Université Pierre et Marie Curie
Master Mathématiques et Applications

Année 2015/2016
Second Semestre

Statistique mathématique

Arnaud Guyader

Remerciements

Je remercie sincèrement Lucien Birgé, précédent responsable du module de Statistique mathématique, de m'avoir prêté le cours qu'il avait rédigé et dont je me suis inspiré pour le présent document. Le Chapitre 2 est, grosso modo, un condensé des Chapitres 2 et 3 du livre *Régression avec R* de Pierre-André Cornillon et Eric Matzner-Lober, que je remercie également à cette occasion.

Table des matières

Chapitre 1

Modélisation statistique

Introduction

Partons d'un exemple jouet. Une pièce a été lancée n fois de suite : à l'issue de cette expérience, on dispose donc du n-uplet $(x_1, \dots, x_n)$ avec la convention $x_i = 1$ si le i-ème lancer a donné Pile et $x_i = 0$ pour Face. Les valeurs x_i peuvent ainsi être considérées comme des réalisations de variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées (en abrégé : iid) selon la loi de Bernoulli de paramètre θ, ce que l'on notera

$$(X_1, \dots, X_n) \overset{iid}{\sim} \mathcal{B}(\theta),$$

où la probabilité $\theta \in]0, 1[$ de tomber sur Pile est inconnue. A partir de cet échantillon $(x_1, \dots, x_n)$, on souhaite par exemple estimer le paramètre θ et tester si la pièce est équilibrée, i.e. si $\theta = 1/2$ ou non.

Ces questions sont typiques de ce que l'on appelle la statistique inférentielle. Il importe de comprendre dès à présent la différence entre probabilités et statistiques. En probabilités, le paramètre θ est supposé connu, donc la loi de $X = X_1$ aussi, et on répond à des questions comme : quelle est la loi du nombre $S_n = X_1 + \cdots + X_n$ de Pile sur les n lancers ? quelle est la limite du rapport S_n/n lorsque n tend vers l'infini ? etc., bref à en déduire des résultats impliquant cette loi de X. En statistiques, c'est le contraire : on dispose d'un échantillon $(x_1, \dots, x_n)$ et on veut remonter à la loi de X, c'est-à-dire au paramètre θ.

Il n'en reste pas moins que les outils utilisés dans les deux domaines sont les mêmes : loi des grands nombres, théorème central limite, inégalités classiques, modes de convergence stochastique, etc. Pour la plupart, ceux-ci ont déjà été vus en cours de probabilités et nous nous contenterons donc de les rappeler brièvement.

1.1 Probabilités : rappels et compléments

Si X est une variable aléatoire, sa loi P_X est définie pour tout borélien B de $\mathbb{R}$ par

$$P_X(B) = \mathbb{P}(X \in B),$$

probabilité que la variable X tombe dans l'ensemble B. Cette loi est complètement déterminée par un objet bien plus simple et maniable : la fonction de répartition F_X, définie pour tout réel x par

$$F_X(x) = \mathbb{P}(X \leq x) = P_X(]-\infty, x]),$$

probabilité que la variable X tombe au-dessous de x. Rappelons que cette fonction est croissante, a pour limites respectives 0 et 1 en $-\infty$ et $+\infty$, et est continue à droite. Elle admet un nombre au plus dénombrable de discontinuités et on a pour tout réel x_0

$$\mathbb{P}(X = x_0) = F_X(x_0) - F_X(x_0^-) = F_X(x_0) - \lim_{x \to x_0, x < x_0} F_X(x).$$

En d'autres termes, F_X présente un saut au point x_0 si et seulement si la probabilité pour X de tomber en x_0 est non nulle, la hauteur du saut correspondant précisément à cette probabilité.

Exemples :

1. Si $X \sim \mathcal{B}(\theta)$, alors
$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1-\theta & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$
2. Si $X \sim \mathcal{N}(0,1)$, loi gaussienne centrée réduite, on note Φ sa fonction de répartition, définie pour tout réel x par
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$
Par symétrie de la loi normale centrée par rapport à 0, il vient $\Phi(-x) = 1 - \Phi(x)$, c'est-à-dire que le point $(0, 1/2)$ est centre de symétrie de la courbe représentant Φ (voir Figure 1.1).

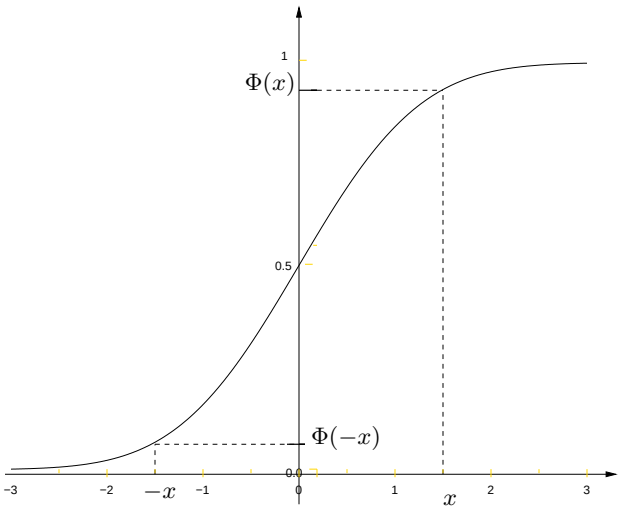

FIGURE 1.1 – Fonction de répartition Φ d'une loi normale $\mathcal{N}(0,1)$ et relation : $\Phi(-x) = 1 - \Phi(x)$.

1.1.1 Modes de convergence

Par rapport aux nombreux modes de convergence vus en cours de probabilités, nous nous focaliserons plus particulièrement sur : la convergence en probabilité, la convergence presque sûre et la convergence en loi.

Définition 1 (Convergence en probabilité)
On dit que la suite (X_n) de variables aléatoires converge en probabilité vers la variable aléatoire X et on note

$$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$$

si

$$\forall \varepsilon > 0 \qquad \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

Si l'on voit une variable aléatoire comme une fonction de Ω dans $\mathbb{R}$, la convergence presque sûre peut quant à elle être considérée comme une version stochastique de la convergence simple d'une suite de fonctions vue en cours d'analyse. Elle est définie par :

$$X_n \xrightarrow[n\to\infty]{p.s.} X \qquad \text{si} \qquad \mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) := \mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1,$$

et elle implique la convergence en probabilité [1] :

$$X_n \xrightarrow[n\to\infty]{p.s.} X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} X.$$

Très souvent, un résultat de convergence presque sûre se démontre par l'intermédiaire du Lemme de Borel-Cantelli. Précisément, si pour tout $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) < \infty,$$

alors (X_n) converge presque sûrement vers X.

Passons maintenant à la convergence en loi, d'usage constant en statistiques en raison du Théorème Central Limite. **Attention** : contrairement aux convergences en probabilité et presque sûre, elle concerne la convergence d'une suite de lois, non la convergence d'une suite de variables !

Définition 2 (Convergence en loi)
On dit que la suite (X_n) de variables aléatoires converge en loi vers (la loi de la variable aléatoire) X et on note

$$X_n \xrightarrow[n\to\infty]{\mathcal{L}} X \qquad ou \qquad X_n \rightsquigarrow X$$

si pour toute fonction continue et bornée φ, on a

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

Exemple et Notation : si pour tout n, $X_n = X \sim \mathcal{N}(0,1)$, alors par symétrie de la loi normale $X' = -X \sim \mathcal{N}(0,1)$, donc

$$X_n = X \xrightarrow[n\to\infty]{\mathcal{L}} X' = -X,$$

mais il n'y a bien sûr pas convergence en probabilité de $(X_n) = X$ vers $X' = -X$. Afin de mettre en évidence le fait que c'est la suite des lois des X_n qui converge, on utilisera souvent l'abus de notation consistant à mettre une loi à la limite. Dans cet exemple, on pourra ainsi écrire

$$X_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Le critère de la définition ci-dessus n'est pas facile à vérifier. Il en existe un parfois plus commode, qui consiste à établir la convergence simple de la suite des fonctions de répartition.

Proposition 1 (Fonctions de répartition & Convergence en loi)
La suite de variables aléatoires (X_n) converge en loi vers X si et seulement si en tout point de continuité x de F_X, on a

$$F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x).$$

1. On peut le voir facilement grâce au Lemme de Fatou.

Notation : pour $a < b$, nous noterons (a, b) l'intervalle allant de a à b sans préciser si les extrémités y appartiennent ou non (donc quatre situations possibles).

Exemple : supposons que (X_n) converge en loi vers X, avec a et b des points de continuité de F_X, alors grâce au résultat précédent, on en déduit que

$$\mathbb{P}\left(X_n \in (a,b)\right) \xrightarrow[n\to\infty]{} \mathbb{P}\left(X \in (a,b)\right).$$

Ceci marche en particulier lorsque X est une variable gaussienne, ce qui sera très souvent le cas pour la convergence en loi.

Proposition 2 (Convergence uniforme d'une suite de fonctions de répartition)
Si la fonction de répartition F_X est continue sur tout $\mathbb{R}$, alors cette convergence simple est en fait uniforme, c'est-à-dire que

$$\|F_{X_n} - F_X\|_\infty := \sup_{x\in\mathbb{R}} |F_{X_n}(x) - F_X(x)| \xrightarrow[n\to\infty]{} 0.$$

Ceci résulte du deuxième théorème de Dini appliqué à notre cadre. A nouveau, ce résultat s'applique en particulier lorsqu'il y a convergence vers une loi normale. Notons que le critère des fonctions de répartition pour vérifier la convergence en loi est pratique lorsque X_n s'écrit comme le minimum ou le maximum de variables aléatoires.

Une autre façon de vérifier la convergence en loi est de passer par les fonctions caractéristiques. Rappelons que la fonction caractéristique d'une variable aléatoire X est la fonction

$$\begin{array}{rccl} \phi_X : & \mathbb{R} & \to \mathbb{C} \\ & t & \mapsto \phi_X(t) = \mathbb{E}\left[e^{itX}\right] \end{array}$$

Comme son nom l'indique, elle caractérise la loi d'une variable, au sens où X et Y ont même loi si et seulement si $\phi_X = \phi_Y$. On a alors l'équivalent de la Proposition 1, c'est-à-dire que la convergence en loi se ramène à la convergence simple d'une suite de fonctions.

Théorème 1 (Critère de convergence de Paul Lévy)
La suite de variables aléatoires (X_n) converge en loi vers X si et seulement si

$$\forall t \in \mathbb{R} \qquad \phi_{X_n}(t) \xrightarrow[n\to\infty]{} \phi_X(t).$$

Puisqu'elle intervient dans de très nombreux phénomènes asymptotiques, il convient de connaître la fonction caractéristique de la gaussienne, à savoir

$$X \sim \mathcal{N}(m, \sigma^2) \iff \phi_X(t) = \exp\left(imt - \frac{\sigma^2 t^2}{2}\right).$$

Ce critère de Paul Lévy est en particulier efficace lorsqu'on a affaire à des sommes de variables aléatoires indépendantes, la fonction caractéristique de la somme étant alors tout simplement égale au produit des fonctions caractéristiques :

$$X \perp Y \implies \phi_{X+Y} = \phi_X \times \phi_Y.$$

Exemple : dans l'exemple introductif, la variable correspondant au nombre de Pile sur les n lancers s'écrit

$$S_n = X_1 + \cdots + X_n \qquad \text{avec} \qquad (X_1, \ldots, X_n) \overset{iid}{\sim} \mathcal{B}(\theta).$$

En appliquant la définition de la fonction caractéristique, on trouve pour la variable X_1 :

$$\phi_{X_1}(t) = (1-\theta) + \theta e^{it}.$$

Celle de la variable S_n s'en déduit donc à peu de frais :

$$\phi_{S_n}(t) = \mathbb{E}\left[e^{itS_n}\right] = \mathbb{E}\left[e^{it(X_1+\cdots+X_n)}\right] = \left(\mathbb{E}\left[e^{itX_1}\right]\right)^n = \left((1-\theta)+\theta e^{it}\right)^n.$$

Puisque S_n suit une loi binomiale $\mathcal{B}(n,\theta)$, on a en fait obtenu la fonction caratéristique de la loi binomiale.

Si un résultat de convergence en loi se démontre généralement grâce à l'un de ces deux critères (fonctions de répartition ou caractéristiques), un résultat de convergence en probabilité ou presque sûre découle typiquement de l'une des inégalités que nous rappelons maintenant.

1.1.2 Majorations classiques

La plupart des inégalités qui suivent quantifient typiquement la probabilité qu'une variable aléatoire s'éloigne de sa moyenne, ou plus généralement qu'elle prenne de grandes valeurs. Leur intérêt est de ne pas faire intervenir la loi de cette variable, qui peut être très compliquée, mais plutôt des moments de celle-ci.

Proposition 3 (Inégalité de Markov)
Soit X une variable aléatoire, alors pour tous réels $c > 0$ et $p > 0$, on a

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}\left[|X|^p\right]}{c^p}.$$

Notons que si $\mathbb{E}\left[|X|^p\right] = +\infty$, cette inégalité reste valide, mais elle ne nous apprend rien... En prenant $c = 2$ et en considérant la variable centrée $X - \mathbb{E}[X]$, on en déduit le résultat suivant.

Corollaire 1 (Inégalité de Bienaymé-Tchebychev)
Soit X une variable aléatoire, alors pour tout réel $c > 0$, on a

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

Exemple : dans l'exemple introductif, un estimateur naturel de θ est la moyenne empirique des X_i, c'est-à-dire

$$\hat{\theta}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

Puisque $S_n \sim \mathcal{B}(n,\theta)$, on a $\mathrm{Var}(S_n) = n\theta(1-\theta)$ donc $\mathrm{Var}(\hat{\theta}_n) = \theta(1-\theta)/n$ et l'inégalité ci-dessus donne

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right) \leq \frac{\theta(1-\theta)}{c^2 n} \leq \frac{1}{4c^2 n}, \tag{1.1}$$

la dernière inégalité venant de ce que $0 < \theta(1-\theta) \leq 1/4$ pour tout $\theta \in]0,1[$. D'après la Définition 1, ceci prouve que la suite des fréquences empiriques $(\hat{\theta}_n)$ tend en probabilité vers la vraie probabilité θ de Pile lorsque le nombre de lancers tend vers l'infini :

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

La borne de Bienaymé-Tchebychev n'est cependant pas suffisamment précise pour montrer la convergence presque sûre via Borel-Cantelli puisque la série $\sum 1/n$ est divergente. Qu'à cela ne tienne, on peut faire bien mieux, comme nous allons le voir maintenant.

Sous réserve d'existence de moments, les inégalités ci-dessus permettent de majorer l'écart à la moyenne par des fonctions polynomiales. Si l'on s'intéresse à des variables bornées, on peut même obtenir des majorations exponentielles. On parle alors d'inégalités de grandes déviations ou de résultats de concentration, car il faut être bien concentré pour pouvoir les comprendre.

Proposition 4 (Inégalité de Hoeffding)
Soit $X_1, \dots, X_n$ des variables aléatoires indépendantes et bornées, avec $a_i \leq X_i \leq b_i$. Notant $S_n = X_1 + \cdots + X_n$ leur somme, on a pour tout réel $c > 0$

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq c\right) \leq \exp\left(-\frac{2c^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

En changeant X_i en $-X_i$, on en déduit aussitôt

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \leq -c\right) \leq \exp\left(-\frac{2c^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

d'où il vient

$$\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \geq c\right) \leq 2\exp\left(-\frac{2c^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Si en plus d'être indépendantes, les variables X_i ont même loi, alors on peut prendre $a_i = a$, $b_i = b$ et en remplaçant c par cn, on en déduit une majoration de l'écart entre la moyenne empirique et la moyenne théorique. Précisément, en notant $m = \mathbb{E}[X_1]$, on obtient

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| \geq c\right) \leq 2\exp\left(-\frac{2c^2 n}{(b-a)^2}\right).$$

Exemple : pour le jeu de Pile ou Face, puisque $a = 0$, $b = 1$ et $m = \theta$, cette inégalité donne

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right) \leq 2\exp\left(-2c^2 n\right), \tag{1.2}$$

laquelle est meilleure que celle de Tchebychev vue en (1.1) dès que $c^2 n \geq 1,08$ (voir Figure 1.2). En particulier, pour tout $c > 0$, on voit que

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right) < \infty,$$

donc par Borel-Cantelli, $(\hat{\theta}_n)$ tend presque sûrement vers θ.

Notons que l'inégalité (1.2) est valable pour toute taille d'échantillon n. A la limite lorsque n tend vers l'infini, on peut faire encore un peu mieux car on connaît asymptotiquement la loi de cet écart entre moyennes empirique et théorique : c'est une gaussienne, comme le spécifie le Théorème Central Limite de la section suivante.

