\sqrt{n}\sum_{i=1}^{n}\psi_{\theta_0}(X_i)$ est asymptotiquement normal par le théorème de la limite central. Sa moyenne asymptotique est $\mathbb{P}\psi_{\theta_0} = \Psi(\theta_0) = 0$ et sa variance est $\mathbb{P}\psi_{\theta_0}^2$. Si on considère le dénominateur, part la loi des grands nombres $\dot{\Psi}_n(\theta_0) \xrightarrow{P} \mathbb{P}\dot{\psi}_{\theta_0}$. Le deuxième converge en probabilité vers 0. Le lemme de Slutsky donnera alors :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\mathbb{P}\psi_{\theta_0}^2}{(\mathbb{P}\dot{\psi}_{\theta_0})^2}\right).$$

Pour rendre ce cheminement rigoureux, il faut imposer des conditions de régularité, notamment pour que $\Psi_n$ soit $\mathcal{C}^1$ et on pourra étendre ce raisonnement aux paramètres de dimension plus grande que 1. Les applications critères sont alors $\Psi_n : \mathbb{R}^k \to \mathbb{R}^k$, $\dot{\Psi}_n(\theta_0)$ est une matrice $k \times k$ qui converge vers la matrice $\mathbb{P}\dot{\psi}_{\theta_0}$ et si cette matrice est inversible, la loi asymptotique devient :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, (\mathbb{P}\dot{\psi}_{\theta_0})^{-1}\mathbb{P}\psi_{\theta_0}\psi_{\theta_0}^T(\mathbb{P}\dot{\psi}_{\theta_0})^{-1}\right).$$

### 2.7.4 Condition classique pour la normalité asymptotique

Soit $X_1, \cdots, X_n$ un échantillon i.i.d. et soit une fonction critère de la forme :

$$\Psi_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \phi_\theta(X_i) = \mathbb{P}_n\psi_\theta,\ \Psi_\theta = \mathbb{P}\psi_\theta.$$

Supposons que l'estimateur $\hat{\theta}_n$ soit un zéro de $\Psi_n$ et converge en probabilité vers un zéro $\theta_0$ de $\Psi$.

**Hypothèses** On suppose que les observations $(X_1, \cdots, X_n)$ sont i.i.d. et que les hypothèses suivantes sont vérifiées :

- H1 Le M-estimateur $\hat{\theta}_n$ est consistant : $\hat{\theta}_n \xrightarrow{P} \theta_0$.
- H2 Il existe un voisinage $B$ du vrai paramètre $\theta_0$ tel que pour tout $\theta \in V$, et tout $x \in \mathbb{R}^d$, la fonction $\theta \mapsto \psi_\theta(x)$ soit continuement dérivable.
- H3 $\mathbb{P}\|\psi_{\theta_0}\|^2 < \infty$.
- H4 La matrice $\mathbb{P}\dot{\psi}_{\theta_0}$ est inversible.

On aura alors le théorème suivant :

**Théorème 26.** *Sous les hypothèses H1,...,H4, la d'estimateurs $\hat{\theta}_n$ vérifiera*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(\mathbb{P}\dot{\psi}_{\theta_0})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i) + o_P(1).$$

*En particulier la suite $\sqrt{n}(\hat{\theta}_n - \theta_0)$ est asymptotiquement normale de moyenne 0 et de matrice de variance-covariance $(\mathbb{P}\dot{\psi}_{\theta_0})^{-1}\mathbb{P}\psi_{\theta_0}\psi_{\theta_0}^T(\mathbb{P}\dot{\psi}_{\theta_0})^{-1}$.*

**Preuve** Par un développement de Taylor et en utilisant le lemme (6) des comparaisons stochastiques, on aura :

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + o_P(\|\hat{\theta}_n - \theta_0)\|)$$

Le premier terme à droite est la moyenne du vecteur aléatoire i.i.d. $\psi_{\theta_0}(X_i)$ qui a pour espérance $\mathbb{P}\psi_{\theta_0} = 0$.

Par le théorème de la limite centrale la suite $\sqrt{n}\Psi_n(\theta_0)$ converge en loi vers une gaussienne de moyenne 0 et de matrice de variance-covariance $\mathbb{P}\psi_{\theta_0}\psi_{\theta_0}^T$. Par la loi des grands nombres $\dot{\Psi}_n(\theta_0)$ converge vers la matrice $\mathbb{P}\dot{\psi}_{\theta_0}$. On aura donc

$$-\Psi_n(\theta_0) = \left(\mathbb{P}\dot{\psi}_{\theta_0} + o_P(1)\right)(\hat{\theta}_n - \theta_0)).$$

La probabilité que la matrice $\mathbb{P}\dot{\psi}_{\theta_0} + o_P(1)$ soit inversible tend vers 1. En multipliant l'équation précédente par $\sqrt{n}(\mathbb{P}\dot{\psi}_{\theta_0} + o_P(1))^{-1}$ on obtient le résultat annoncé.

Les hypothèses du théorème précédents sont restrictives. Pour le simple exemple de la médiane on a $\psi_\theta(x) = sign(x - \theta)$ n'est pas $\mathcal{C}^1$, ni même $\mathcal{C}^0$. Pourtant, sous de bonnes hypothèses, la mediane est asymptotiquement normale. Les propriétés des M-estimateurs peuvent être obtenu sous des conditions plus faibles. En effet, si le critère asymptotique est $\mathcal{C}^2$, on aura

$$\mathbb{P}m_\theta = \mathbb{P}m_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2).$$

C'est ce développement qui est utilisé dans le théorème suivant, cependant on a besoin de deux lemmes avancés (qui seront admis) sur les processus empiriques.

**Lemme 11.** *Pour tout $\theta$ dans un ensemble ouvert $\Theta \subset \mathbb{R}^k$ soit $x \mapsto m_\theta(x)$ une fonction mesurable telle que $\theta \mapsto m_\theta(x)$ soit différentiable en $\theta_0$ pour presque tout $x$. On notera $\dot{m}_{\theta_0}(x)$ cette différentielle. On suppose qu'il existe une fonction $\dot{m}$, avec $P\dot{m}^2 < \infty$, et un voisinage $B$ de $\theta_0$ tels que pour tout $\theta_1$ et $\theta_2$ dans $B$, on a*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)\|\theta_1 - \theta_2\|.$$

*Alors, si on note $\sqrt{n}(\mathbb{P}_n - \mathbb{P})f$ en $\mathbb{G}_n f$, le processus empirique évalué en $f$ et $\tilde{h}_n$ une suite aléatoire bornée en probabilité :*

$$\mathbb{G}_n\left[r_n\left(m_{\theta_0 + \frac{\tilde{h}_n}{r_n}} - m_{\theta_0}\right) - \tilde{h}_n^T \dot{m}_{\theta_0}\right] \xrightarrow{P} 0.$$

**Lemme 12.** *Pour tout $\theta$ dans un ensemble ouvert $\Theta \subset \mathbb{R}^k$ soit $x \mapsto m_\theta(x)$ une fonction mesurable telle que $\theta \mapsto m_\theta(x)$ soit différentiable en $\theta_0$ pour presque tout $x$. On notera $\dot{m}_{\theta_0}(x)$ cette différentielle. On suppose qu'il existe une fonction $\dot{m}$, avec $P\dot{m}^2 < \infty$, et un voisinage $B$ de $\theta_0$ tels que pour tout $\theta_1$ et $\theta_2$ dans $B$, on a*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)\|\theta_1 - \theta_2\|.$$

*De plus, on suppose que l'application $\theta \mapsto \mathbb{P}m_\theta$ admet un développement de Taylor à l'ordre deux :*

$$\mathbb{P}m_\theta = \mathbb{P}m_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2),$$

*avec une matrice $V_{\theta_0}$ inversible. Si $\hat{\theta}_n$ est un quasi-M-estimateur : $\mathbb{P}_n m_{\hat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - o_P(n^{-1})$ et $\hat{\theta}_n \xrightarrow{P} \theta_0$ alors :*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = O_P(1).$$

**Théorème 27.** *Pour tout $\theta$ dans un ensemble ouvert $\Theta \subset \mathbb{R}^k$ soit $x \mapsto m_\theta(x)$ une fonction mesurable telle que $\theta \mapsto m_\theta(x)$ soit différentiable en $\theta_0$ pour presque tout $x$. On notera $\dot{m}_{\theta_0}(x)$ cette différentielle. On suppose qu'il existe une fonction $\dot{m}$, avec $P\dot{m}^2 < \infty$, et un voisinage $B$ de $\theta_0$ tels que pour tout $\theta_1$ et $\theta_2$ dans $B$, on a*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)\|\theta_1 - \theta_2\|.$$

*De plus, on suppose que l'application $\theta \mapsto \mathbb{P}m_\theta$ admet un développement de Taylor à l'ordre deux :*

$$\mathbb{P}m_\theta = \mathbb{P}m_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2),$$

*avec une matrice $V_{\theta_0}$ inversible. Si $\hat{\theta}_n$ est un quasi-M-estimateur : $\mathbb{P}_n m_{\hat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - o_P(n^{-1})$ et $\hat{\theta}_n \xrightarrow{P} \theta_0$ alors :*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) + o_P(1)$$

*En particulier la suite $\sqrt{n}(\hat{\theta}_n - \theta_0)$ est asymptotiquement normale de moyenne 0 et de matrice de variance-covariance $V_{\theta_0}^{-1} \mathbb{P} \dot{m}_{\theta_0} \dot{m}_{\theta_0}^T V_{\theta_0}^{-1}$.*

**Preuve** Pour une fonction fixée $f$, on note $\sqrt{n}(\mathbb{P}_n - \mathbb{P})f$ en $\mathbb{G}_n f$, le processus empirique évalué en $f$. Si $\tilde{h}_n$ est une suite déterministe bornée

$$\mathbb{G}_n \left[ \sqrt{n} \left( m_{\theta_0 + \frac{\tilde{h}_n}{\sqrt{n}}} - m_{\theta_0} \right) - \tilde{h}_n^T \dot{m}_{\theta_0} \right] \xrightarrow{P} 0.$$

Car c'est une variable aléatoire de moyenne nulle dont la variance converge vers 0. En effet, par définition, pour toute fonction $f$ intégrable, $\mathbb{E}\mathbb{G}_n f = 0$, de plus, comme $\theta \mapsto m_\theta(x)$ est différentiable en $\theta_0$ pour presque tout $x$ :

$$\mathbb{P} \| \sqrt{n} \left( m_{\theta_0 + \frac{\tilde{h}_n}{\sqrt{n}}} - m_{\theta_0} \right) - \tilde{h}_n^T \dot{m}_{\theta_0} \|^2 = \mathbb{P} \| \tilde{h}_n^T \dot{m}_{\theta_0} + o(1) - \tilde{h}_n^T \dot{m}_{\theta_0} \|^2 \longrightarrow 0$$