Remarque : Méthode de Chernoff. L'hypothèse fondamentale dans l'inégalité de Hoeffding est l'aspect borné des variables aléatoires. On peut parfois obtenir des bornes exponentielles explicites en supposant "seulement" que la variable X admet des moments exponentiels, c'est-à-dire que $\mathbb{E}[\exp(\lambda X)] < \infty$ pour $\lambda \geq 0$. Dans ce cas, pour tout $c > 0$ et tout $\lambda > 0$, la croissance de la fonction $x \mapsto \exp(\lambda x)$ et l'inégalité de Markov permettent d'écrire

$$\mathbb{P}(X \geq c) = \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda c)) \leq \exp(-\lambda c))\mathbb{E}[\exp(\lambda X)] =: \varphi(\lambda),$$

et si on sait minimiser φ, ceci donne

$$\mathbb{P}(X \geq c) \leq \min_{\lambda \geq 0} \varphi(\lambda) = \varphi(\lambda_0) = \exp(-\lambda_0 c))\mathbb{E}[\exp(\lambda_0 X)].$$

Cette ruse aussi simple que puissante est connue sous le nom de méthode de Chernoff.

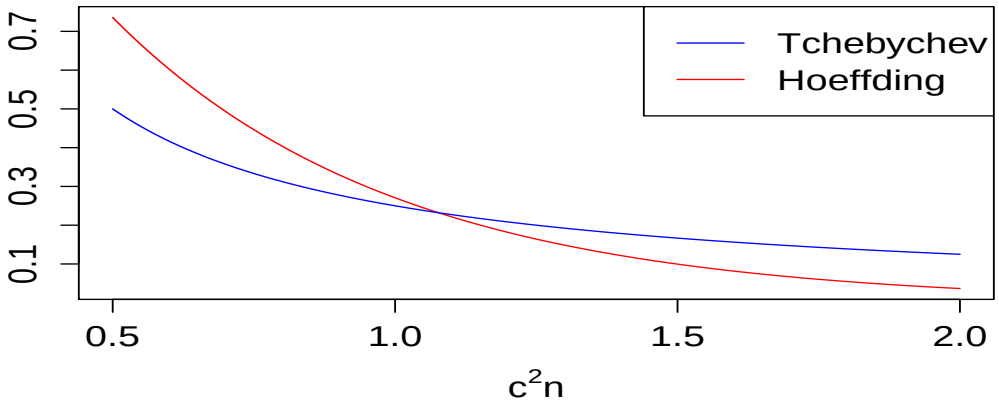

FIGURE 1.2 – Bornes de Bienaymé-Tchebychev et de Hoeffding pour $\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right)$.

1.1.3 Théorèmes asymptotiques

On revient à notre exemple : on veut estimer la probabilité θ de tomber sur Pile. Comme on l'a dit, un estimateur naturel est celui de la fréquence empirique d'apparition de Pile au cours des n premiers lancers, c'est-à-dire

$$\hat{\theta}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

On a démontré en section précédente que, lorsque le nombre de lancers tend vers l'infini, cette fréquence empirique $(\hat{\theta}_n)$ tend presque sûrement vers la fréquence théorique θ. Nous avons fait la démonstration "à la main", via Hoeffding et Borel-Cantelli. Il y a en fait un argument massue qui permettait directement de conclure : la Loi des Grands Nombres, qui est le premier grand résultat de convergence.

Théorème 2 (Loi des Grands Nombres)
Soit (X_n) une suite de variables aléatoires iid admettant une moyenne $m = \mathbb{E}[X_1]$, alors

$$\frac{S_n}{n} := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P} \text{ \& } p.s.} m.$$

On parle de Loi Forte des Grands Nombres pour la convergence presque sûre et de Loi faible des Grands Nombres pour la convergence en probabilité.

Exemple : dans notre exemple, les X_i étant effectivement iid avec $\mathbb{E}[X_1] = \theta$, on retrouve bien

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P} \text{ \& } p.s.} \theta.$$

La figure 1.3 représente une trajectoire $(\hat{\theta}_n)$ pour une pièce déséquilibrée (2 fois plus de chances de tomber sur Face que sur Pile).

En analyse, i.e. dans un cadre déterministe, une fois établi qu'une suite de nombres est convergente, l'étape suivante consiste à déterminer la vitesse de convergence vers cette limite. On peut se poser la même question dans un contexte stochastique. A quelle vitesse la suite de moyennes empiriques

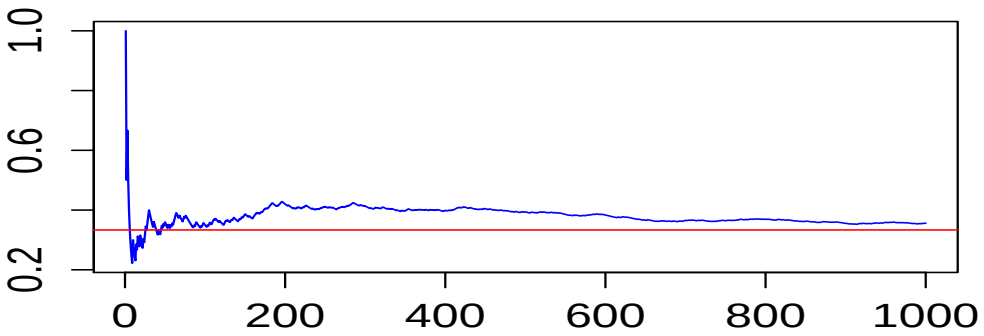

FIGURE 1.3 – Une réalisation de $\hat{\theta}_n$ pour $1 \leq n \leq 10^3$ lorsque $\theta = 1/3$.

(S_n/n) converge-t-elle vers la vraie moyenne m ? De façon générale, dès lors que les variables admettent un moment d'ordre 2 (c'est-à-dire hormis pour les lois à queues lourdes de type Cauchy, Pareto, etc.), cette vitesse est en $1/\sqrt{n}$, comme le montre le Théorème Central Limite, second grand résultat de convergence.

Théorème 3 (Théorème Central Limite)
Soit (X_n) une suite de variables aléatoires iid admettant une variance $\sigma^2 = \mathrm{Var}(X_1) > 0$, alors

$$\sqrt{n}\left(\frac{S_n}{n} - m\right) = \frac{S_n - nm}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - m) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

ce qui est équivalent à dire que

$$\frac{\sqrt{n}}{\sqrt{\sigma}}\left(\frac{S_n}{n} - m\right) = \frac{S_n - nm}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - m}{\sigma} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Remarques :

1. Noter que, par convention, le second paramètre de la gaussienne désignera toujours la variance, non l'écart-type. Ceci n'est pas le cas pour tous les logiciels, ainsi R adopte la convention inverse.
2. Le cas $\sigma^2 = 0$ est trivial, puisqu'alors $X_i = m$ presque sûrement et il en va de même pour S_n/n, donc la loi de $\sqrt{n}(S_n/n - m)$ est dégénérée : c'est un Dirac en 0.

Le fait que la loi normale apparaisse ainsi de façon universelle[2] comme limite de somme de variables convenablement centrées et normalisées est franchement remarquable. Le centrage et la normalisation ne recèlent quant à eux aucun mystère : en effet, puisque les X_i sont iid, on a

$$\mathbb{E}[S_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = nm \qquad \& \qquad \mathrm{Var}(S_n) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2.$$

2. en fait quasi-universelle, puisqu'elle suppose que les X_i admettent un moment d'ordre 2. Si on lève cette hypothèse, d'autres vitesses et d'autres lois limites apparaissent...

Le TCL nous dit que, si l'on additionne un grand nombre de variables iid, cette somme s'approche d'une gaussienne, donc de façon hautement non rigoureuse on écrirait que pour n "grand",

$$S_n = \sum_{i=1}^{n} X_i \overset{\mathcal{L}}{\approx} \mathcal{N}(nm, n\sigma^2)$$

écriture que l'on rend rigoureuse en centrant (soustraction de nm), réduisant (division par l'écart-type $\sigma\sqrt{n}$) et en passant à la limite, c'est-à-dire

$$\frac{S_n - nm}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

qui est exactement le Théorème Central Limite.

Exemple : dans notre exemple, on a donc

$$\frac{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

En particulier, en notant F_n la fonction de répartition de la variable de gauche (qui se déduit de celle d'une loi binomiale $\mathcal{B}(n,\theta)$ par translation et changement d'échelle), on déduit de la Proposition 1 que pour tout réel x,

$$F_n(x) := \mathbb{P}\left(\frac{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} \leq x\right) \xrightarrow[n\to\infty]{} \Phi(x).$$

Cette convergence simple, qui est en fait uniforme via la Proposition 2, est illustrée Figure 1.4.

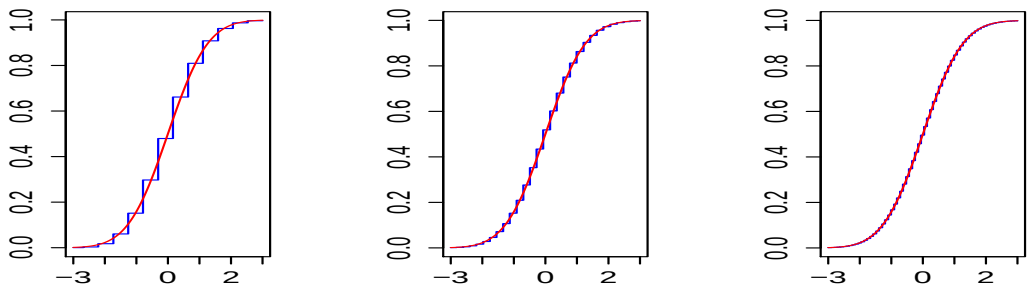

FIGURE 1.4 – Illustration du TCL via la convergence des fonctions de répartition F_n vers Φ pour le Pile ou Face avec $\theta = 1/3$ et respectivement $n = 20$, $n = 100$ et $n = 500$.

Supposons qu'on puisse appliquer le TCL avec la loi limite, alors avec un léger abus de notation, on aurait

$$\mathbb{P}\left(\left|\frac{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}{\sqrt{\theta(1-\theta)}}\right| \geq c\right) \approx \mathbb{P}\left(|\mathcal{N}(0,1)| \geq c\right) = 2(1 - \Phi(c)),$$

et en remplaçant c par $c\sqrt{n}/\sqrt{\theta(1-\theta)}$, il vient

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right) \approx 2\left(1 - \Phi(c\sqrt{n}/\sqrt{\theta(1-\theta)})\right).$$

Il est temps de rappeler l'encadrement de la queue de la gaussienne (voir Figure 1.5) :

$$\forall x>0 \qquad \left(\frac{1}{x}-\frac{1}{x^3}\right)\times\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\leq 1-\Phi(x)\leq\frac{1}{x}\times\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}. \tag{1.3}$$

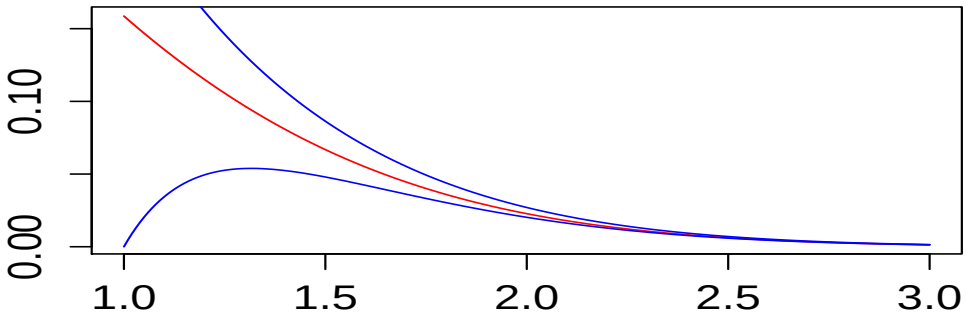

FIGURE 1.5 – Encadrement (1.3) de la queue de la gaussienne $\mathbb{P}(\mathcal{N}(0,1)\geq x)=1-\Phi(x)$.

Le minorant ne nous sera pas utile ici, il sert juste à montrer que si l'on note f la densité de la loi normale centrée réduite, l'équivalent $1-\Phi(x)\sim f(x)/x$ pour $x\to+\infty$ est très précis puisque l'erreur relative est en $1/x^2$. Dans notre situation, puisque $0<\theta(1-\theta)\leq 1/4$, ceci implique

$$\mathbb{P}\left(\left|\hat{\theta}_n-\theta\right|\geq c\right)\leq\frac{1}{\sqrt{n}}\times\sqrt{\frac{2\theta(1-\theta)}{\pi c^2}}\exp\left(-\frac{c^2 n}{2\theta(1-\theta)}\right)\leq\frac{1}{\sqrt{n}}\times\frac{1}{\sqrt{2\pi c^2}}\exp\left(-2c^2 n\right).$$

Ainsi, par rapport à la majoration (1.2) obtenue via Hoeffding et abstraction faite des constantes multiplicatives, le TCL permet de gagner un facteur $1/\sqrt{c^2 n}$.

Les calculs précédents supposent néanmoins que la moyenne centrée normalisée à gauche du TCL suit la loi normale standard de droite, ce qui n'est vrai qu'à la limite... Le résultat suivant permet de contrôler la vitesse de convergence vers la gaussienne.

Théorème 4 (Berry-Esseen)
Sous les mêmes hypothèses que dans le Théorème Central Limite, en supposant de plus que les variables X_i admettent un moment d'ordre 3, alors

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_n-nm}{\sigma\sqrt{n}}\leq x\right)-\Phi(x)\right|\leq\frac{\mathbb{E}\left[|X_1-m|^3\right]}{\sigma^3\sqrt{n}},$$

où Φ désigne la fonction de réparition de la loi normale centrée réduite.

La borne optimale est en fait un peu meilleure en ce sens qu'on peut prendre comme majorant $C\times\mathbb{E}\left[|X_1-m|^3\right]/(\sigma^3\sqrt{n})$, avec $C<1$. Cette constante C fait l'objet d'améliorations constantes : la meilleure valeur obtenue à ce jour est de l'ordre de $1/2$. Ceci importe peu : il convient surtout de voir que l'écart vertical maximal entre la fonction de répartition à l'étape n et celle de la gaussienne est en $1/\sqrt{n}$.

Exemple : puisque

$$\mathbb{E}\left[|X_1-\theta|^3\right] = \theta(1-\theta)\left(\theta^2+(1-\theta)^2\right) \leq \theta(1-\theta),$$

cette inégalité donne pour tout réel x et tout $n \in \mathbb{N}^*$,

$$\left|\mathbb{P}\left(\sqrt{n}\,\frac{\hat{\theta}_n-\theta}{\sigma} \leq x\right) - \Phi(x)\right| \leq \frac{1}{\sqrt{\theta(1-\theta)n}},$$

ce qui n'est pas reluisant : au mieux en $2/\sqrt{n}$ pour $\theta = 1/2$, et franchement désastreux lorsque θ est proche du bord de l'intervalle $[0,1]$. On paye ici le fait d'avoir une borne universelle (dès lors qu'il y a un moment d'ordre 3) et non asymptotique, la majoration par Berry-Esseen étant valable pour tout n.

1.1.4 Opérations sur les limites

Nous avons vu en Section 1.1.1 que la convergence presque sûre implique la convergence en probabilité. Quid du lien entre cette dernière et la convergence en loi ?

Proposition 5 (Convergence en probabilité ⇒ Convergence en loi)
Si la suite de variables aléatoires (X_n) converge en probabilité vers la variable X, alors (X_n) converge en loi vers X :

$$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to\infty]{\mathcal{L}} X.$$

La réciproque est fausse en générale, mais vraie si la limite est une constante :

$$X_n \xrightarrow[n\to\infty]{\mathcal{L}} a \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} a.$$

Dire que (X_n) tend en loi vers la constante a signifie que la loi des X_n tend vers un Dirac au point a, ou encore que pour toute fonction continue et bornée φ,

$$\mathbb{E}[\varphi(X_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[\varphi(a)] = \varphi(a).$$

Exercice : grâce à un développement limité à l'ordre 1 de la fonction caractéristique de $\hat{\theta}_n$, retrouver le fait que $(\hat{\theta}_n)$ converge en probabilité vers θ.

Une suite de variables aléatoires (X_n) étant donnée, il arrive souvent qu'on s'intéresse à son image par une fonction g, c'est-à-dire à la suite de variables aléatoires $(g(X_n))$. Question : si (X_n) converge en un certain sens, cette convergence est-elle préservée pour $(g(X_n))$? La réponse est oui si g est continue, comme le montre le résultat suivant, connu en anglais sous le nom de *Continuous Mapping Theorem*.

Théorème 5 (Théorème de continuité)
Soit (X_n) une suite de variables aléatoires, X une variable aléatoire, g une fonction dont l'ensemble des points de discontinuité est noté D_g. Si $\mathbb{P}(X \in D_g) = 0$, alors la suite $(g(X_n))$ hérite du mode de convergence de la suite (X_n) :

(a) Si (X_n) converge p.s. vers X, alors $(g(X_n))$ converge p.s. vers $g(X)$.