Si la séquence $\tilde{h}_n$ est aléatoire, c'est une conséquence du lemme (11). Comme les hypothèses du théoreme permettent d'appliquer le lemme (12), la suite $\sqrt{n}(\hat{\theta}_n - \theta_0)$ est bornée en probabilité. Comme l'application $\theta \mapsto \mathbb{P} m_\theta$ est deux fois dérivable, on aura :

$$\begin{aligned} & n(\mathbb{P}_n - P) \left( m_{\theta_0 + \frac{\tilde{h}_n}{\sqrt{n}}} - m_{\theta_0} \right) = \sqrt{n}(\mathbb{P}_n - P) \tilde{h}_n^T \dot{m}_{\theta_0} + o_P(1) \\ \Leftrightarrow\ & n \mathbb{P}_n \left( m_{\theta_0 + \frac{\tilde{h}_n}{\sqrt{n}}} - m_{\theta_0} \right) = \tfrac{1}{2} \tilde{h}_n^T V_{\theta_0} \tilde{h}_n + \tilde{h}_n^T \mathbb{G}_n \dot{m}_{\theta_0} + o_P(1) \end{aligned}$$

Comme la suite $\hat{\theta}_n$ converge vers $\theta_0$ à la vitesse $\sqrt{n}$, l'égalité précédente est valide pour $\tilde{h}_n$ valant $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ ou bien $\tilde{h}_n = -V_{\theta_0}^{-1} \mathbb{G}_n \dot{m}_{\theta_0}$. On obtient donc les équations :

$$\begin{aligned} & n \mathbb{P}_n \left( m_{\theta_0 + \frac{\hat{h}_n}{\sqrt{n}}} - m_{\theta_0} \right) = \tfrac{1}{2} \hat{h}_n^T V_{\theta_0} \hat{h}_n + \hat{h}_n^T \mathbb{G}_n \dot{m}_{\theta_0} + o_P(1) \\ & \text{et} \\ & n \mathbb{P}_n \left( m_{\theta_0 + \frac{\hat{h}_n}{\sqrt{n}}} - m_{\theta_0} \right) = -\tfrac{1}{2} \mathbb{G}_n \dot{m}_{\theta_0}^T V_{\theta_0}^{-1} \mathbb{G}_n \dot{m}_{\theta_0} + o_P(1) \end{aligned}$$

Comme $\hat{\theta}_n$ est un M-estimateur (à un $o_P(1)$ près) le terme de gauche de la première équation est plus grand que celui de la deuxième équation, c'est donc aussi vrai pour les termes de droites. En prenant la différence des deux équations on obtiendra

$$\frac{1}{2} \left( \hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \dot{m}_{\theta_0} \right)^T V_{\theta_0} \left( \hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \dot{m}_{\theta_0} \right) + o_P(1) \geq 0$$

Comme la matrice $V_{\theta_0}$ est définie négative, ce n'est possible que si $\| \hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \dot{m}_{\theta_0} \|$ converge vers 0.

**Exemple (médiane)** Soit $X_1, \cdots, X_n$ des variables réelles i.i.d. L'estimateur de la médiane maximise la fonction critère $\theta \mapsto -\sum_{i=1}^n |X_i - \theta|$. Supposons que fonction de répartition $F$ des observations est dérivable en sa médiane $\theta_0$, avec pour dérivée strictement positive $f(\theta_0)$ et que l'ensemble de paramètres possibles $\Theta$ est compact. Alors l'estimateur de la médiane est asymptotiquement normale. Cela vient du théorème précédent appliqué à $m_\theta(x) = |x - \theta| - |x|$. En effet, on aura

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| = ||x - \theta_1| - |x| - (|x - \theta_2| - |x|)| \leq |\theta_1 - \theta_2|$$

car $|\theta_1 - \theta_2| = |\theta_1 - x - (\theta_2 - x)| \geq |\theta_1 - x| - |\theta_2 - x|$. Cette fonction est donc Lipschitz avec $\dot{m}(x) = 1$. De plus, l'application $\theta \mapsto m_\theta(x)$ est différentiable en $\theta_0$ sauf en $x = \theta_0$, avec $\dot{m}_{\theta_0}(x) = -sign(x - \theta_0)$. On aura de plus :

$\mathbb{P}m_\theta = \int_{-\infty}^{\infty} m_\theta(x)dF(x) = \int_{-\infty}^{\infty} |x-\theta| - |x|dF(x) =$
$\int_{-\infty}^{\theta} -x+\theta dF(x) + \int_{\theta}^{\infty} x-\theta dF(x) - \int_{-\infty}^{0} -xdF(x) - \int_0^{\infty} xdF(x) =$
$\int_{-\infty}^{0} -x+\theta dF(x) + \int_0^{\theta} -x+\theta dF(x) + \int_0^{\infty} x-\theta dF(x) + \int_\theta^0 x-\theta dF(x) - \int_{-\infty}^0 -xdF(x) - \int_0^\infty xdF(x) =$
$\int_{-\infty}^{0} \theta dF(x) + \int_0^{\theta} -x+\theta dF(x) + \int_0^{\infty} -\theta dF(x) + \int_\theta^0 x-\theta dF(x) =$
$\theta F(0) - \theta(1-F(0)) + 2\int_0^\theta \theta - xdF(x) = \theta F(0) - \theta(1-F(0)) + 2([(\theta-x)F(x)]_0^\theta + \int_0^\theta F(x)dx) =$
$\theta F(0) - \theta(1-F(0)) - 2\theta F(0) + 2\int_0^\theta F(x)dx = 2\int_0^\theta F(x)dx - \theta.$

Si $F$ est differentiable en $\theta_0$, $\mathbb{P}m_\theta$ admet un developpement de Taylor au voisinage de $\theta_0$ :

$$\mathbb{P}m_\theta = \mathbb{P}m_{\theta_0} + \frac{1}{2}(\theta-\theta_0)^2 2f(\theta_0) + o(|\theta-\theta_0|^2).$$

En notant $V_{\theta_0} = 2f(\theta_0)$, comme $\mathbb{P}\dot{m}^2_{theta_0} = \mathbb{E}1 = 1$, la variance asymptotique de l'estimateur de la médiane sera $\left(\frac{1}{2f(\theta_0)}\right)^2$.

## 2.8 Régions de confiance

En pratique, estimer un paramètre le plus souvent ne suffit pas. On aimerait connaître plus précisément quelle marge de sécurité on a sur la connaissance de ce paramètre.

**Définition 29.** *On se place dans le cadre d'un modèle paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$. Soit $\alpha \in ]0,1[$ un nombre fixé a priori. On appelle région de confiance du paramètre $\theta$ de niveau $1-\alpha$ un sous-ensemble aléatoire $R_{1-\alpha}$ inclus dans $\mathbb{R}^p$ et défini sur $((\Omega')^{\mathbb{N}}, \mathcal{A}'_{\mathbb{N}})$, tel que pour tout $\theta \in \Theta$, $\{(x_1,\dots,x_n) \in (\Omega')^{\mathbb{N}}, \theta \in R_{1-\alpha}(x_1,\dots,x_n)\} \in \mathcal{A}'_n$ et :*

$$\inf_{\theta\in\Theta} \{\mathbb{P}_\theta(\theta \in R_{1-\alpha})\} \geq 1-\alpha. \tag{6}$$

*Si un échantillon observé $(X_1(\omega),\dots,X_n(\omega))$ est connu, $R_{1-\alpha}(X_1(\omega),\dots,X_n(\omega))$ est appelé région de confiance observé. Dans le cas où le paramètre est un réel ($p=1$), on pourra obtenir un intervalle de confiance.*

Comment déterminer une région de confiance ? En premier lieu, il est clair que pour tout $\alpha \in ]0,1[$, $R_{1-\alpha} \subset \Theta$ (en général, on choisit $\alpha$ proche de 0, et en particulier $\alpha = 0.05$ est très souvent utilisé). Une démarche possible pour la construction de région de confiance est la suivante : naturellement, on désirerait utiliser un estimateur $\widehat{T}$ convergent de $\theta$, mais sa loi dépend en général de $\theta$ ce qui rend difficile (à part quelques exceptions) son utilisation directe. On préférera donc utiliser ce que l'on appelle une fonction pivotale $\pi(\widehat{T},\theta)$, qui est une fonction mesurable d'un estimateur et de $\theta$ et qui est une statistique libre. On essayera alors d'écrire la propriété (6) sous la forme

$$\inf_{\theta\in\Theta} \left\{\mathbb{P}_\theta(\pi(\widehat{T},\theta) \in C_\alpha)\right\} \geq 1-\alpha,$$

où $C_\alpha$ est une région déterministe. Aussi pourra-t-on ensuite construire la région de confiance en fonction des quantiles (souvent à $\alpha/2$ et $1-\alpha/2$) de la loi de la fonction pivotale.

**Exemple 10.** *Si le modèle est régulier, sous les conditions du théorème de normalité asymptotique du maximum de vraisemblance, on peut également montrer (théorème de Sltuski) que*

$$\pi(\widehat{\theta}_n,\theta_0) = \sqrt{n}\cdot(I_1(\widehat{\theta}_n))^{1/2}\cdot\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_d(0,I_p),$$

*où $I_d$ est la matrice identité de taille $p$ et $(I_1(\theta))^{1/2}\cdot(I_1(\theta))^{1/2} = I_1(\theta)$ pour tout $\theta \in \Theta$. Ainsi, si $n$ est grand, on pourra assimiler la loi de $\pi(\widehat{\theta}_n,\theta_0)$ avec la loi normale centrée réduite multidimensionnelle. Or si $Z \sim \mathcal{N}_p(0,I_p)$, avec $q_{1-\alpha/2} > 0$ le quantile d'une loi normale centrée réduite réelle de niveau $1-\alpha/2$, tel que $P(Z \in [-q_{1-\alpha/2}, q_{1-\alpha/2}]^d) \geq 1-\alpha$. Aussi le polyèdre $n^{-1/2}\cdot(I_1(\widehat{\theta}_n))^{-1/2}\cdot[-q_{1-\alpha/2}, q_{1-\alpha/2}]^d$ recentré autour de $\widehat{\theta}_n$ formera la région de confiance cherchée.*

# 3 Tests paramétriques

## 3.1 Principes d'un test

Un test permet, à partir d'une réalisation d'un échantillon, de décider entre deux hypothèses, en mettant en avant une hypothèse privilégiée, appelée hypothèse $H_0$, et une hypothèse alternative, appelée $H_1$. On associe à un test un niveau $\alpha$ (avec souvent $\alpha \simeq 0.05$) et une puissance $1-\beta$. La plupart du temps, $\alpha$ est fixé a priori et $\beta$ s'en déduit. Plus précisément,

**Définition 30.** *On se place dans le cadre d'un modèle paramétrique dominé $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$ et soit $\theta$ la "vraie" valeur du paramètre. Un problème de test est un choix entre deux hypothèses :*

$$\begin{cases} H_0 : \theta \in \Theta_0 & : \textit{ hypothèse dite nulle} \\ \\ H_1 : \theta \in \Theta_1 & : \textit{ hypothèse dite alternative,} \end{cases} \tag{7}$$

*où $\Theta_0 \subset \mathbb{R}^p$, $\Theta_1 \subset \mathbb{R}^d$ et $\Theta_0 \cap \Theta_1 = \emptyset$.*

Ceci posé, on peut préciser deux types de problèmes de tests suivant les constitutions de $\Theta_0$ et $\Theta_1$ :