(b) Si (X_n) converge en probabilité vers X, alors $(g(X_n))$ converge en probabilité vers $g(X)$.

(c) Si (X_n) converge en loi vers X, alors $(g(X_n))$ converge en loi vers $g(X)$.

Si g est continue sur $\mathbb{R}$, aucun souci à se faire, mais cette condition est inutilement forte : ce qui importe à la limite, c'est la continuité de g là où la variable X a des chances de tomber. Or la condition $\mathbb{P}(X \in D_g) = 0$ assure justement que X ne tombe jamais là où g pose des problèmes, donc tout se passe bien.

Exemple : Si (X_n) tend vers X de l'une des trois manières ci-dessus et si X est presque sûrement non nulle, alors $(1/X_n)$ tend vers $1/X$ de la même manière.

Lorsque (x_n) et (y_n) sont deux suites de nombres réels tendant respectivement vers x et y, alors la suite $(x_n + y_n)$ tend vers $x + y$ et la suite $(x_n y_n)$ vers xy. Le Théorème de Slutsky propose un analogue de ce résultat pour la convergence en loi.

Théorème 6 (Théorème de Slutsky)
Si (X_n) converge en loi vers X et si (Y_n) converge en probabilité vers la constante a, alors

$$X_n + Y_n \xrightarrow[n\to\infty]{\mathcal{L}} X + a \qquad \text{et} \qquad X_n Y_n \xrightarrow[n\to\infty]{\mathcal{L}} aX.$$

Exemple : l'application du TCL à notre exemple a donné

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

On aimerait en déduire des intervalles de confiance pour θ, mais ce n'est pas possible sous cette forme car le dénominateur fait intervenir le paramètre θ inconnu. L'idée naturelle est de le remplacer par son estimateur $\hat{\theta}_n$ et, par conséquent, de considérer la suite de variables

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}.$$

Que dire de sa convergence ? Puisque $\hat{\theta}_n$ tend en probabilité vers θ par la loi (faible) des grands nombres, le Théorème de continuité assure que

$$\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)} \xrightarrow[n\to\infty]{\mathbb{P}} \sqrt{\theta(1-\theta)},$$

ou encore que

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

Il reste à appliquer le Théorème de Slutsky :

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} = \sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \times \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Ceci permet de construire des intervalles de confiance asymptotiques, encore faut-il être vigilant sur ladite asymptotique... Nous y reviendrons.

Le résultat suivant n'a rien d'étonnant et montre grosso modo qu'un TCL implique une convergence en probabilité.

Proposition 6
Soit (X_n) une suite de variables aléatoires, X une variable aléatoire et a un nombre réel tels que

$$\sqrt{n}\,(X_n - a) \xrightarrow[n\to\infty]{\mathcal{L}} X,$$

alors (X_n) converge en probabilité vers a.

Remarque : cet énoncé peut se généraliser en remplaçant $\sqrt{n}$ par une suite (v_n) de réels tendant vers $+\infty$. C'est du reste sous cette forme qu'on l'utilisera dans la Delta méthode ci-dessous.

Entre autres choses, cette Delta méthode explique l'action d'une application dérivable sur un résultat de type TCL. Elle précise en fait le premier terme non constant d'un développement limité aléatoire. En effet, par rapport au Théorème 5 qui est un résultat de continuité, celui-ci peut se voir comme un résultat de dérivabilité.

Théorème 7 (Delta méthode)
Soit (X_n) une suite de variables aléatoires et (v_n) une suite de réels tendant vers $+\infty$. Supposons qu'il existe un réel a et une variable X tels que

$$v_n\,(X_n - a) \xrightarrow[n\to\infty]{\mathcal{L}} X.$$

Si g est une fonction dérivable au point a, alors

$$v_n\,(g(X_n) - g(a)) \xrightarrow[n\to\infty]{\mathcal{L}} g'(a)X.$$

En particulier, si $X \sim \mathcal{N}(0, \sigma^2)$ et $g'(a) \neq 0$, alors

$$v_n\,(g(X_n) - g(a)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, (\sigma g'(a))^2).$$

Preuve. D'après la Proposition 6 et la remarque qui la suit,

$$v_n\,(X_n - a) \xrightarrow[n\to\infty]{\mathcal{L}} X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} a.$$

Dire que g est dérivable en a signifie qu'il existe une fonction r telle que

$$g(x) = g(a) + (x - a)(g'(a) + r(x)),$$

avec $\lim_{x\to a} r(x) = 0$. En d'autres termes, la fonction r est prolongeable par continuité en a, et ce en posant $r(a) = 0$. Puisque (X_n) converge en probabilité vers a, on déduit du Théorème de continuité que la suite $(r(X_n))$ converge en probabilité vers $r(a) = 0$. Nous avons donc le développement limité aléatoire

$$g(X_n) = g(a) + (X_n - a)(g'(a) + r(X_n)),$$

avec

$$g'(a) + r(X_n) \xrightarrow[n\to\infty]{\mathbb{P}} g'(a).$$

Il ne reste plus qu'à appliquer le Théorème de Slutsky :

$$v_n\,(g(X_n) - g(a)) = (g'(a) + r(X_n)) \times v_n\,(X_n - a) \xrightarrow[n\to\infty]{\mathcal{L}} g'(a)X.$$

■

Remarque : si $g'(a) = 0$, alors $g'(a)X = 0$ et la limite est la loi de Dirac en 0.

La plupart du temps, ce résultat sera appliqué avec $v_n = \sqrt{n}$ et $X \sim \mathcal{N}(0, \sigma^2)$. Voici une façon heuristique de le retrouver (le degré zéro de la rigueur, mais l'idée est là) :

$$g(X_n) \approx g\left(a + \frac{X}{\sqrt{n}}\right) \approx g(a) + g'(a)\frac{X}{\sqrt{n}} \implies \sqrt{n}\,(g(X_n) - g(a)) \approx g'(a)X \sim \mathcal{N}(0, (g'(a)\sigma)^2).$$

Exemple : nous avons vu que

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta)).$$

La convergence en loi de la suite de variables aléatoires $(1/\hat{\theta}_n)$ est alors une conséquence directe de la Delta méthode :

$$\sqrt{n}\left(\frac{1}{\hat{\theta}_n} - \frac{1}{\theta}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, (1-\theta)/\theta^3).$$

1.1.5 Absolue continuité et densités

Cette section rappelle quelques résultats de théorie de la mesure utiles dans la suite pour définir la notion de modèle statistique dominé. De façon très générale, on considère un ensemble E muni d'une tribu (ou σ-algèbre) $\mathcal{E}$ et d'une mesure positive μ, c'est-à-dire une application de $\mathcal{E}$ dans $[0, +\infty]$ vérifiant $\mu(\emptyset) = 0$ et, pour toute suite (A_n) d'ensembles de $\mathcal{E}$ deux à deux disjoints, la propriété de σ-additivité :

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Deux exemples d'espaces mesurés nous intéresseront plus particulièrement dans tout ce cours, l'un relatif aux variables discrètes, l'autre aux variables à densité.

Exemples :

1. Mesure de comptage : $(E, \mathcal{E}, \mu) = (\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, où $\mathcal{P}(\mathbb{N})$ désigne l'ensemble de toutes les parties de $\mathbb{N}$ et μ la mesure de comptage, qui à un ensemble A associe son cardinal, noté $|A|$ et éventuellement infini. On peut décrire μ par l'intermédiaire des mesures de Dirac :

$$\mu = \sum_{n=0}^{+\infty} \delta_n \implies \mu(A) = \sum_{n=0}^{+\infty} \delta_n(A) = |A|.$$

2. Mesure de Lebesgue : $(E, \mathcal{E}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, où $\mathcal{B}(\mathbb{R})$ désigne la tribu borélienne de la droite réelle (i.e. la tribu engendrée par les intervalles ouverts) et λ la mesure de Lebesgue, qui à un intervalle associe sa longueur, éventuellement infinie. Avec la notation (a, b) définie précédemment, ceci s'écrit :

$$-\infty \leq a < b \leq +\infty \implies \lambda((a, b)) = b - a,$$

avec la convention classique : $+\infty - a = +\infty - (-\infty) = b - (-\infty) = +\infty$.

Ces deux mesures ne sont pas finies puisque $\mu(\mathbb{N}) = \lambda(\mathbb{R}) = \infty$, mais elles sont σ-finies.

Définition 3 (Mesure σ-finie)
Soit $(E, \mathcal{E}, \mu)$ un espace mesuré. On dit que la mesure μ est σ-finie s'il existe une suite (E_n) d'ensembles mesurables tels que $\mu(E_n) < \infty$ pour tout n et

$$E = \bigcup_{n=1}^{\infty} E_n.$$

Autrement dit, il existe un recouvrement de E par des sous-ensembles de mesures finies.

Exemples :

1. Mesure de comptage : il suffit de prendre $E_n = \{0, \ldots, n\}$.
2. Mesure de Lebesgue : les intervalles $E_n = [-n, n]$ font l'affaire.

La notion suivante correspond à une relation de préordre (réfléxivité et transitivité) entre mesures.

Définition 4 (Absolue continuité)
Soit $(E, \mathcal{E})$ un espace mesurable, λ et μ deux mesures positives sur cet espace. On dit que μ est absolument continue par rapport à λ, noté $\mu \ll \lambda$, si tout ensemble mesurable A négligeable pour λ l'est aussi pour μ :

$$\forall A \in \mathcal{E} \qquad \lambda(A) = 0 \implies \mu(A) = 0.$$

On dit que λ et μ sont équivalentes si $\mu \ll \lambda$ et $\lambda \ll \mu$, auquel cas elles ont les mêmes ensembles négligeables.

Lorsque les mesures λ et μ sont σ-finies, on retrouve la notion de densité de μ par rapport à λ, bien connue pour les variables aléatoires.

Théorème 8 (Radon-Nikodym)
Soit $(E, \mathcal{E})$ un espace mesurable, λ et μ deux mesures positives σ-finies sur cet espace. Si μ est absolument continue par rapport à λ, alors μ a une densité par rapport à λ, c'est-à-dire qu'il existe une fonction f mesurable et positive, notée $f = d\mu/d\lambda$, telle que pour toute fonction μ-intégrable φ, on ait

$$\int_E \varphi(x)\mu(dx) = \int_E \varphi(x)\frac{d\mu}{d\lambda}(x)\lambda(dx) = \int_E \varphi(x)f(x)\lambda(dx).$$

Notation : dans ce cas on note $\mu = f \cdot \lambda$.

Remarque : ça ne marche plus si les mesures ne sont pas supposées σ-finies. En effet, considérons $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et λ la mesure de comptage sur cet espace, c'est-à-dire que $\lambda(B) = |B|$ pour tout borélien B. Ainsi $\lambda(B) = 0$ si et seulement si B est l'ensemble vide. Dès lors, toute mesure sur $(E, \mathcal{E})$ est absolument continue par rapport à λ. Ceci est en particulier vrai pour la mesure de Lebesgue $\mu(dx) = dx$. Pourtant, celle-ci n'admet pas de densité par rapport à la mesure de comptage, sinon il existerait une fonction f telle que pour toute indicatrice $\varphi = \mathbb{1}_a$, on ait

$$0 = \int_{\mathbb{R}} \mathbb{1}_a(x)dx = \int_{\mathbb{R}} \mathbb{1}_a(x)\mu(dx) = \int_{\mathbb{R}} \mathbb{1}_a(x)f(x)\lambda(dx) = f(a),$$

d'où, pour $\varphi = \mathbb{1}_{[0,1]}$,

$$1 = \int_0^1 dx = \int_{\mathbb{R}} \mathbb{1}_{[0,1]}(x)dx = \int_{\mathbb{R}} \mathbb{1}_{[0,1]}(x)f(x)\lambda(dx) = 0,$$

ce qui est absurde.

Exemples :

1. Mesure de comptage : soit X une variable aléatoire discrète, c'est-à-dire à valeurs dans $\mathbb{N}$ ou un sous-ensemble de $\mathbb{N}$. Sa loi P_X définit une mesure de probabilité sur $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. D'après la remarque ci-dessus, il est clair que cette mesure est absolument continue par rapport à la mesure de comptage μ sur $\mathbb{N}$, la densité f correspondant aux poids $p_n = P_X(\{n\}) = \mathbb{P}(X = n)$ puisque

$$\int_{\mathbb{N}} \varphi(x)P_X(dx) = \sum_{n=0}^{\infty} \varphi(n)\mathbb{P}(X = n) = \int_{\mathbb{N}} \varphi(x)\mathbb{P}(X = x)\mu(dx).$$

2. Mesure de Lebesgue : une variable aléatoire est dite absolument continue (sous-entendu : par rapport à la mesure de Lebesgue) ou à densité (même sous-entendu) s'il existe une fonction f mesurable positive d'intégrale 1 par rapport à la mesure de Lebesgue $\lambda(dx) = dx$ et telle que pour toute fonction P_X-intégrable φ, on ait

$$\int_{\mathbb{R}} \varphi(x)P_X(dx) = \int_{\mathbb{R}} \varphi(x)f(x)dx.$$

Notons enfin les deux propriétés suivantes :

1. Mesures équivalentes : les mesures σ-finies λ et μ sont équivalentes si et seulement si leurs densités respectives sont presque partout [3] strictement positives et inverses l'une de l'autre :

$$\frac{d\mu}{d\lambda}(x) \times \frac{d\lambda}{d\mu}(x) = 1 \qquad p.p.$$

3. au sens de λ comme de μ, puisqu'elles sont équivalentes.

2. Transitivité : soit λ, μ et ν trois mesures σ-finies, alors si $\nu \ll \mu$ et $\mu \ll \lambda$, on a $\nu \ll \lambda$ avec

$$\frac{d\nu}{d\lambda}(x) = \frac{d\nu}{d\mu}(x) \times \frac{d\mu}{d\lambda}(x) \qquad \nu - p.p.$$

Remarque. Dans toute la suite de ce cours, même si ce n'est pas précisé, toutes les mesures considérées seront supposées sigma-finies, de même que toutes les fonctions considérées seront supposées mesurables.

1.2 Modèles statistiques

La démarche statistique comporte généralement deux étapes. La première est une phase de modélisation, qui consiste à mettre un phénomène réel sous forme mathématique. En pratique, ceci revient à supposer que l'observation $\mathbf{X}$ est un objet aléatoire dont la loi $P_{\mathbf{X}}$ (inconnue !) appartient à une famille de lois $(P_\theta)_{\theta\in\Theta}$ que l'on spécifie. Cette étape préliminaire, cruciale, est en grande partie une affaire de praticien : pour chaque domaine d'application (biologie, physique, chimie, etc.), ce sont les spécialistes du domaine qui fourniront cette modélisation.

Ceci étant supposé acquis, la seconde étape est celle qui nous occupe dans ce cours, à savoir l'inférence statistique, ou statistique inférentielle. Il s'agit, à partir du modèle $(P_\theta)_{\theta\in\Theta}$ et de l'observation $\mathbf{X}$, de retirer l'information la plus pertinente possible sur les paramètres en jeu dans le modèle, c'est-à-dire dans la loi de $\mathbf{X}$. On rappelle que si $\mathbf{X}$ est un objet aléatoire (variable, vecteur, processus) à valeurs dans un espace mesurable $(E, \mathcal{E})$, sa loi $P_{\mathbf{X}}$ est définie pour tout A de $\mathcal{E}$ par :

$$P_{\mathbf{X}}(A) = \mathbb{P}(\mathbf{X} \in A) = \mathbb{P}(\{\omega \in \Omega :\ \mathbf{X}(\omega) \in A\}),$$

probabilité que l'objet aléatoire $\mathbf{X}$ tombe dans l'ensemble A. Résumons ce qui vient d'être dit.

Définition 5 (Expérience statistique)
Une expérience statistique est la donnée d'un objet aléatoire $\mathbf{X}$ à valeurs dans un espace mesurable $(E, \mathcal{E})$ et d'une famille de lois $(P_\theta)_{\theta\in\Theta}$ sur cet espace, supposée contenir la loi $P_{\mathbf{X}}$, et appelée modèle statistique pour la loi de $\mathbf{X}$.

Dans cette définition, l'hypothèse fondamentale est bien entendu qu'il existe une valeur $\theta \in \Theta$ telle que $P_{\mathbf{X}} = P_\theta$. Le vrai paramètre θ est inconnu mais l'espace Θ dans lequel il vit est, lui, supposé connu.