**Définition 31.** *Une hypothèse ($H_0$ ou $H_1$) est dite simple si elle est associée à un singleton ($\Theta_0$ ou $\Theta_1$). Sinon, elle sera dite composite. Dans le cas réel ($\Theta \subset \mathbb{R}$), si $H_0$ est simple de la forme $\theta = \theta_0$, et si $H_1$ est composite de la forme $\theta > \theta_0$ ou $\theta < \theta_0$, on parlera de test unilatéral ; si $H_1$ est composite de la forme $\theta \neq \theta_0$, on parlera de test bilatéral.*

Comment faire pour choisir entre les deux hypothèses $H_1$ et $H_2$ ? Il faudra partir de ce que l'on peut connaître du modèle, c'est-à-dire généralement un échantillon observé $(X_1, \ldots, X_n)$. Pour cela, on définit une statistique qui sera la clé de voûte du test :

**Définition 32.** *Dans le cadre du problème de test (7, soit $\widehat{T}$ une statistique (donc une fonction mesurable d'un échantillon $(X_1, \cdots, X_n)$ issu du modèle) à valeurs dans $\mathbb{R}^d$, qui sera appelée statistique du test. Le test sera défini par la fonction $\widehat{\phi} = \mathbb{I}_{\widehat{T} \in W}$, où $W$ est une partie de $\mathbb{R}^p$ appelée région critique du test (et sa partie complémentaire dans $\mathbb{R}^p$ est appelée région d'acceptation du test). Si $\widehat{\phi} = 1$, on choisira $H_1$, sinon on décidera plutôt $H_0$.*

Donc, à chaque hypothèse $H_0$ et $H_1$, on associe une partie de $\mathbb{R}^p$ pour la statistique de test $\widehat{T}$. En général, ces parties ne sont pas $\Theta_0$ et $\Theta_1$. Pour pouvoir précisément déterminer la région $W$, dans un cadre théorique (qui n'est pas le même que le cadre pratique, voir plus bas), on peut commencer par associer une fonction puissance à la statistique de test, puis définir les erreurs de premier espèce $\alpha$ et de deuxième espèce $\beta$ :

**Définition 33.** *Pour la statistique de test $\widehat{T}$, on associe :*

- *une fonction puissance, qui est la probabilité de choisir $H_1$ : $\theta \in \Theta_1 \mapsto \mathbb{P}_\theta(\widehat{T} \notin W)$.*
- *une erreur de première espèce : $P_{H_0}(\textit{Choisir } H_1) = \alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\widehat{T} \in W)$ ;*
- *une erreur de seconde espèce : $P_{H_1}(\textit{Choisir } H_0) = \beta = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta(\widehat{T} \notin W)$.*

*La puissance du test est $1-\beta$.*

Cependant, ce qui vient d'être écrit reste théorique. En pratique, on utilisera plutôt la démarche suivante :

**Construction concrète d'un test :** On suppose le problème de test (7). On pose également a priori $\alpha$ qui dépend du problème posé (mais en général $\alpha = 0.05$), et $1-\alpha$ est appelé le niveau du test. Par la suite, on réalise :

1. L'expression quantitative des hypothèses $H_0$ et $H_1$.
2. Le choix de la statistique $\widehat{T}$ du test.
3. La construction d'une région critique $W$ à l'hypothèse $H_1$ par rapport à $\widehat{T}$.
4. La détermination explicite de $W$ en fonction de $\alpha$.
5. Le calcul (si possible) de la puissance du test $1-\beta$.
6. Pour la réalisation de l'échantillon, rejet ou acceptation de $H_0$.

**Remarque :**Cependant, en pratique on ne procède pas ainsi. On a donc deux types d'erreur. Le choix de l'hypothèse privilégiée est donc fondamental car le résultat d'un test n'est pas symétrique. Par exemple, supposons que l'on ait pour modèle $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathcal{N}(\theta,1)^{\otimes n}, \theta \in \mathbb{R})$ et que l'on veuille tester $H_0 : \theta = 0$ contre $H_1 : \theta = 1$ à partir d'un échantillon $(X_1, \cdots, X_n)$ du modèle. Nous verrons pourquoi un peu plus loin, $\overline{X}_n$ est une statistique de test pertinente. Par exemple, si $n = 1$, et $X_1(\omega) = \overline{X}_1(\omega) = 0.8$, que va-t-on choisir entre $H_0$ et $H_1$ ? Naturellement, une région critique sera de la forme $[s, +\infty[$, où $s \in \mathbb{R}$, car $\overline{X}_n$ est un estimateur de $\theta$. On détermine $s$ à l'aide de $\alpha$, puisque $P_{H_0}(\text{Choisir } H_1) = \alpha = P_0(\overline{X}_1 \geq s)$, donc par exemple, si $\alpha = 0.05$, $s \simeq 1.65$. Par suite, si $\overline{X}_1(\omega) = 0.8$, on accepte $H_0$ et l'erreur de seconde espèce est $P_1(\overline{X}_1 < s) \simeq 0.74$, donc très élevée : le test n'est pas très discriminant. Maintenant, si on inverse $H_0$ et $H_1$, soit $H_0 : \theta = 1$ contre $H_1 : \theta = 0$, le même résultat $X_1(\omega) = 0.8$, conduit à accepter $H_0$, avec une erreur de second espèce encore $\simeq 0.74$. On obtient donc deux résultats opposés pour la même expérience aléatoire. Les hypothèses $H_0$ et $H_1$ ne sont clairement pas interchangeable.

La question qui se pose maintenant est de savoir comment trouver une statistique de test. Une idée naturelle dans ce cadre paramétrique serait d'utiliser un estimateur du maximum de vraisemblance.

## 3.2 Test de Wald

Un estimateur du maximum de vraisemblance permet d'associer à chaque hypothèse du test un ensemble de même "forme" que $\Theta_0$ et $\Theta_1$. Cependant, la difficulté est trouver la loi de l'estimateur du maximum de vraisemblance $\widehat{\theta}$ à $n$ fixé. Si cela est possible, on utilisera directement $\widehat{\theta}$ comme statistique de test.

Sinon, de manière plus générale, on connaît la loi asymptotique de $\widehat{\theta}_n$ quand le modèle est régulier. Donc quand $n$ est grand, on pourrait utiliser une loi normale comme approximation de la loi de $\widehat{\theta}_n$. Mais, un nouvel obstacle apparaît : la matrice de covariance asymptotique, qui est la matrice d'information de Fisher inverse, dépend du paramètre $\theta$. Aussi va-t-on préférer utiliser la statistique de test $\widehat{T}$ suivante :

**Définition 34.** *Pour un modèle paramétrique dominé* **régulier** $((\Omega')^n, \mathcal{A}'^{\otimes n}, \mathbb{P}_\theta, \theta \in \Theta)$*, où* $\Theta \subset \mathbb{R}^p$*. La statistique de Wald* $\widehat{T}$ *pour le test* $H_0 : \theta = \theta_0$ *contre* $H_1 : \theta \in \Theta_1$ *est :* $\widehat{T}_n = n \cdot {}^t(\widehat{\theta}_n - \theta) \cdot I(\theta) \cdot (\widehat{\theta}_n - \theta)$*.*

Pour montrer "théoriquement" la pertinence de ce test, on va donc considérer la suite de tests $(\widehat{T}_n)$ en se plaçant dans le "grand" modèle asymptotique :

**Théorème 28.** *Dans le cadre d'un modèle paramétrique* $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\otimes \mathbb{N}}_{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$*, où* $\Theta \subset \mathbb{R}^p$*, dominé par une mesure* $\mu$ *et régulier, pour le problème de test* $H_0 : \theta = \theta_0$ *contre* $H_1 : \theta \neq \theta_0$*, alors, en notant* $\widehat{T}_n$ *la statistique de test de Wald pour le modèle projeté de taille* $n$ *sous l'hypothèse* $H_0$*,*

$$\widehat{T}_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(p).$$

*La région de rejet asymptotique du test sera donc de la forme* $\widehat{T}_n > s_\alpha$*, où* $s_\alpha$ *est le quantile d'ordre* $1-\alpha$ *de la loi du* $\chi^2(p)$*. La suite de test* $(\widehat{T}_n)_n$ *a donc une puissance qui tend vers* 1 *lorsque* $\alpha$ *est fixé.*

*Démonstration :* La loi asymptotique de $\widehat{\theta}_n$ induit la loi asymptotique de $\widehat{T}_n$, car $\sqrt{n} \cdot I(\theta)^{1/2} \cdot (\widehat{\theta}_n - \theta)$ suit asymptotiquement une loi $\mathcal{N}(0, I_d)$ sous l'hypothèse $H_0$ et $\widehat{T}_n = \|\sqrt{n} \cdot I(\theta)^{1/2} \cdot (\widehat{\theta}_n - \theta)\|^2$. ■

Voici donc un premier type de test, qui sous certaines conditions de régularités du modèle et pour certaines hypothèses de tests est intéressant. Mais pourrait-on faire mieux ? Et en quel sens ? Désormais, il nous faut donc définir un moyen de comparaison entre deux tests.

## 3.3 Test du rapport de vraisemblance

**Définition 35.** *Sous les hypothèses et notations précédentes, on dira qu'un test* $\phi$ *est uniformément le plus puissant (U.P.P.) au seuil* $\alpha$ *si le niveau de* $\widehat{\phi}$ *associé à la statistique* $\widehat{T}$ *est inférieur ou égal à* $\alpha$ *et si pour tout autre test* $\widehat{\phi}'$ *associé à la statistique* $\widehat{T}'$ *de niveau inférieur ou égal à* $\alpha$*,* $\forall \theta \in \Theta_1$*,*

$$\mathbb{E}_\theta(\widehat{\phi}) = 1 - \mathbb{P}_\theta(\widehat{T} \notin W) \leq 1 - \mathbb{P}_\theta(\widehat{T}' \notin W') = \mathbb{E}_\theta(\widehat{\phi}').$$

**Définition 36.** *Sous les hypothèses précédentes, si $L_\theta(.)$ est la vraisemblance, on appellera test du rapport de vraisemblance (test de Neyman-Person dans le cas d'hypothèses simples) un test de statistique $\widehat{T}$ telle que :*

$$\widehat{T} = \frac{\sup_{\theta\in\Theta_0} L_\theta(X_1,\cdots,X_n)}{\sup_{\theta\in\Theta_1} L_\theta(X_1,\cdots,X_n)}.$$

*La région critique $W$ associée à un tel test est de la forme $W = ]+\infty, K[$ (donc si $\widehat{T} < K$, on rejette $H_0$).*

Une des vertus du test du rapport de vraisemblance par rapport au test de Wald est qu'il peut être utilisé dans un modèle non régulier (mais la question de sa loi, ou de la loi d'une fonctionnelle de ce test, demeure). De plus, la propriété suivante confirme l'intérêt de cette statistique de test :

**Propriété 21** (Principe de Lehmann (admis))**.** *Dans le cas du test de deux hypothèses simples, ou d'un test unilatéral ($\Theta \subset \mathbb{R}$), ce test est U.P.P. Dans le cas d'un test bilatéral, il n'existe pas forcément de test U.P.P.*

Enfin, un tel test pour un modèle régulier, va pouvoir être traité de manière générale grâce à la normalité asymptotique de l'estimateur du maximum de vraisemblance :