Exemples :

1. Dans le jeu de Pile ou Face, on a donc $E = \{0, 1\}^n$. Puisque E est fini, on le munit naturellement de la tribu $\mathcal{E} = \mathcal{P}(E)$ de toutes les parties de E. L'objet aléatoire est ici le n-uplet $\mathbf{X} = (X_1, \ldots, X_n)$. Comme le résultat de chaque lancer suit une loi de Bernoulli $\mathcal{B}(\theta)$, pour un certain paramètre inconnu $\theta \in \Theta =]0, 1[$, et puisque ces lancers sont iid, le modèle statistique est la famille de lois

$$(P_\theta)_{\theta\in\Theta} = \left(\mathcal{B}(\theta)^{\otimes n}\right)_{\theta\in]0,1[}.$$

Autrement dit, toute réalisation $\mathbf{x} = (x_1, \ldots, x_n)$ de $\mathbf{X}$ a, sous P_θ, la probabilité

$$P_\theta(\mathbf{x}) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{s_n}(1-\theta)^{n-s_n},$$

où $s_n = x_1 + \cdots + x_n$ correspond au nombre de Pile dans le n-uplet $\mathbf{x} = (x_1, \ldots, x_n)$.

2. Dans une population donnée, la taille des hommes adultes est modélisée par une loi normale de moyenne et variance inconnues. On veut estimer ces dernières à partir d'un échantillon de n hommes pris au hasard dans la population. On considère cette fois $E = \mathbb{R}^n$ muni de la tribu borélienne $\mathcal{E} = \mathcal{B}(\mathbb{R}^n)$. L'objet aléatoire est le n-uplet $\mathbf{X} = (X_1, \ldots, X_n)$ avec les X_i iid suivant une certaine loi normale $\mathcal{N}(m, \sigma^2)$. Dans ce cas, $\theta = (m, \sigma^2)$ et $\Theta = \mathbb{R} \times \mathbb{R}_+^\star$. La famille de lois est donc

$$(P_\theta)_{\theta\in\Theta} = \left(\mathcal{N}(m, \sigma^2)^{\otimes n}\right)_{(m,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^\star}.$$

Notons qu'on peut aussi prendre $\theta = (m, \sigma)$ en fonction du contexte.

Dans ces deux exemples, le vecteur $\mathbf{X} = (X_1, \ldots, X_n)$ est un échantillon de variables X_i iid appelées des observations [4]. Lorsque, comme dans ces exemples, ces variables sont iid de loi commune Q_θ, c'est-à-dire que

$$(P_\theta)_{\theta\in\Theta} = \left(Q_\theta^{\otimes n}\right)_{\theta\in\Theta},$$

on parle de modèle d'échantillonnage. Dans ce cas, on appellera indifféremment $(P_\theta)_{\theta\in\Theta}$ ou $(Q_\theta)_{\theta\in\Theta}$ le modèle statistique en question. Ce n'est bien sûr pas le seul cadre envisageable, comme nous le verrons sur le modèle de régression linéaire du chapitre suivant. Par ailleurs, ces deux exemples ont un autre point commun : la taille de l'espace des paramètres.

Définition 6 (Modèle paramétrique)
Si l'espace Θ des paramètres du modèle statistique $(P_\theta)_{\theta\in\Theta}$ est contenu dans $\mathbb{R}^k$ pour un certain $k \in \mathbb{N}^\star$, on parle de modèle paramétrique. Sinon, il est non paramétrique.

Exemples :

1. Jeu de Pile ou Face : $\Theta =]0, 1[\subseteq \mathbb{R}$, donc c'est un problème paramétrique unidimensionnel.
2. Taille : $\Theta = \mathbb{R} \times \mathbb{R}_+^\star \subseteq \mathbb{R}^2$, problème paramétrique bidimensionnel.
3. Considérons que la taille des hommes ne soit pas supposée suivre une loi normale, mais une loi inconnue sur $[0.5; 2.5]$. On suppose, ce qui est raisonnable, que cette loi a une densité f par rapport à la mesure de Lebesgue. Dans ce cas, Θ correspond à l'ensemble des densités sur $[0.5; 2.5]$, qui est clairement de dimension infinie. C'est donc un modèle non paramétrique. Dans ce genre de situation, afin d'éviter des espaces fonctionnels trop gros, on met en général des contraintes supplémentaires sur la densité, typiquement des hypothèses de régularité.

Remarque : tout modèle statistique est un modèle **approché** de la réalité. Lorsqu'on suppose par exemple que la répartition des tailles suit une loi normale, il y a a priori incompatibilité entre le fait qu'une gaussienne est à valeurs dans $\mathbb{R}$ tout entier et le fait que ladite taille est à valeurs dans $\mathbb{R}_+$ (et même dans $[0.5; 2.5]$). Ceci pourrait faire croire que le modèle adopté est inadapté, sauf que cet argument n'en est pas un, car les variables gaussiennes sont essentiellement bornées (voir Figure 1.6). En effet, si $X \sim \mathcal{N}(0, 1)$, la probabilité que X ne tombe pas dans l'intervalle $[-8, 8]$ est de l'ordre [5] de 10^{-15}. Ainsi, même en considérant un échantillon d'un milliard de gaussiennes, la probabilité que l'une d'entre elles sorte de cet intervalle est inférieure à une chance sur un million. Bref, pour les valeurs de n que l'on considère en pratique, un échantillon de n gaussiennes est indiscernable d'une suite de variables à support dans $[-8, 8]$. De façon générale, un modèle statistique est toujours une approximation de la réalité, mais ceci n'est pas un problème tant que les conclusions que l'on tire de ce modèle approché restent fiables.

Passons à un autre point. Notre but étant d'approcher la vraie valeur θ du paramètre, encore faut-il que celui-ci soit défini sans ambiguïté. C'est le principe d'identifiabilité qui est ici à l'œuvre.

4. avec un léger abus de langage, le même terme servant à qualifier $\mathbf{X}$, voire $\mathbf{x} = (x_1, \ldots, x_n)$.
5. en R, il suffit de taper la commande : `2*(1-pnorm(8))`

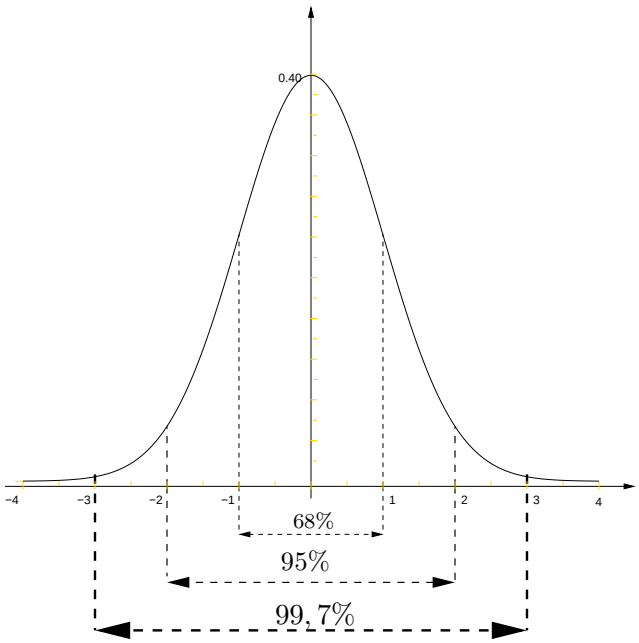

FIGURE 1.6 – Concentration de la loi normale standard autour de sa moyenne.

Définition 7 (Identifiabilité)
Le modèle statistique $(P_\theta)_{\theta\in\Theta}$ est dit identifiable si l'application $\theta \mapsto P_\theta$ est injective, c'est-à-dire si deux paramètres distincts ne peuvent correspondre à la même loi.

Exemple : le modèle gaussien $\left(\mathcal{N}(m,\sigma^2)\right)_{m\in\mathbb{R},\sigma>0}$ est identifiable. Par contre, le modèle alternatif $\left(\mathcal{N}(m,\sigma^2)\right)_{m\in\mathbb{R},\sigma\neq 0}$ ne l'est pas puisque $\mathcal{N}(m,\sigma^2) = \mathcal{N}(m,(-\sigma)^2)$.

Dans toute la suite, tous les modèles seront supposés identifiables. Nous concluons cette section par une définition permettant de ramener une famille de lois à une famille de densités. Elle fait appel aux notions rappelées en Section 1.1.5.

Définition 8 (Modèle statistique dominé)
Le modèle statistique $(P_\theta)_{\theta\in\Theta}$ sur $(E,\mathcal{E})$ est dit dominé s'il existe une mesure σ-finie λ sur $(E,\mathcal{E})$ telle que, pour tout $\theta\in\Theta$, on ait $P_\theta \ll \lambda$. La mesure λ est alors appelée mesure dominante.

Dans le classique modèle d'échantillonnage où $P_\theta = Q_\theta^{\otimes n}$, il est clair que $Q_\theta \ll \lambda$ si et seulement si $P_\theta \ll \lambda^{\otimes n}$. On parlera donc de mesure dominante aussi bien pour P_θ que pour Q_θ. En particulier, si $Q_\theta = f_\theta\cdot\lambda$, alors la loi P_θ a pour densité $f_\theta(x_1)\times\cdots\times f_\theta(x_n)$ par rapport à la mesure dominante $\lambda^{\otimes n}$.

Exemples :

1. Jeu de Pile ou Face : une mesure dominante de $Q_\theta = (1-\theta)\delta_0 + \theta\delta_1$ est $\lambda = \delta_0+\delta_1$, mesure de comptage sur $\{0,1\}$.
2. Taille : le modèle est dominé par la mesure de Lebesgue sur $\mathbb{R}$.
3. Si $(E,\mathcal{E}) = (\mathbb{R},\mathcal{B}(\mathbb{R}))$, le modèle $(\delta_\theta)_{\theta\in\mathbb{R}}$ des mesures de Dirac ne peut être dominé. En effet, supposons qu'il existe une mesure σ-finie λ telle que $\delta_\theta \ll \lambda$ pour tout réel θ. Alors, d'après le Théorème de Radon-Nikodym, il existe une fonction f_θ telle que $\delta_\theta = f_\theta\cdot\lambda$, d'où en particulier
$$1 = \delta_\theta(\{\theta\}) = f_\theta(\theta)\times\lambda(\{\theta\}) \implies \lambda(\{\theta\}) > 0.$$

Puisque λ est σ-finie, il existe un recouvrement de $\mathbb{R}$ par une suite (E_n) de boréliens tels que $\lambda(E_n) < \infty$ pour tout n. Or, puisque $\lambda(\{\theta\}) > 0$ pour tout θ, la somme

$$\lambda(E_n) = \sum_{\theta \in E_n} \lambda(\{\theta\})$$

ne peut être finie que si E_n est au plus dénombrable. Une union d'ensembles au plus dénombrables étant au plus dénombrable, l'union des E_n ne peut être égale à $\mathbb{R}$.

En pratique, deux mesures dominantes nous serviront constamment : la mesure de comptage si E est au plus dénombrable, la mesure de Lebesgue si $E = \mathbb{R}^d$.

1.3 Les problèmes statistiques classiques

Dans toute cette section, on considère le cadre d'une expérience statistique telle que décrite dans la Définition 5 et on inventorie quelques questions classiques en statistique inférentielle. Comme précédemment, l'exemple du jeu de Pile ou Face servira de fil rouge pour illustrer le propos.

1.3.1 Estimation

La première question que l'on se pose est celle de l'estimation du vrai paramètre θ.

Définition 9 (Statistique et Estimateur)
Une statistique $T(\mathbf{X})$ est une fonction mesurable de l'objet aléatoire $\mathbf{X}$ et éventuellement de paramètres connus, mais qui ne dépend pas de θ. Un estimateur de θ est une statistique $\hat{\theta} = \hat{\theta}(\mathbf{X})$ destinée à approcher θ.

Exemple : pour le jeu de Pile ou Face, la variable

$$S_n = X_1 + \cdots + X_n$$

est bien une statistique, puisqu'elle ne dépend que de l'observation $\mathbf{X} = (X_1, \ldots, X_n)$, mais ce n'est clairement pas un estimateur de θ, contrairement à la fréquence empirique

$$\hat{\theta}_n = \frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n},$$

qui est effectivement une approximation aléatoire de θ.

Remarques :

1. Un estimateur est censé approcher le paramètre d'intérêt, le rôle plus général d'une statistique étant de fournir des informations de diverses natures.
2. Dans la pratique, c'est la réalisation de l'estimateur qui fournit une estimation de θ : on l'appelle parfois l'estimée. Ainsi, si $\mathbf{x} = (x_1, \ldots, x_n)$ est une réalisation de $\mathbf{X} = (X_1, \ldots, X_n)$ de loi P_θ, on peut calculer l'approximation $\hat{\theta}(\mathbf{x})$ de θ.
3. On peut vouloir estimer une fonction $g(\theta)$ du paramètre θ, par exemple $g(\theta) = \theta^2$. Dans ce cas, un estimateur sera une statistique $\hat{g}(\mathbf{X})$. Si g est régulière et que l'on dispose déjà d'un "bon" estimateur $\hat{\theta}$ du paramètre θ, un estimateur naturel est $\hat{g}(\mathbf{X}) = g(\hat{\theta})$.

Le but de l'estimateur $\hat{\theta}$ étant d'approcher θ, encore faut-il préciser en quel sens. Une manière classique de quantifier la précision d'un estimateur est de passer par son risque quadratique.

Définition 10 (Risque quadratique)
Etant donné une expérience statistique telle que $\Theta \subseteq \mathbb{R}$, le risque quadratique, ou erreur quadratique moyenne, de l'estimateur $\hat{\theta}$ est défini pour tout $\theta \in \Theta$ par

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

Remarques :

1. Dans cette définition, le calcul d'espérance se fait en supposant que l'observation $\mathbf{X}$ suit la loi P_θ, c'est-à-dire que

$$R(\hat{\theta}, \theta) = \mathbb{E}\left[\left(\hat{\theta}(\mathbf{X}) - \theta\right)^2\right] = \int_E \left(\hat{\theta}(\mathbf{x}) - \theta\right)^2 P_\theta(\mathbf{dx}).$$

   C'est pourquoi on note parfois $\mathbb{E}_\theta$ au lieu de $\mathbb{E}$, Var_θ au lieu de Var et $\mathbb{P}_\theta$ au lieu de $\mathbb{P}$. **Afin d'alléger les écritures**, nous n'adoptons pas cette convention, mais il convient de garder constamment en tête la valeur du paramètre par rapport à laquelle on calcule les probabilités, espérances et variances.

2. Lorsque Θ est un espace métrique muni de la distance d, cette définition se généralise sans problème :

$$R(\hat{\theta}, \theta) = \mathbb{E}\left[d\left(\theta, \hat{\theta}\right)^2\right].$$

   L'exemple le plus courant est celui où $\Theta \subseteq \mathbb{R}^k$ avec d correspondant à la distance euclidienne.

L'inégalité de Markov de la Proposition 3 avec $p = 2$ et $X = (\hat{\theta} - \theta)$ donne

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

Par conséquent, si le risque quadratique est petit, l'estimateur $\hat{\theta}$ est proche de θ avec une grande probabilité. D'autre part, le risque quadratique admet la décomposition fondamentale suivante, dite de biais-variance.

Lemme 1 (Décomposition biais-variance)
Avec les notations de la définition du risque quadratique, on a

$$R(\hat{\theta}, \theta) = \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 + \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] =: B(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta}).$$

Le terme $B(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$ est appelé biais de l'estimateur $\hat{\theta}$. S'il est nul, on dit que l'estimateur est sans biais ou non biaisé.

Preuve : Il suffit de l'écrire

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

Dans cette expression, le terme $B(\hat{\theta}) := \left(\mathbb{E}[\hat{\theta}] - \theta\right)$ est déterministe donc en prenant l'espérance, il vient

$$R(\hat{\theta}, \theta) = \left(\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right]\right)^2 + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 = \mathrm{Var}(\hat{\theta}) + B(\hat{\theta})^2.$$

■

Remarques :

1. Si le paramètre θ a une unité, le biais se mesure avec cette même unité, tandis que la variance se mesure avec cette unité au carré. Ne serait-ce que pour des raisons d'homogénéité des grandeurs, il est donc logique d'ajouter le carré du biais à la variance.
2. Le biais mesure l'erreur moyenne faite par l'estimateur $\hat{\theta}$, tandis que le terme de variance mesure les fluctuations de $\hat{\theta}$ autour de sa moyenne. Un estimateur sera donc d'autant meilleur que son biais et sa variance sont **tous deux** faibles.
3. une inspection rapide de la preuve montre que cette décomposition biais-variance se généralise en dimension supérieure lorsque $\Theta \subseteq \mathbb{R}^k$ est équipé de la distance euclidienne, notée $\|\cdot\|$. Elle s'écrit alors

$$R(\hat{\theta},\theta) = \mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|^2\right] = \left\|\mathbb{E}[\hat{\theta}]-\theta\right\|^2 + \mathbb{E}\left[\left\|\hat{\theta}-\mathbb{E}[\hat{\theta}]\right\|^2\right] = \sum_{i=1}^{k}\left(B(\hat{\theta}_i)^2 + \mathrm{Var}(\hat{\theta}_i)\right),$$

ce qui donne finalement

$$R(\hat{\theta},\theta) = \sum_{i=1}^{k} R(\hat{\theta}_i,\theta_i),$$

c'est-à-dire que l'erreur quadratique globale est la somme des erreurs quadratiques sur chaque composante.

Exemple : dans l'exemple du Pile ou Face, $\hat{\theta} = \hat{\theta}_n$ et tous les calculs ont déjà été faits. Nous avons vu que $\mathbb{E}[\hat{\theta}_n] = \theta$ donc il est sans biais, d'où un risque quadratique égal à

$$R(\hat{\theta}_n,\theta) = \mathrm{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n} \leq \frac{1}{4n} \xrightarrow[n\to\infty]{} 0.$$

1.3.2 Intervalles de confiance

Toujours dans l'exemple du jeu de Pile ou Face, supposons qu'on vous dise : après n lancers, on a obtenu 60% de Pile. Devez-vous en déduire que la pièce n'est pas équilibrée ? Il est clair que votre réponse dépendra du nombre n de lancers. En effet, si $n = 10$, alors même si la pièce est équilibrée, la variables S_n du nombre de Pile suit une loi binomiale $\mathcal{B}(10, 0.5)$ et la probabilité d'observer au moins 6 Pile est environ égale à 38%. Bref, on ne peut rien en conclure.

A contrario, si $n = 1000$, on a cette fois $S_n \sim \mathcal{B}(1000, 0.5)$, laquelle est très bien approchée par une loi gaussienne. Précisément, le Théorème Central Limite nous assure que $(S_n - 500)/\sqrt{250}$ suit approximativement une loi normale centrée réduite donc, modulo cette approximation [6],

$$\mathbb{P}(S_n \geq 600) = \mathbb{P}\left(\frac{S_n - 500}{\sqrt{250}} \geq \frac{100}{\sqrt{250}}\right) \approx \mathbb{P}(\mathcal{N}(0,1) \geq 6.32) \approx 10^{-10}.$$

Cette fois, le doute n'est plus permis : il est à peu près certain que le dé est déséquilibré.

Au final, on voit que notre confiance dans l'estimateur est très fortement liée à sa loi et, par là, à la taille de l'échantillon dont on dispose. L'objet des intervalles de confiance est justement de formaliser ce point.

Définition 11 (Intervalle de confiance)
Supposons $\Theta \subseteq \mathbb{R}$ et fixons $\alpha \in]0,1[$ (petit, par exemple 5%). On appelle intervalle de confiance pour θ de niveau $(1-\alpha)$ tout intervalle aléatoire $(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X}))$ dont les deux bornes sont des statistiques et tel que, pour tout $\theta \in \Theta$,

$$\mathbb{P}(\theta \in (\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X}))) \geq 1-\alpha.$$

6. qui est en fait excellente car $\theta = 1/2$.

Achtung ! Il ne faut pas confondre l'intervalle de confiance (qui est aléatoire) et sa réalisation $(\underline{\theta}(\mathbf{x}), \overline{\theta}(\mathbf{x}))$, qui ne l'est pas ! Ainsi, écrire

$$\mathbb{P}(0.48 \leq \theta \leq 0.52) = 0.95$$

n'a strictement aucun sens puisque cette probabilité vaut 0 ou 1. On se contentera de dire que $[0.48; 0.52]$ est un intervalle de confiance à 95% pour θ.

Remarques :

1. Les deux critères de qualité d'un intervalle de confiance sont sa longueur et son niveau. Ceux-ci étant antagonistes, il s'agit de réaliser un compromis. Ainsi, pour un niveau de confiance donné (par exemple 95%), on cherchera un intervalle de confiance de plus petite longueur possible.
2. Si l'on ne suppose plus $\Theta \subseteq \mathbb{R}$, on appelle domaine (ou région) de confiance de niveau $(1-\alpha)$ tout ensemble aléatoire $D(\mathbf{X})$ ne dépendant ni de θ ni d'autres quantités inconnues et tel que
$$\forall \theta \in \Theta \qquad \mathbb{P}(\theta \in D(\mathbf{X})) \geq 1 - \alpha.$$

Exemple : on revient au jeu de Pile ou Face, pour lequel on applique les bornes vues en Section 1.1.2. L'inégalité de Tchebychev nous a permis d'écrire que, pour tout $c > 0$,

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right) \leq \frac{\theta(1-\theta)}{c^2 n} \leq \frac{1}{4c^2 n} \iff \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \leq c\right) \geq 1 - \frac{1}{4c^2 n}.$$

En prenant $c = 1/(2\sqrt{n\alpha})$, on en déduit que

$$\mathbb{P}\left(\hat{\theta}_n - \frac{1}{2\sqrt{n\alpha}} \leq \theta \leq \hat{\theta}_n + \frac{1}{2\sqrt{n\alpha}}\right) \geq 1 - \alpha,$$

c'est-à-dire que $[\hat{\theta}_n - 1/(2\sqrt{n\alpha}), \hat{\theta}_n + 1/(2\sqrt{n\alpha})]$ est un intervalle de confiance de niveau $(1-\alpha)$ pour θ. Ceci donne, pour $\alpha = 5\%$, un intervalle de confiance de largeur $2.24/\sqrt{n}$.

Par l'inégalité de Hoeffding, nous avons obtenu

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right) \leq 2\exp\left(-2c^2 n\right) \iff \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \leq c\right) \geq 1 - 2\exp\left(-2c^2 n\right),$$

donc en posant $c = \sqrt{-\log(\alpha/2)/(2n)}$, on obtient le nouvel intervalle de confiance

$$\mathbb{P}\left(\hat{\theta}_n - \sqrt{\frac{-\log(\alpha/2)}{2n}} \leq \theta \leq \hat{\theta}_n + \sqrt{\frac{-\log(\alpha/2)}{2n}}\right) \geq 1 - \alpha.$$

Cet intervalle est plus petit que celui donné par Tchebychev si et seulement si

$$\sqrt{\frac{-\log(\alpha/2)}{2n}} \leq \frac{1}{2\sqrt{n\alpha}} \iff -2\alpha\log(\alpha/2) \leq 1 \iff 0 < \alpha \leq 0.23,$$

ce qui correspond bien aux valeurs de α pertinentes pour des intervalles de confiance à 90, 95 ou 99%. A titre d'exemple, l'intervalle de confiance à 95% fourni par Hoeffding est de rayon $1.36/\sqrt{n}$, effectivement plus petit que celui obtenu par Tchebychev.

Ces intervalles de confiance sont valables pour tout n. Lorsque n est suffisamment grand et que l'on dispose d'un résultat de convergence en loi de type normalité asymptotique, on se sert des quantiles de la loi normale pour construire des intervalles de confiance **asymptotiques**, au sens où ils sont valables pour $n \to \infty$. Illustrons l'idée sur l'exemple du Pile ou Face.

Exemple : le Théorème Central Limite a permis d'établir la convergence en loi

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Notons $q_{1-\alpha/2}$ le quantile d'ordre $(1-\alpha/2)$ de la loi normale centrée réduite, c'est-à-dire en notant Φ^{-1} la réciproque de sa fonction de répartition (encore appelée fonction quantile),

$$q_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2) \iff \mathbb{P}\left(\mathcal{N}(0,1) \leq q_{1-\alpha/2}\right) = 1-\alpha/2 \iff \mathbb{P}\left(|\mathcal{N}(0,1)| \leq q_{1-\alpha/2}\right) = 1-\alpha.$$

Le quantile le plus connu est bien sûr $q_{0.975} = 1.96... \approx 2$, qui sert à construire des intervalles de confiance à 95%. On a donc

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \leq q_{1-\alpha/2}\frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Le paramètre inconnu θ apparaissant dans les bornes de l'intervalle, deux solutions s'offrent à nous pour pouvoir poursuivre : ou bien on lâche du lest en se souvenant que $0 < \theta(1-\theta) \leq 1/4$, pour aboutir à

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \leq \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right) \geq 1-\alpha.$$

Ou bien on fait ce qu'on appelle en anglais du *plug-in* : dans les bornes, on remplace θ par son estimateur $\hat{\theta}_n$, ce qui est justifié par le Théorème de Slutsky puisque (voir Section 1.1.4)

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1), \tag{1.4}$$

et mène à l'intervalle de confiance asymptotique

$$\left[\hat{\theta}_n - q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}, \hat{\theta}_n + q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\right]. \tag{1.5}$$

Il faut cependant garder à l'esprit que la convergence (1.4) fait intervenir une double asymptotique : ceci devient problématique lorsque θ est proche de 0, puisque la probabilité que $\hat{\theta}_n = 0$ n'est alors pas négligeable[7]. Dans ce cas, pour que l'intervalle (1.5) ait un sens, la prudence incite à prendre n au moins de l'ordre de $5/\theta$. La même remarque s'applique, mutatis mutandis, au cas où θ est proche de 1.

Quoi qu'il en soit, puisque $0 \leq \hat{\theta}_n(1-\hat{\theta}_n) \leq 1/4$, on obtient à nouveau un rayon inférieur à $q_{1-\alpha/2}/(2\sqrt{n})$. En particulier, pour $\alpha = 0.05$, il vaut donc $1/\sqrt{n}$, à comparer au $1.36/\sqrt{n}$ obtenu par Hoeffding.

Remarque : tout ce qui vient d'être dit s'applique en politique dans le cadre des sondages aléatoires simples. Ainsi, pour un échantillon de 1000 personnes prises au hasard dans la population, la précision est de l'ordre de $\pm 3\%$. Néanmoins, en pratique, les instituts de sondage utilisent des méthodes d'échantillonnage par quotas, et tout se complique pour l'estimation de la précision...

7. de l'ordre de $\exp(-n\theta)$ si $n \approx 1/\theta$, cf. par exemple l'approximation de la binomiale par la loi de Poisson.

1.3.3 Tests d'hypothèses

Le principe d'un test d'hypothèse est de répondre de façon binaire (i.e. par oui ou non) à une question sur le paramètre de l'expérience statistique en jeu. Dans le cadre du Pile ou Face, ce sera par exemple : la pièce est-elle oui ou non équilibrée ? Dans le cadre des sondages politiques, ce sera plutôt : Alice va-elle être élue plutôt que Bob ?

Ceci revient à se donner une partition de Θ en deux sous-ensembles Θ_0 et Θ_1, c'est-à-dire que

$$\Theta_0 \cup \Theta_1 = \Theta \qquad \text{et} \qquad \Theta_0 \cap \Theta_1 = \emptyset.$$

Puis, à partir d'une observation $\mathbf{X} \sim P_\theta$, à décider si le vrai paramètre θ appartient à Θ_0 ou à Θ_1. On définit ainsi :

— $H_0 : \theta \in \Theta_0$, hypothèse nulle ;
— $H_1 : \theta \in \Theta_1$, hypothèse alternative.

Exemples :

1. Pour le jeu de Pile ou Face, on veut tester $H_0 : \theta = 1/2$, c'est-à-dire $\Theta_0 = \{1/2\}$ (hypothèse simple), contre $H_1 : \theta \neq 1/2$ donc $\Theta_1 =]0, 1/2[\cup]1/2, 1[$ (hypothèse bilatère).
2. Dans le cadre des élections, notant θ la vraie proportion de votants pour Alice dans la population complète, on veut tester $H_0 : \theta \geq 1/2$, c'est-à-dire $\Theta_0 = [1/2, 1]$ (hypothèse unilatère), contre $H_1 : \theta < 1/2$, c'est-à-dire $\Theta_1 = [0, 1/2[$.

Définition 12 (Test d'hypothèse)
Un test d'hypothèse est une statistique $T(\mathbf{X})$ à valeurs dans $\{0, 1\}$ associée à la stratégie suivante : pour l'observation $\mathbf{X}$, l'hypothèse H_0 est acceptée (respectivement rejetée) si $T(\mathbf{X}) = 0$ (respectivement $T(\mathbf{X}) = 1$). Le domaine

$$\mathcal{R} = T^{-1}(\{1\}) = \{\mathbf{x} \in E,\ T(\mathbf{x}) = 1\}$$

est appelé région de rejet du test, et $\mathcal{R}^c$ la région d'acceptation.

Très souvent, la statistique de test est elle-même basée sur un estimateur $\hat{\theta} = \hat{\theta}(\mathbf{X})$ du paramètre θ et

$$T(\mathbf{X}) = \mathbb{1}_{\mathbf{X} \in \mathcal{R}} = \mathbb{1}_{\hat{\theta} \in \mathcal{R}'}.$$

Par abus de langage, on appelle encore $\mathcal{R}'$ la région de rejet associée à la statistique de test. Tous les exemples qui suivent se situent d'ailleurs dans ce cadre. A première vue, on pourrait penser au choix naturel $\mathcal{R}' = \Theta_1$ comme région de rejet de H_0, mais ce n'est pas une bonne idée, comme on le verra sur l'exemple ci-dessous.

En pratique, on dispose seulement d'une réalisation $\mathbf{x}$ de $\mathbf{X}$ et la procédure est la suivante : si $\hat{\theta} = \hat{\theta}(\mathbf{x}) \in \mathcal{R}'$, on rejette H_0, sinon on l'accepte.

Définition 13 (Risques, niveau et puissance d'un test)
On appelle :

— *risque (ou erreur) de première espèce l'application*

$$\begin{array}{rccl} \underline{\alpha} : & \Theta_0 & \to & [0,1] \\ & \theta & \mapsto & \mathbb{E}[T(\mathbf{X})] = \mathbb{P}(T(\mathbf{X}) = 1) \end{array}$$

— *taille du test le réel*

$$\alpha^\star = \sup_{\theta \in \Theta_0} \underline{\alpha}(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}(T(\mathbf{X}) = 1).$$

Etant donné $\alpha \in [0, 1]$, le test est dit de niveau α si sa taille est majorée par α.

— *risque (ou erreur) de deuxième espèce l'application*

$$\begin{array}{rccl} \underline{\beta}: & \Theta_1 & \to & [0,1] \\ & \theta & \mapsto & 1-\mathbb{E}[T(\mathbf{X})] = \mathbb{P}(T(\mathbf{X})=0) \end{array}$$

— *fonction puissance du test l'application*

$$\begin{array}{rccl} \pi: & \Theta & \to & [0,1] \\ & \theta & \mapsto & \mathbb{E}[T(\mathbf{X})] = \mathbb{P}(T(\mathbf{X})=1) \end{array}$$

Ces définitions reflètent le fait que, lors d'un test d'hypothèse, on peut se tromper de deux façons :
— ou bien en rejetant H_0 alors qu'elle est vraie, ce qui arrive avec probabilité $\underline{\alpha}(\theta)$ pour $\theta \in \Theta_0$: on parle de faux positif ;
— ou bien en conservant H_0 alors qu'elle est fausse, ce qui arrive avec probabilité $\underline{\beta}(\theta)$ pour $\theta \in \Theta_1$: on parle de faux négatif.

Clairement, la fonction puissance permet de retrouver les deux types de risques : sur Θ_0 on a $\pi(\theta) = \underline{\alpha}(\theta)$, tandis que sur Θ_1 on a $\pi(\theta) = 1 - \underline{\beta}(\theta)$. Idéalement, on aimerait que cette fonction puissance soit proche de 0 lorsque $\theta \in \Theta_0$ et proche de 1 lorsque $\theta \in \Theta_1$. Malheureusement, ceci est en général impossible puisque, dans la plupart des cas, les ensembles Θ_0 et Θ_1 ont une frontière non vide et la fonction π est continue.

Exemple : on considère $\mathbf{X} = (X_1, \ldots, X_n)$ iid selon une loi normale $\mathcal{N}(\theta, 1)$. On veut tester

$$H_0 \ : \ \theta \geq 0 \quad \text{contre} \quad H_1 \ : \ \theta < 0$$

ce qui revient, en notant $\Theta_0 = [0, +\infty[$ et $\Theta_1 =]-\infty, 0[$, à tester

$$H_0 \ : \ \theta \in \Theta_0 \quad \text{contre} \quad H_1 \ : \ \theta \in \Theta_1.$$

Une façon naturelle de procéder est de se baser sur la moyenne empirique

$$\hat{\theta}_n = \hat{\theta}(\mathbf{X}) = \frac{X_1 + \cdots + X_n}{n}$$

et de considérer la région de rejet $\mathcal{R}' =]-\infty, 0[$. Calculons la fontion puissance de ce test. Quel que soit le réel θ, la loi de l'estimateur est connue :

$$\hat{\theta}_n \sim \mathcal{N}(\theta, 1/n).$$

Par conséquent, quel que soit le réel θ,

$$\pi(\theta) = \mathbb{P}(\hat{\theta}_n < 0) = \Phi(-\theta\sqrt{n}),$$

dont la représentation se déduit de celle de Φ (voir Figure 1.7). L'erreur de première espèce et la taille du test s'en déduisent immédiatement :

$$\forall \theta \geq 0 \qquad \underline{\alpha}(\theta) = \mathbb{P}(\hat{\theta}_n < 0) = \Phi(-\theta\sqrt{n}) \implies \alpha^\star = \sup_{\theta \geq 0} \underline{\alpha}(\theta) = \sup_{\theta \geq 0} \pi(\theta) = \Phi(0) = \frac{1}{2},$$

donc on a construit un test de niveau 1/2, ce qui n'est pas glorieux... Voyons comment faire mieux.