**Théorème 29.** *Dans le cadre d'un modèle paramétrique $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure $\mu$ et régulier, pour le problème de test $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1 \supset \Theta_0$, alors, en notant $\widehat{T}_n$ la statistique du rapport de vraisemblance,*

$$-2\log(\widehat{T}_n) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(dim(\Theta_1) - dim(\Theta_0)).$$

*La région de rejet asymptotique du test sera donc de la forme $-2\log(\widehat{T}_n) > s_\alpha$, où $s_\alpha$ est le quantile d'ordre $1-\alpha$ de la loi du $\chi^2(dim(\Theta_1) - dim(\Theta_0))$. La suite de test $(\widehat{T}_n)_n$ a donc une puissance qui tend vers 1 lorsque $\alpha$ est fixé. On remarquera que si $\Theta_0$ est un singleton $dim(\Theta_0) = 0$.*

*Idée de démonstration :* la démonstration est basée sur un développement de Taylor autour du vrai paramètre.

# 4 Sélection de modèles

On sait que sous de bonne condition l'estimateur du maximum de vraisemblance est asymptotiquement sans biais et asymptotiquement efficace, on peut même préciser un peu la valeur de la log-vraisemblance obtenue, ce qui donne des critères de sélection du modèle.

## 4.1 Critère d'akaike

On suppose dans cette section que le modèle est régulier. Soit $(Y_1,\cdots,Y_n)$ des variables aléatoires i.i.d. (éventuellement vectorielles si on veut y inclure des variables explicatives), soit $\hat{\theta} \in \Theta$, l'estimateur du maximum de vraisembance du vrai paramètre $\theta_0$. Pour une variable aléatoire $Y$ de même loi que le phénomène étudié mais indépendante des observations $(Y_1,\cdots,Y_n)$, soit

$$D = 2E_{\theta_0}\left[\log L_{\theta_0}(Y) - \log L_{\hat{\theta}}(Y)\right] = 2K\left(P_{\theta_0}, P_{\hat{\theta}}\right)$$

où $K\left(P_{\theta_0}, P_{\hat{\theta}}\right)$ est la distance de Kullback de la vraie mesure de probabilité (calculée avec $\theta_0$) et la mesure de probabilité calculée avec le paramètre estimé $\hat{\theta}$.

Alors

$$D = 2 \times E\left(\log L_{\theta_0}(Y) - \frac{1}{n}\log L_{\hat{\theta}}(Y_1,\cdots,Y_n)\right) + 2\frac{q}{n} + o\left(\frac{1}{n}\right)$$

où $q$ est le nombre de paramètre libres du modèle et $\hat{L}_{\hat{\theta}}$ est l'estimation de la vraisemblance de la vraiemblance pour le paramètre estimé. Ainsi pour trouver (asymptotiquement) le modèle le plus proche pour la distance de Kullback de $P_{\theta_0}$ il faudra minimiser en $\theta$

$$AIC = -\frac{1}{n}\log L_{\hat{\theta}}(Y_1,\cdots,Y_n) + \frac{q}{n}$$

**Preuve** On a

$$2 \times E\left(\log L_{\theta_0}(Y) - \frac{1}{n}\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right)\right) =$$
$$2 \times E\left(\log L_{\theta_0}(Y) - \log L_{\hat{\theta}}(Y) + \log L_{\hat{\theta}}(Y) - \frac{1}{n}\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right)\right).$$

Il suffit donc de prouver que

$$2 \times E\left(\log L_{\hat{\theta}}(Y) - \frac{1}{n}\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right)\right) = -2\frac{q}{n} + o\left(\frac{1}{n}\right)$$

On a l'approximation :

$$2\log L_{\hat{\theta}}(Y) = 2\log L_{\theta_0}(Y) + 2\left(\hat{\theta} - \theta_0\right)^T \frac{\partial \log L_{\theta_0}(Y)}{\partial \theta} + \left(\hat{\theta} - \theta_0\right)^T \frac{\partial^2 \log L_{\theta_0}(Y)}{\partial \theta^2}\left(\hat{\theta} - \theta_0\right) + o_P\left(\left\|\hat{\theta}_q - \theta_0\right\|^2\right)$$

$$= 2\log L_{\theta_0}(Y) + 2\left(\hat{\theta} - \theta_0\right)^T \frac{\partial \log L_{\theta_0}(Y)}{\partial \theta} + tr\left(\frac{\partial^2 \log L_{\theta_0}(Y)}{\partial \theta^2}\left(\hat{\theta} - \theta_0\right)\left(\hat{\theta} - \theta_0\right)^T\right) + o_P\left(\left\|\hat{\theta}_q - \theta_0\right\|^2\right)$$

Comme $\theta_0$ maximise l'espérance de la log-vraisemblance, on aura

$$E_{\theta_0}\left[\frac{\partial \log L_{\theta_0}(Y)}{\partial \theta}\right] = 0$$

d'où

$$2E_{\theta_0}\left[\log L_{\theta_0}(Y) - \log L_{\hat{\theta}}(Y)\right] = E_{\theta_0}\left[tr\left(-\frac{\partial^2 \log L_{\theta_0}(Y)}{\partial \theta^2}\left(\hat{\theta} - \theta_0\right)\left(\hat{\theta} - \theta_0\right)^T\right)\right] + o\left(\frac{1}{n}\right)$$
$$= tr\left(I_{\theta_0} Var\left(\hat{\theta}\right)\right) + o\left(\frac{1}{n}\right)$$

et comme $Var\left(\hat{\theta}\right) = \frac{1}{n} I_{\theta_0}^{-1}$, on aura :

$$2E_{\theta_0}\left[\log L_{\theta_0}(Y) - \log L_{\hat{\theta}}(Y)\right] = \frac{q}{n} + o\left(\frac{1}{n}\right)$$