Dissymétrisation (Neyman & Pearson) : pour sortir de cette impasse, une méthode classique est de privilégier l'une des hypothèses par rapport à l'autre, par convention H_0 par rapport à H_1, et de contrôler avant tout la probabilité de rejeter H_0 alors qu'elle est vraie, i.e. l'erreur de première espèce. Typiquement, on prendra pour H_0 :
— une hypothèse communément admise ;

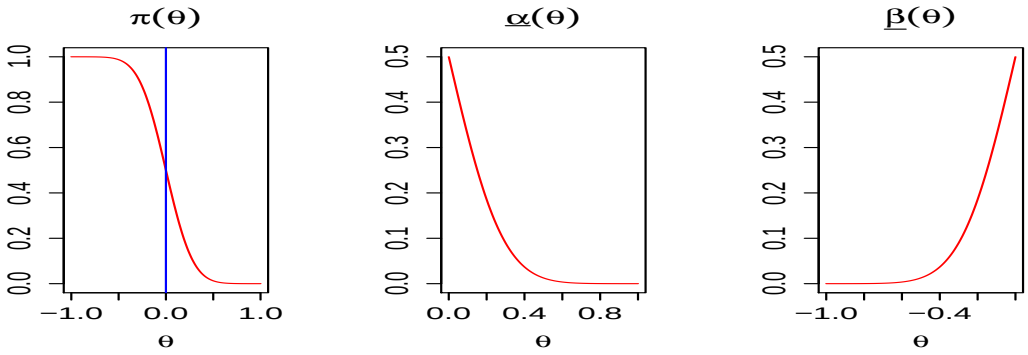

FIGURE 1.7 – Fonction puissance, risque de première espèce, risque de deuxième espèce ($n = 20$).

— une hypothèse de prudence ;
— une hypothèse facile à formuler, etc.

Le plan de vol consiste alors à se fixer un niveau α petit (inférieur à 10%) et à chercher un test de niveau α avec une fonction puissance qui tend aussi vite que possible vers 1 quand $\theta \in \Theta_1$ s'éloigne de Θ_0.

Exemple : reprenons l'exemple précédent avec la statistique de test basée sur l'estimateur $\hat{\theta}_n$. Le niveau $\alpha \in]0,1[$ étant fixé (par exemple 5%), l'idée est de se donner une marge de sécurité sur la région de rejet en considérant $\mathcal{R}'_\alpha =]-\infty, -c_\alpha[$, avec $c_\alpha > 0$. Dit autrement, pour décider que le vrai paramètre θ est négatif, la négativité de l'estimateur $\hat{\theta}_n$ ne suffit pas à nous convaincre : il faut que ce dernier soit inférieur à $-c_\alpha$, constante elle-même strictement négative. Reste à déterminer c_α. Pour ce faire, il suffit d'écrire la condition sur le niveau du test :

$$\sup_{\theta \geq 0} \mathbb{P}(\hat{\theta}_n < -c_\alpha) \leq \alpha \iff \Phi(-c_\alpha \sqrt{n}) \leq \alpha.$$

En notant $q_\alpha = \Phi^{-1}(\alpha)$ le quantile d'ordre α de la normale centrée réduite (e.g. $q_\alpha = -1.64$ si $\alpha = 5\%$), il suffit donc de prendre $c_\alpha = -q_\alpha/\sqrt{n} = q_{1-\alpha}/\sqrt{n}$. Ainsi, au niveau 5%, on rejettera H_0 si la moyenne des X_i est inférieure à $-1.64/\sqrt{n}$.

On peut alors calculer la fonction puissance du test ainsi construit. Pour tout réel θ, on a

$$\pi(\theta) = \mathbb{P}(\hat{\theta}_n < q_\alpha/\sqrt{n}) = \Phi(q_\alpha - \theta\sqrt{n}).$$

Comme attendu, cette fonction est majorée par α sur $[0, +\infty[$. Sur $]-\infty, 0[$, elle est décroissante et tend vers 1 lorsque θ s'éloigne du point frontière 0 (voir Figure 1.8).

La connaissance d'intervalles de confiance permet de construire des tests d'hypothèse. C'est ce que garantit le résultat suivant, aussi élémentaire a priori qu'efficace en pratique.

Lemme 2 (Intervalles et tests)
Soit $\alpha \in [0,1]$ fixé. Si, pour tout $\theta \in \Theta$, $(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X}))$ est un intervalle de confiance de niveau $(1-\alpha)$ pour θ, alors le test $T(\mathbf{X})$ tel que $T(\mathbf{X}) = 1$ si et seulement si $(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X})) \cap \Theta_0 = \emptyset$, est un test de niveau α.

Preuve. Il suffit de noter que, pour tout $\theta \in \Theta_0$,

$$(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X})) \cap \Theta_0 = \emptyset \implies \theta \notin (\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X})).$$

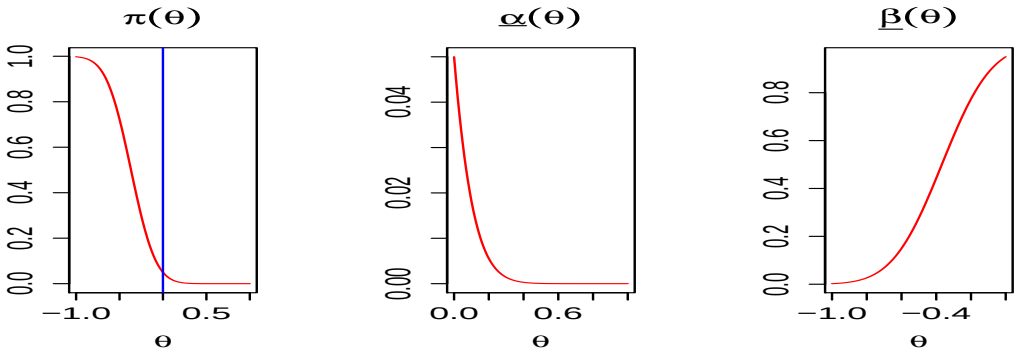

FIGURE 1.8 – Puissance, risque de première espèce, risque de deuxième espèce ($n = 20$ et $\alpha = 5\%$).

Par conséquent

$$\mathbb{P}(T(\mathbf{X}) = 1) = \mathbb{P}((\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X})) \cap \Theta_0 = \emptyset) \leq \mathbb{P}(\theta \notin (\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X}))) \leq \alpha,$$

la dernière inégalité venant de la définition même de l'intervalle de confiance. Puisque cette inégalité est valable pour tout $\theta \in \Theta_0$, elle reste vérifiée pour le supremum :

$$\alpha^\star = \sup_{\theta \in \Theta_0} \mathbb{P}(T(\mathbf{X}) = 1) \leq \alpha,$$

et le test T est bien de niveau α.

■

Exemples :

1. Dans l'exemple de l'échantillon gaussien, puisque $\hat{\theta}_n \sim \mathcal{N}(\theta, 1/n)$, on voit que pour tout réel θ, l'intervalle $]-\infty, \hat{\theta}_n + q_{1-\alpha}/\sqrt{n}]$ est un intervalle de confiance unilatère de niveau $(1-\alpha)$ pour θ. D'après ce qui vient d'être dit, on rejette H_0 si

$$\left]-\infty, \hat{\theta}_n + q_{1-\alpha}/\sqrt{n}\right] \cap [0, +\infty[= \emptyset \iff \hat{\theta}_n < -q_{1-\alpha}/\sqrt{n},$$

ce qui est précisément la condition à laquelle on avait abouti ci-dessus. Au passage, notons que $[\hat{\theta}_n - q_{1-\alpha}/\sqrt{n}, +\infty[$ est aussi un intervalle de confiance de niveau $(1-\alpha)$ pour θ, donc le test consistant à rejeter H_0 si

$$\left[\hat{\theta}_n - q_{1-\alpha}/\sqrt{n}, +\infty, \right[\cap [0, +\infty[= \emptyset$$

est aussi de niveau α. Clairement, cette condition n'est jamais réalisée : un test ne rejetant jamais H_0 ne rejette jamais H_0 à tort donc est de niveau α pour tout $\alpha \in [0, 1]$. Il n'en reste pas moins qu'il n'a aucun intérêt...

2. Pour l'exemple des élections, θ est la vraie proportion de votants pour Alice dans la population totale et on souhaite confronter les hypothèses

$$H_0 \ : \ \theta \geq \frac{1}{2} \quad \text{contre} \quad H_1 \ : \ \theta < \frac{1}{2}.$$

D'après (1.4), nous savons que

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

donc un intervalle de confiance unilatère et asymptotique de niveau $(1-\alpha)$ pour θ est

$$\left[0, \hat{\theta}_n + q_{1-\alpha}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\right].$$

Toujours par le résultat du lemme précédent, on rejette donc H_0 si

$$\hat{\theta}_n + q_{1-\alpha}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}} < \frac{1}{2}.$$

Il faut cependant noter qu'on a construit ici un test T_n de niveau **asympotique** α, au sens où

$$\sup_{\theta\in\Theta_0} \lim_{n\to+\infty} \mathbb{P}(T_n(\mathbf{X}) = 1) \leq \alpha.$$

3. Revenons à l'exemple du jeu de Pile ou Face, où nous disposons de $\mathbf{X} = (X_1, \ldots, X_n)$ iid selon la loi $\mathcal{B}(\theta)$. On veut construire un test d'hypothèse pour décider si la pièce est, oui ou non, équilibrée :

$$H_0 \ : \ \theta = \frac{1}{2} \quad \text{contre} \quad H_1 \ : \ \theta \neq \frac{1}{2}.$$

Nous avons vu en (1.5) qu'un intervalle bilatère et asymptotique de niveau $(1-\alpha)$ est donné par

$$\left[\hat{\theta}_n - q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}, \hat{\theta}_n + q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\right].$$

A partir de là, le test consistant à conserver H_0 si $1/2$ appartient à cet intervalle est asympotiquement de niveau α puisque, si la pièce est équilibrée,

$$\mathbb{P}\left(\frac{1}{2} \notin \left[\hat{\theta}_n - q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}, \hat{\theta}_n + q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\right]\right) \xrightarrow[n\to\infty]{} \alpha.$$

Prenons par exemple $n = 1000$ et $\alpha = 5\%$, donc $q_{1-\alpha/2} = 1.96... \approx 2$. En recyclant la majoration $\hat{\theta}_n(1-\hat{\theta}_n) \leq 1/4$, on rejette H_0 si $|\hat{\theta}_n - 1/2| > 0.03$.

Remarque : il ressort de ces exemples que si l'on veut construire un test unilatère (respectivement bilatère), on part d'intervalles de confiance unilatères (respectivement bilatères).

Dans ce qui précède, le choix du niveau α est fixé a priori, par exemple $\alpha = 5\%$. Puis, une réalisation $\mathbf{x}$ étant donnée, on regarde si au vu de celle-ci on rejette H_0 ou non. On peut en fait procéder de façon duale : partant de $\mathbf{x}$ et d'une famille $\mathcal{R}_\alpha$ (ou $\mathcal{R}'_\alpha$) de régions de rejet, on peut se demander à quel point la réalisation est en (dés)accord avec H_0.

Exemple : on revient sur l'exemple de l'échantillon gaussien. Supposons que l'on observe $\mathbf{x} = (x_1, \ldots, x_{100})$ de moyenne empirique $\hat{\theta}_n(\mathbf{x}) = -0.3$. A partir de là, conserve-t-on H_0 au niveau 10% ? 5% ? 1% ? La réponse est donnée par la procédure de test : celle-ci spécifie en effet que l'on rejette H_0 au niveau α si et seulement si

$$\hat{\theta}_n(\mathbf{x}) < \Phi^{-1}(\alpha)/\sqrt{n} \iff \alpha > \Phi(\sqrt{n}\hat{\theta}_n(\mathbf{x})) = \Phi(-3) \approx 10^{-3}.$$

En particulier, on rejette H_0 au niveau de risque 10%, 5%, 1%, et en fait à tout niveau supérieur à 1‰. La notion de p-value permet de formaliser cette idée.

Revenons donc au cas général. Notant $\mathcal{R}_\alpha$ la région de rejet de niveau α pour la statistique de test $T(\mathbf{X})$, on rejette H_0 si

$$T(\mathbf{X}) = 1 \iff \mathbf{X} \in \mathcal{R}_\alpha.$$

Si cette statistique de test est basée sur un estimateur $\hat{\theta} = \hat{\theta}(\mathbf{X})$, ceci s'exprime encore

$$T(\mathbf{X}) = 1 \iff \hat{\theta} \in \mathcal{R}'_\alpha.$$

Ce qui se passe dans quasiment tous les cas, et ce que nous supposerons dans la suite, c'est que les régions de rejet sont emboîtées, c'est-à-dire que

$$0 \leq \alpha_1 \leq \alpha_2 \leq 1 \iff \mathcal{R}_{\alpha_1} \subseteq \mathcal{R}_{\alpha_2} \iff \mathcal{R}'_{\alpha_1} \subseteq \mathcal{R}'_{\alpha_2}.$$

En pratique, on dispose d'une réalisation $\mathbf{x}$ et on veut décider si, au vu de cette réalisation, on accepte H_0 ou si on la rejette, et ce en précisant le niveau de risque.

Définition 14 (Niveau de significativité, probabilité critique, p-value)
Pour une réalisation $\mathbf{x}$*, on appelle niveau de significativité (ou probabilité critique, ou p-value) du test associé aux régions de rejet* $\mathcal{R}_\alpha$ *la quantité*

$$\alpha_0(\mathbf{x}) = \inf\{\alpha \in [0,1], \mathbf{x} \in \mathcal{R}_\alpha\} = \inf\{\alpha \in [0,1],\ H_0 \text{ est rejetée au niveau } \alpha\}.$$

Take-home message : c'est cette valeur $\alpha_0(\mathbf{x})$ qui est usuellement donnée par les logiciels de statistiques en sortie d'un test d'hypothèse. Comme son nom en français l'indique, cette p-value reflète à quel point il est significatif de rejeter H_0. Si $\alpha_0(\mathbf{x})$ est très proche de 0 (disons inférieur à $1/100$), on rejette H_0 sans scrupules. Si au contraire $\alpha_0(\mathbf{x})$ est grand (disons supérieur à $1/10$), il est raisonnable de conserver H_0. Pour des valeurs intermédiaires de $\alpha_0(\mathbf{x})$, plus rien n'est clair, la décision dépend du problème en question.

Revenons à l'exemple de l'échantillon gaussien où a été observée, pour $n = 100$, une moyenne empirique $\hat{\theta}_n(\mathbf{x}) = -0.3$, correspondant à une p-value d'environ 10^{-3}. Une autre façon de retrouver ce résultat est de se dire que si H_0 était vraie, c'est-à-dire $\theta \geq 0$, le scénario le plus vraisemblable pour observer une valeur négative de $\hat{\theta}_n(\mathbf{x})$ est que $\theta = 0$. Or si $\theta = 0$, l'estimateur $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$ suit une loi normale $\mathcal{N}(0, 1/n)$ et la probabilité qu'une telle variable soit inférieure ou égale à -0.3 est, avec $n = 100$,

$$\mathbb{P}(\mathcal{N}(0, 1/100) \leq -0.3) = \mathbb{P}(\mathcal{N}(0,1) \leq -3) = \Phi(-3) \approx 10^{-3}.$$

Ceci permet d'interpréter la p-value comme une probabilité (et au passage de comprendre le "p" de p-value) : elle correspond à la probabilité qu'on aurait d'observer une valeur au moins aussi négative de $\hat{\theta}_n$ si H_0 était vraie. Le "au moins aussi négative" vient du test fait ici et de H_0, qui suppose $\theta \geq 0$. Pour un autre test, il faudra adapter le vocabulaire, comme l'illustre l'exemple suivant.

Exemple : nous revenons à l'exemple du Pile ou Face, où l'on veut tester

$$H_0 \ :\ \theta = \frac{1}{2} \quad \text{contre} \quad H_1 \ :\ \theta \neq \frac{1}{2}.$$

On observe $\mathbf{x} = (x_1, \ldots, x_n)$: quelle est la p-value associée ? Sous H_0, on a $\theta = 1/2$ donc, en supposant n suffisamment grand pour pouvoir appliquer le TCL, on a avec un abus de notation

$$2\sqrt{n}\left(\hat{\theta}_n - \frac{1}{2}\right) \overset{\mathcal{L}}{\approx} \mathcal{N}(0,1).$$

Dès lors, il est clair que le test consistant à rejeter H_0 si

$$2\sqrt{n}\left|\hat{\theta}_n - \frac{1}{2}\right| > q_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$$

est de niveau α. Notant $\hat{\theta}_n(\mathbf{x})$ la fréquence empirique observée, la p-value est donc par définition

$$\alpha_0(\mathbf{x}) = \inf\{\alpha \in [0,1], \mathbf{x} \in \mathcal{R}_\alpha\} = \inf\{\alpha \in [0,1],\ 2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2| > \Phi^{-1}(1-\alpha/2)\}.$$

Or la croissance de Φ permet d'écrire

$$2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2| > \Phi^{-1}(1-\alpha/2) \iff \alpha > 2(1 - \Phi(2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2|))$$

d'où

$$\alpha_0(\mathbf{x}) = 2(1 - \Phi(2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2|).$$

Puisque, de façon générale, on a pour tout $c \geq 0$,

$$\mathbb{P}(|\mathcal{N}(0,1)| \geq c) = 2(1 - \Phi(c)),$$

on peut aussi écrire

$$\alpha_0(\mathbf{x}) = \mathbb{P}(|\mathcal{N}(0,1)| \geq 2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2|).$$

Or, sous H_0,

$$\theta = 1/2 \implies \hat{\theta}_n(\mathbf{X}) \sim \mathcal{N}(1/2, 1/(4n)) \implies 2\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - 1/2) \sim \mathcal{N}(0,1)$$

et l'équation précédente est encore équivalente à

$$\alpha_0(\mathbf{x}) = \mathbb{P}(2\sqrt{n}|\hat{\theta}_n(\mathbf{X}) - 1/2| \geq 2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2|) = \mathbb{P}(|\hat{\theta}_n(\mathbf{X}) - 1/2| \geq |\hat{\theta}_n(\mathbf{x}) - 1/2|).$$

La p-value correspond donc à la probabilité d'observer un écart à $1/2$ au moins aussi grand que $|\hat{\theta}_n(\mathbf{x}) - 1/2|$ si la pièce est équilibrée.

Généralisation : pour voir la p-value comme une probabilité, il faut considérer que le test $T(\mathbf{X})$ est obtenu par le seuillage d'une statistique $S(\mathbf{X})$, c'est-à-dire que l'on rejette H_0 au niveau α si et seulement si $S(\mathbf{X}) > c_\alpha$. Les exemples que nous avons déjà rencontrés, et en fait tous ceux que nous croiserons, ne procèdent pas autrement :

- Echantillon gaussien : $S(\mathbf{x}) = -\sqrt{n}\hat{\theta}_n(\mathbf{x})$ et $c_\alpha = q_{1-\alpha}$.
- Alice et Bob :
$$S(\mathbf{x}) = -\sqrt{n}\frac{\hat{\theta}_n(\mathbf{x}) - \frac{1}{2}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \qquad \text{et} \qquad c_\alpha = q_{1-\alpha}.$$
- Pile ou Face : $S(\mathbf{x}) = 2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2|$ et $c_\alpha = q_{1-\alpha/2}$.

Une réalisation $\mathbf{x}$ étant donnée, on peut alors montrer que, sous des hypothèses très générales, la p-value se reformule comme suit :

$$\alpha_0(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}(S(\mathbf{X}) \geq S(\mathbf{x})),$$

où, pour chaque valeur de $\theta \in \Theta_0$, $\mathbf{X}$ (aléatoire !) a pour loi P_θ. On résume souvent ceci par la phrase : "La p-value est la probabilité, sous H_0, d'obtenir une statistique de test au moins aussi extrême que celle observée."

Chapitre 2

Le modèle linéaire gaussien

Introduction

Le principe de la régression est de modéliser une variable y, dite variable à expliquer, comme une fonction d'un certain nombre de variables $\mathbf{x} = [x_1, \dots, x_p]'$, dites variables explicatives :

$$y = g(\mathbf{x}) = g(x_1, \dots, x_p).$$

On dispose d'un échantillon de taille n de $(p+1)$-uplets $(\mathbf{x}, y)$ et le but est de retrouver la fonction g. Le modèle le plus simple est celui où g est linéaire, c'est-à-dire qu'il existe un vecteur de coefficients $\beta = [\beta_1, \dots, \beta_p]'$ tel que

$$y = \mathbf{x}'\beta = \beta_1 x_1 + \cdots + \beta_p x_p.$$

En pratique, ceci ne marche pas, ou bien parce que ce modèle est approché (la liaison n'est pas réellement linéaire) ou bien en raison des erreurs de mesure. L'idée est alors de voir y comme la réalisation d'une variable aléatoire Y tenant compte de cette inadéquation. Concrètement, ceci revient à réécrire le modèle sous la forme

$$Y = \mathbf{x}'\beta + \varepsilon = \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

où la variable aléatoire ε est supposée centrée et de variance inconnue σ^2. On parle alors de modèle de régression linéaire. Partant de notre échantillon, l'objectif est ainsi d'estimer le paramètre β ainsi que la variance σ^2 de l'erreur ε. On a donc affaire à un problème d'inférence statistique, paramétrique au sens de la Définition 6. Les exemples d'applications de la régression linéaire foisonnent, on se contente ici d'en mentionner quelques-uns :

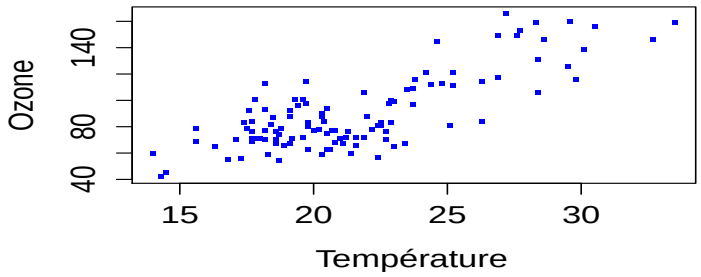

FIGURE 2.1 – Maximum journalier de l'ozone en fonction de la température à midi.

1. **Concentration de l'ozone** : dans ce domaine, on cherche à expliquer le maximum journalier de la concentration en ozone, notée O_3 (en $\mu g/m^3$), en fonction de la température à midi T. Le nuage de points de la Figure 2.1 correspond à 112 données relevées durant l'été 2001 à Rennes. On propose le modèle :

$$O_3 = \beta_1 + \beta_2 T + \varepsilon.$$

Lorsqu'il n'y a, comme ici, qu'une "vraie" variable explicative (la température), on parle de régression linéaire simple. On peut affiner ce modèle en tenant compte de la nébulosité [1] N à midi et de la projection V du vecteur vitesse du vent sur l'axe Est-Ouest, ce qui donne

$$O_3 = \beta_1 + \beta_2 T + \beta_3 V + \beta_4 N + \varepsilon,$$

et on parle alors de régression linéaire multiple.

2. **Hauteur d'un eucalyptus** : la Figure 2.2 correspond à environ 1400 couples (x_i, y_i) où x_i correspond à la circonférence du tronc à 1 mètre du sol (en centimètres) et y_i à la hauteur de l'arbre (en mètres). Au vu de ce nuage de points, on peut proposer le modèle

$$Y = \beta_1 + \beta_2 x + \beta_3 \sqrt{x} + \varepsilon.$$

On voit sur cet exemple que le modèle de régression linéaire est linéaire en les paramètres inconnus β_j, non en la variable x !

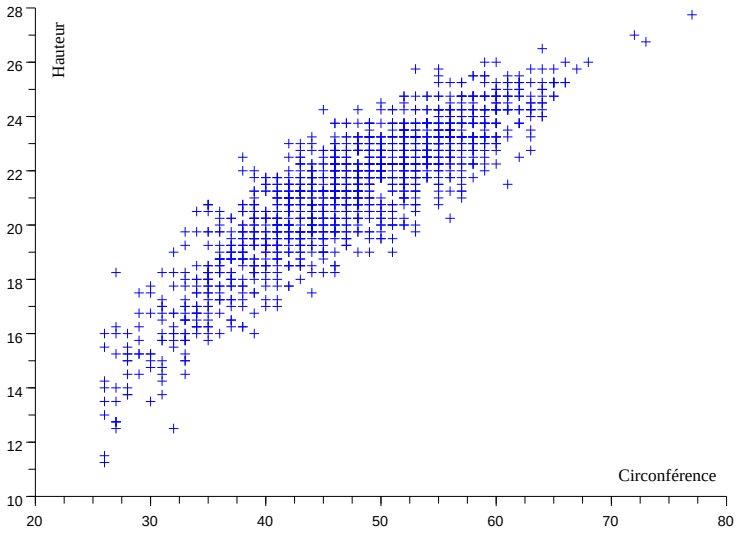

FIGURE 2.2 – Nuage de points pour les eucalyptus.

3. **Modèle de Cobb-Douglas** : énoncé en 1928 dans l'article *A Theory of Production*, le principe est de décrire, sur l'ensemble des Etats-Unis, la production P en fonction du capital K (valeur des usines, etc.) et du travail T (nombre de travailleurs). Les auteurs proposèrent le modèle suivant

$$P = \alpha_1 K^{\alpha_2} T^{\alpha_3}.$$

En passant au logarithme, en notant $(\beta_1, \beta_2, \beta_3) = (\log \alpha_1, \alpha_2, \alpha_3)$ et en tenant compte de l'erreur du modèle, on aboutit donc à

$$\log P = \beta_1 + \beta_2 \log K + \beta_3 \log T + \varepsilon.$$

1. celle-ci prend des valeurs entières de 0 à 8, pour un ciel allant de très dégagé à très couvert.

A partir de données sur 24 années consécutives, de 1899 à 1922, ils estimèrent $\alpha_2 = 1/4$ et $\alpha_3 = 3/4$. Ici, partant d'un modèle de régression non-linéaire en α_2 et α_3, on a pu le linéariser grâce à une simple transformation logarithmique. Ce n'est bien sûr pas toujours le cas...

2.1 Régression linéaire multiple

2.1.1 Modélisation

Nous supposons que les données collectées suivent le modèle suivant :

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2.1}$$

où :
- les Y_i sont des variables aléatoires dont on observe les réalisations y_i ;
- les x_{ij} sont connus, non aléatoires, la variable x_{i1} valant souvent 1 pour tout i ;
- les paramètres β_j du modèle sont inconnus, mais non aléatoires ;
- les ε_i sont des variables aléatoires inconnues, i.e. non observées contrairement aux Y_i.

Remarque : comme la constante appartient généralement au modèle, beaucoup d'auteurs l'écrivent plutôt sous la forme

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

de sorte que p correspond toujours au nombre de "vraies" variables explicatives. Avec notre convention d'écriture (2.1), si x_{i1} vaut 1 pour tout i, p est le nombre de paramètres à estimer, tandis que le nombre de variables explicatives est, à proprement parler, $(p-1)$.

En adoptant l'écriture matricielle de (2.1), nous obtenons la définition suivante :

Définition 15 (Modèle de régression linéaire multiple)
Un modèle de régression linéaire est défini par une équation de la forme :

$$Y = X\beta + \boldsymbol{\varepsilon}$$

où :
- *Y est un vecteur aléatoire de dimension n,*
- *X est une matrice de taille $n \times p$ connue, appelée matrice du plan d'expérience,*
- *β est le vecteur de dimension p des paramètres inconnus du modèle,*
- *$\boldsymbol{\varepsilon}$, de dimension n, est le vecteur aléatoire et inconnu des erreurs.*

Les hypothèses concernant le modèle sont

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \text{les } \varepsilon_i \text{ sont i.i.d. avec } \mathbb{E}[\boldsymbol{\varepsilon}] = 0 \text{ et } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{cases}$$

L'hypothèse $(\mathcal{H}_1)$ assure que le modèle est identifiable, nous y reviendrons en Section 2.2 pour l'étude du modèle gaussien. Pour l'instant, contentons-nous de noter qu'elle implique $p \leq n$ et qu'elle est équivalente à supposer la matrice $X'X$ inversible. Supposons en effet qu'il existe un vecteur β de $\mathbb{R}^p$ tel que $(X'X)\beta = 0$. Ceci impliquerait que $\|X\beta\|^2 = \beta'(X'X)\beta = 0$, donc $X\beta = 0$, d'où $\beta = 0$ puisque $\mathrm{rg}(X) = p$. Autrement dit, la matrice symétrique $X'X$ est définie positive.

En $(\mathcal{H}_2)$, supposer les erreurs centrées est naturel : si tel n'était pas le cas, leur moyenne m passerait dans la partie déterministe du modèle, quitte éventuellement à ajouter un paramètre $\beta_0 = m$ si la

constante n'est pas déjà présente dans le modèle. Par ailleurs, dans toute cette section 2.1, nous pourrions en fait nous contenter de supposer que les erreurs ε_i sont décorrélées, centrées et de même variance σ^2 (on parle alors d'homoscédasticité).

Notation. On notera $X = [X_1|\dots|X_p]$, où X_j est le vecteur colonne de taille n correspondant à la j-ème variable. La i-ème ligne de la matrice X sera quant à elle notée $\mathbf{x}_i' = [x_{i1}, \dots, x_{ip}]$ et elle correspond au i-ème "individu" de l'échantillon. La matrice X du plan d'expérience est aussi appelée matrice "individus $\times$ variables". Par conséquent, l'équation (2.1) s'écrit encore

$$Y_i = \mathbf{x}_i'\beta + \varepsilon_i \qquad \forall i \in \{1, \dots, n\}.$$

2.1.2 Estimateurs des Moindres Carrés

Notre but est tout d'abord d'estimer β. Mathématiquement, l'estimateur le plus simple à calculer et à étudier est celui dit des Moindres Carrés. Lorsque les erreurs ε_i sont gaussiennes, il correspond d'ailleurs à celui du maximum de vraisemblance. Nous y reviendrons.

Définition 16 (Estimateur des Moindres Carrés)
L'estimateur des moindres carrés $\hat{\beta}$ est défini comme suit :

$$\hat{\beta} = \underset{\alpha\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \alpha_j x_{ij} \right)^2 = \underset{\alpha\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i'\alpha \right)^2 = \underset{\alpha\in\mathbb{R}^p}{\operatorname{argmin}} \left\| Y - \sum_{j=1}^{p} \alpha_j X_j \right\|^2$$
$$= \underset{\alpha\in\mathbb{R}^p}{\operatorname{argmin}} \|Y - X\alpha\|^2. \tag{2.2}$$

Pour déterminer $\hat{\beta}$, il suffit de raisonner géométriquement. La matrice $X = [X_1|\dots|X_p]$ du plan d'expérience est formée de p vecteurs colonnes dans $\mathbb{R}^n$ (la première étant généralement constituée de 1). Le sous-espace de $\mathbb{R}^n$ engendré par ces p vecteurs colonnes est appelé espace image, ou espace des solutions, et noté

$$\mathcal{M}_X = \operatorname{Im}(X) = \operatorname{Vect}(X_1, \dots, X_p).$$

Il est de dimension p par l'hypothèse $(\mathcal{H}_1)$ et tout vecteur de cet espace est de la forme $X\alpha$, où α est un vecteur de $\mathbb{R}^p$:

$$X\alpha = \alpha_1 X_1 + \dots + \alpha_p X_p.$$

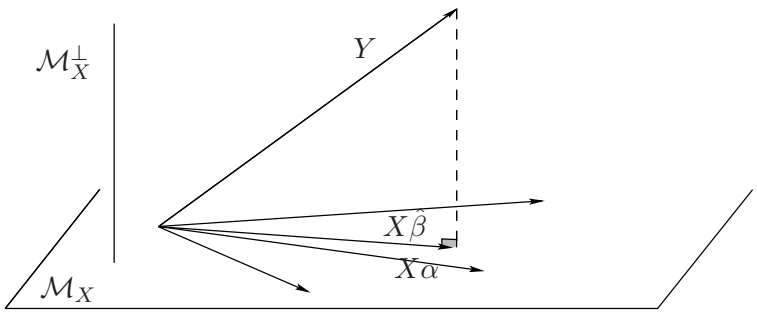

FIGURE 2.3 – Représentation de $X\hat{\beta}$ dans l'espace des variables.