Soit

$$2E_{\theta_0}\left[\log L_{\hat{\theta}}(Y)\right] = 2E_{\theta_0}\left[\log L_{\theta_0}(Y)\right] - \frac{q}{n} + o\left(\frac{1}{n}\right)$$

De plus, par le théorème du test de rapport de vraisemblance

$$2\left(\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right) - \log L_{\theta_0}\left(Y_1, \cdots, Y_n\right)\right) \xrightarrow{\mathcal{L}} \chi^2(q)$$

On aura

$$2E_{\theta_0}\left[\left(\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right) - \log L_{\theta_0}\left(Y_1, \cdots, Y_n\right)\right)\right] = q + o(1)$$

et comme

$$2E_{\theta_0}\left[\frac{1}{n}\log L_{\theta_0}\left(Y_1, \cdots, Y_n\right) - \log L_{\theta_0}(Y)\right] = 0$$

$$2 \times E\left(\log L_{\hat{\theta}}(Y) - \frac{1}{n}\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right)\right) =$$
$$2 \times E\left(\log L_{\hat{\theta}}(Y) - \log L_{\theta_0}(Y) + \log L_{\theta_0}(Y) - \frac{1}{n}\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right)\right) =$$
$$2 \times E\left(\log L_{\hat{\theta}}(Y) - \log L_{\theta_0}(Y) + \frac{1}{n}\log L_{\theta_0}\left(Y_1, \cdots, Y_n\right) - \frac{1}{n}\log L_{\hat{\theta}}\left(Y_1, \cdots, Y_n\right)\right) = -2\frac{q}{n} + o\left(\frac{1}{n}\right)$$

■

Donc, asymptotiquement, si on trouve

$$-\frac{1}{n}\log L_{\hat{\theta}_{q_1}}\left(X_1, \cdots, X_n\right) + \frac{q_1}{n} < -\frac{1}{n}\log L_{\hat{\theta}_{q_2}}\left(X_1, \cdots, X_n\right) + \frac{q_2}{n}$$

alors

$$E\left(\log L_{\theta_0}(Y) - \log L_{\hat{\theta}_{q_1}}(Y)\right) < E\left(\log L_{\theta_0}(Y) - \log L_{\hat{\theta}_{q_2}}(Y)\right)$$

et le modèle avec $q_1$ paramètre libres est en moyenne plus proche du vrai modèle que celui avec $q_2$ paramètre libre.

### 4.2 Inconvénient de AIC pour la sélection de modèles.

Rappelons que l'objectif est de choisir la dimension $q$ de $\theta_q$ telle que

$$E\left(\log L_{\theta_0}(Y) - \log L_{\hat{\theta}_q}(Y)\right)$$

soit minimale. Soit $q_0$ la vraie dimension de $\theta_0$, si $q < q_0$, on aura

$$AIC(q) - AIC(q_0) \simeq 2n\left(\log L_{\theta_0}(Y) - \log L_{\hat{\theta}_q}(Y)\right) - 2(q_0 - q) \overset{r\to+\infty}{\longrightarrow} +\infty$$

car $\inf_{\theta_q} \log L_{\theta_0}(Y) - \log L_{\theta_q}(Y) \geq \varepsilon > 0$ si l'ensemble des paramètres possibles est compact. Ainsi les modèles de dimension plus petite que $q_0$ sont asymptotiquement disqualifiés. Par contre si $q > q_0$,

$$AIC(q) - AIC(q_0) \simeq -\chi^2(q - q_0) + 2(q - q_0)$$

et la probabilité de disqualifier les modèles de dimension supérieure à $q_0$ ne tend pas vers 1, puisque le terme issu des pénalités $2(q - q_0)$ ne diverge pas quand $n$ tend vers l'infini. AIC ne sélectionne donc pas, avec probabilité 1 le vrai modèle du dimension $q_0$.

## 5 Le critère BIC

### 5.1 Consistance

Une solution pour obtenir un critère qui sélectionne asymptotiquement la vraie dimension du modèle $q_0$ est de pénaliser "plus fort" sans pour autant pénaliser "trop fort". Un compromis est par exemple d'utiliser le critère BIC :

$$BIC(q) = -2\log L_{\hat{\theta}_q}(X_1, \cdots, X_n) + q\ln(n)$$

ainsi, si $q < q_0$, on aura

$$BIC(q) - BIC(q_0) \simeq 2n\left(\log L_{\theta_0}(Y) - \log L_{\hat{\theta}_q}(Y)\right) - (q_0 - q)\ln(n) \overset{n\to+\infty}{\longrightarrow} +\infty$$

car $\inf_{\theta_q} \log L_{\theta_0}(Y) - \log L_{\theta_q}(Y) \geq \varepsilon > 0$ si l'ensemble des paramètres possibles est compact. Ainsi les modèles de dimension plus petite que $q_0$ sont asymptotiquement disqualifiés. De plus, si $q > q_0$,

$$BIC(q) - BIC(q_0) \simeq -\chi^2(q - q_0) + (q - q_0)\ln(n)$$

et la probabilité de disqualifier les modèles de dimension supérieure à $q_0$ tend vers 1, puisque le terme issu des pénalités $(q - q_0)\ln(n)$ diverge quand $n$ tend vers l'infini. BIC sélectionne donc, avec probabilité 1 le vrai modèle du dimension $q_0$. On pourra remarque que pour être consistant le terme de pénalité $p_n(q)$ de la vraisemblance doit vérifié :

$$\frac{p_n(q)}{n} \overset{n\to\infty}{\longrightarrow} 0 \text{ et } p_n(q) \overset{n\to\infty}{\longrightarrow} \infty$$

ce qui laisse un grand nombre de choix possibles !