Selon le modèle de la Définition 15, le vecteur Y est la somme d'un élément $X\beta$ de $\mathcal{M}_X$ et d'une erreur $\boldsymbol{\varepsilon}$, laquelle n'a aucune raison d'appartenir à $\mathcal{M}_X$. Minimiser $\|Y - X\alpha\|^2$ revient à chercher

l'élément de $\mathcal{M}_X$ qui soit le plus proche de Y au sens de la norme euclidienne. Cet élément, unique, est par définition le projeté orthogonal de Y sur $\mathcal{M}_X$. Il sera noté $\hat{Y} = P_X Y$, où P_X est la matrice de projection orthogonale sur $\mathcal{M}_X$. Il peut aussi s'écrire sous la forme $\hat{Y} = X\hat{\beta}$, où $\hat{\beta}$ est l'estimateur des moindres carrés de β. L'espace orthogonal à $\mathcal{M}_X$, noté $\mathcal{M}_X^\perp$, est souvent appelé espace des résidus. En tant que supplémentaire orthogonal, il est de dimension

$$\dim(\mathcal{M}_X^\perp) = \dim(\mathbb{R}^n) - \dim(\mathcal{M}_X) = n - p.$$

Les expressions de $\hat{\beta}$ et P_X données maintenant sont sans aucun doute les plus importantes de tout ce chapitre, puisqu'on peut quasiment tout retrouver à partir de celles-ci.

Proposition 7 (Expression de $\hat{\beta}$)
L'estimateur $\hat{\beta}$ des moindres carrés a pour expression :

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

et la matrice P_X de projection orthogonale sur $\mathcal{M}_X$ s'écrit :

$$P_X = X(X'X)^{-1}X'.$$

Preuve. On peut montrer ce résultat de plusieurs façons.

1. Par différentiation : on cherche $\alpha \in \mathbb{R}^p$ qui minimise la fonction

   $$S(\alpha) = \|Y - X\alpha\|^2 = \alpha'(X'X)\alpha - 2Y'X\alpha + \|Y\|^2.$$

   Or S est de type quadratique en α, avec $X'X$ symétrique définie positive, donc le problème admet une unique solution $\hat{\beta}$: c'est le point où le gradient de S est nul. Géométriquement, en dimension 2, c'est le sommet du paraboloïde défini par S. Ceci s'écrit :

   $$\nabla S(\hat{\beta}) = 2\hat{\beta}'X'X - 2Y'X = 0 \iff (X'X)\hat{\beta} = X'Y.$$

   La matrice $X'X$ étant inversible par $(\mathcal{H}_1)$, ceci donne $\hat{\beta} = (X'X)^{-1}X'Y$. Puisque par définition $\hat{Y} = P_X Y = X\hat{\beta} = X(X'X)^{-1}X'Y$ et que cette relation est valable pour tout $Y \in \mathbb{R}^n$, on en déduit que $P_X = X(X'X)^{-1}X'$.

2. Par projection : une autre méthode consiste à dire que le projeté orthogonal $\hat{Y} = X\hat{\beta}$ est défini comme l'unique vecteur tel que $(Y - \hat{Y})$ soit orthogonal à $\mathcal{M}_X$. Puisque $\mathcal{M}_X$ est engendré par les vecteurs $X_1, \ldots, X_p$, ceci revient à dire que $(Y - \hat{Y})$ est orthogonal à chacun des X_i :

   $$\begin{cases} \langle X_1, Y - X\hat{\beta}\rangle = X_1'(Y - X\hat{\beta}) = 0 \\ \qquad\qquad\qquad\vdots \\ \langle X_p, Y - X\hat{\beta}\rangle = X_p'(Y - X\hat{\beta}) = 0 \end{cases}$$

   Ces p équations se regroupent en une seule : $X'(Y - X\hat{\beta}) = 0$, d'où l'on déduit bien l'expression de $\hat{\beta}$, puis celle de P_X.

■

Exemple. Revenons à l'exemple des eucalyptus. La courbe des moindres carrés, de la forme $y = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3\sqrt{x}$, est représentée Figure 2.4.

Dorénavant nous noterons $P_X = X(X'X)^{-1}X'$ la matrice de projection orthogonale sur $\mathcal{M}_X$ et $P_{X^\perp} = (I_n - P_X)$ la matrice de projection orthogonale sur $\mathcal{M}_X^\perp$. La décomposition

$$Y = \hat{Y} + (Y - \hat{Y}) = P_X Y + (I_n - P_X)Y = P_X Y + P_{X^\perp} Y$$

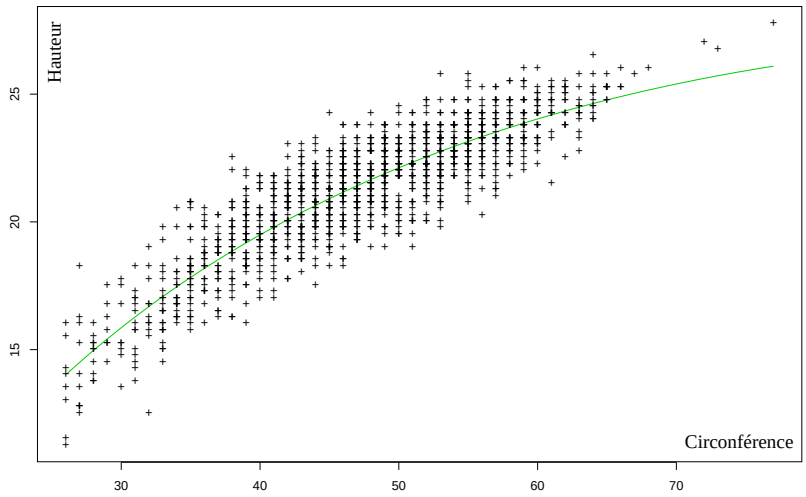

FIGURE 2.4 – Nuage de points et courbe des moindres carrés pour les eucalyptus.

n'est donc rien de plus qu'une décomposition orthogonale de Y sur $\mathcal{M}_X$ et $\mathcal{M}_X^\perp$.

Achtung ! La décomposition

$$\hat{Y} = X\hat{\beta} = \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

signifie que les $\hat{\beta}_i$ sont les coordonnées de $\hat{Y}$ dans la base $(X_1, \ldots, X_p)$ de $\mathcal{M}_X$. Il ne faudrait pas croire pour autant que les $\hat{\beta}_i$ sont les coordonnées des projections de Y sur les X_i : ceci n'est vrai que si la base $(X_1, \ldots, X_p)$ est orthogonale, ce qui n'est pas le cas en général.

Rappels sur les projecteurs : soit P une matrice carrée de taille n. On dit que P est une matrice de projection si $P^2 = P$. Ce nom est dû au fait que pour tout vecteur x de $\mathbb{R}^n$, Px est la projection de x sur $\mathrm{Im}(P)$ parallèlement à $\mathrm{Ker}(P)$. Si en plus de vérifier $P^2 = P$, la matrice P est symétrique (i.e. $P' = P$), alors Px est la projection **orthogonale** de x sur $\mathrm{Im}(P)$ parallèlement à $\mathrm{Ker}(P)$, c'est-à-dire qu'on a la décomposition

$$x = Px + (x - Px) \qquad \text{avec} \qquad Px \perp x - Px.$$

C'est ce cas de figure qui nous concernera dans ce cours. Toute matrice symétrique réelle étant diagonalisable en base orthonormée, il existe une matrice orthogonale Q (i.e. $QQ' = I_n$, ce qui signifie que les colonnes de Q forment une base orthonormée de $\mathbb{R}^n$) et une matrice diagonale Δ telles que $P = Q\Delta Q'$. On voit alors facilement que la diagonale de Δ est composée de p "1" et de $(n-p)$ "0", où p est la dimension de $\mathrm{Im}(P)$, espace sur lequel on projette. En particulier la trace de P, qui est égale à celle de Δ, vaut tout simplement p.

Revenons à nos moutons : on a vu que $P_X = X(X'X)^{-1}X'$. On vérifie bien que $P_X^2 = P_X$ et que P_X est symétrique. Ce qui précède assure également que $\mathrm{Tr}(P_X) = p$ et $\mathrm{Tr}(P_{X^\perp}) = n - p$. Cette dernière remarque nous sera utile pour construire un estimateur sans biais de σ^2. D'autre part, la matrice P_X est souvent notée H (comme *Hat*) dans la littérature anglo-saxonne, car elle met un chapeau sur le vecteur Y : $P_X Y = HY = \hat{Y}$.

Nous allons maintenant nous intéresser au biais et à la matrice de covariance de l'estimateur $\hat{\beta}$ des moindres carrés. On rappelle que la matrice de covariance du vecteur aléatoire $\hat{\beta}$, ou matrice de variance-covariance, ou matrice de dispersion, est par définition :

$$\mathrm{Cov}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])'] = \mathbb{E}[\hat{\beta}\hat{\beta}'] - \mathbb{E}[\hat{\beta}]\mathbb{E}[\hat{\beta}]'.$$

Puisque β est de dimension p, elle est de dimension $p \times p$. De plus, pour pour toute matrice A de taille $m \times p$ et tout vecteur b de dimension m déterministes, on a

$$\mathbb{E}[A\hat{\beta} + b] = A\mathbb{E}[\hat{\beta}] + b \qquad \text{et} \qquad \text{Cov}(A\hat{\beta} + b) = A\text{Cov}(\hat{\beta})A'.$$

Ces propriétés élémentaires seront très souvent appliquées dans la suite.

Proposition 8 (Biais et matrice de covariance)
L'estimateur $\hat{\beta}$ des moindres carrés est sans biais, i.e. $\mathbb{E}[\hat{\beta}] = \beta$, et sa matrice de covariance est

$$\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

Preuve. Pour le biais il suffit d'écrire :

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(X'X)^{-1}X'Y] = (X'X)^{-1}X'\mathbb{E}[Y] = (X'X)^{-1}X'\mathbb{E}[X\beta + \boldsymbol{\varepsilon}] = (X'X)^{-1}X'(X\beta + \mathbb{E}[\boldsymbol{\varepsilon}]),$$

et puisque $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$, il vient

$$\mathbb{E}[\hat{\beta}] = (X'X)^{-1}X'X\beta = \beta.$$

Pour la variance, on procède de même :

$$\text{Cov}(\hat{\beta}) = \text{Cov}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\text{Cov}(Y)X(X'X)^{-1},$$

or $\text{Cov}(Y) = \text{Cov}(X\beta + \boldsymbol{\varepsilon}) = \text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, donc

$$\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

■

Les résidus sont définis par

$$\hat{\boldsymbol{\varepsilon}} = [\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n]' = Y - X\hat{\beta} = Y - \hat{Y} = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\boldsymbol{\varepsilon},$$

car $Y = X\beta + \boldsymbol{\varepsilon}$ et $X\beta \in \mathcal{M}_X$. Si $\hat{\beta}$ estime bien β, alors les résidus $\hat{\boldsymbol{\varepsilon}} = Y - \hat{Y} = Y - X\hat{\beta}$ estiment bien les erreurs, donc un estimateur "naturel" de la variance résiduelle σ^2 est donné par :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\|\hat{\boldsymbol{\varepsilon}}\|^2 = \frac{SCR}{n},$$

où $SCR = \|\hat{\boldsymbol{\varepsilon}}\|^2$ est appelée somme des carrés résiduelle. En fait, comme on va le voir, cet estimateur est biaisé. Ce biais est néanmoins facilement corrigeable, comme le montre le résultat suivant.

Proposition 9 (Estimateur de la variance résiduelle)
La statistique

$$\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p} = \frac{SCR}{n-p}$$

est un estimateur sans biais de σ^2.

Preuve. Nous calculons tout bonnement la moyenne de la somme des carrés résiduelle, en tenant compte du fait que $P_{X^\perp}$ est un projecteur orthogonal :

$$\mathbb{E}[\|\hat{\boldsymbol{\varepsilon}}\|^2] = \mathbb{E}[\|P_{X^\perp}\boldsymbol{\varepsilon}\|^2] = \mathbb{E}[\boldsymbol{\varepsilon}'P'_{X^\perp}P_{X^\perp}\boldsymbol{\varepsilon}] = \mathbb{E}[\boldsymbol{\varepsilon}'P_{X^\perp}\boldsymbol{\varepsilon}] = \mathbb{E}\left[\sum_{1\leq i,j\leq n} P_{X^\perp}(i,j)\varepsilon_i\varepsilon_j\right],$$

Par linéarité de l'espérance et indépendance des erreurs, il vient :

$$\mathbb{E}[\|\hat{\boldsymbol{\varepsilon}}\|^2] = \sum_{1\leq i,j\leq n} P_{X^\perp}(i,j)\mathbb{E}\left[\varepsilon_i\varepsilon_j\right] = \sigma^2\sum_{1\leq i\leq n} P_{X^\perp}(i,i) = \sigma^2\text{Tr}(P_{X^\perp}).$$

Et comme $P_{X^\perp}$ projette sur un sous-espace de dimension $(n-p)$, on a bien :

$$\mathbb{E}[\|\hat{\boldsymbol{\varepsilon}}\|^2] = (n-p)\sigma^2.$$

■

On déduit de cet estimateur de $\hat{\sigma}^2$ de la variance résiduelle σ^2 un estimateur de la covariance de β, valant comme on l'a vu $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$:

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p}(X'X)^{-1} = \frac{SCR}{n-p}(X'X)^{-1}.$$

En particulier, un estimateur de l'écart-type de l'estimateur $\hat{\beta}_j$ du j-ème coefficient de la régression est tout simplement

$$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}.$$

Afin d'alléger les notations, on écrira parfois $\hat{\sigma}_j$ pour $\hat{\sigma}_{\hat{\beta}_j}$.

Cautious ! L'écriture $\left[(X'X)^{-1}\right]_{jj}$ signifie "le j-ème terme diagonal de la matrice $(X'X)^{-1}$", et non "l'inverse du j-ème terme diagonal de la matrice $(X'X)$". Afin d'alléger les écritures, nous écrirons souvent $(X'X)^{-1}_{jj}$ au lieu de $\left[(X'X)^{-1}\right]_{jj}$.

2.2 Le modèle gaussien

Rappelons le contexte de la section précédente. Nous avons supposé un modèle de la forme :

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

que nous avons réécrit en termes matriciels :

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

où les dimensions sont indiquées en indices. Les hypothèses concernant le modèle étaient :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \text{les } \varepsilon_i \text{ sont i.i.d. avec } \mathbb{E}[\boldsymbol{\varepsilon}] = 0 \text{ et } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{cases}$$

Nous allons désormais faire une hypothèse plus forte, à savoir celle de gaussianité des résidus. Nous supposerons donc jusqu'à la fin de ce chapitre :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

L'intérêt de supposer les résidus gaussiens est de pouvoir en déduire les lois de nos estimateurs, donc de construire des régions de confiance et des tests d'hypothèses. Par ailleurs, même si l'on peut bien entendu trouver des exemples ne rentrant pas dans ce cadre, modéliser les erreurs par une loi gaussienne n'est généralement pas farfelu au vu du Théorème Central Limite.

Remarques :

1. Si l'on reprend la Définition 5 d'une expérience statistique, l'objet aléatoire est ici le vecteur Y de $\mathbb{R}^n$, de loi normale $\mathcal{N}(X\beta, \sigma^2 I_n)$. En accord avec la Définition 7, le modèle statistique

   $$(P_\theta)_{\theta\in\Theta} = (\mathcal{N}(X\beta, \sigma^2 I_n))_{\beta\in\mathbb{R}^p, \sigma^2>0}$$

   n'est cependant identifiable que si l'application $(\beta, \sigma^2) \mapsto \mathcal{N}(X\beta, \sigma^2 I_n)$ est injective, or ceci n'est vrai que si X est injective, donc de rang p, d'où l'hypothèse $(\mathcal{H}_1)$.
2. Contrairement à tous les exemples du Chapitre 1, nous ne sommes plus dans un modèle d'échantillonnage puisque toutes les variables Y_i n'ont pas la même loi : $Y_i \sim \mathcal{N}(\mathbf{x}_i'\beta, \sigma^2)$, c'est-à-dire qu'elles ont même variance mais pas même moyenne.

2.2.1 Quelques rappels

Commençons par quelques rappels sur les vecteurs gaussiens. Un vecteur aléatoire Y de $\mathbb{R}^n$ est dit gaussien si toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne. Ce vecteur admet alors une espérance $\mu = \mathbb{E}[Y]$ et une matrice de variance-covariance $\Sigma_Y = \mathbb{E}[(Y-\mu)(Y-\mu)']$ qui caractérisent complètement sa loi. On note dans ce cas $Y \sim \mathcal{N}(\mu, \Sigma_Y)$.

Plusieurs aspects rendent les vecteurs gaussiens particulièrement sympathiques. Le premier concerne leur stabilité par transformation affine : Si A et b sont respectivement une matrice et un vecteur déterministes de tailles adéquates, alors

$$Y \sim \mathcal{N}(\mu, \Sigma_Y) \implies AY + b \sim \mathcal{N}(A\mu + b, A\Sigma_Y A').$$

Le second point agréable est la facilité avec laquelle on peut vérifier l'indépendance : en effet, les composantes d'un vecteur gaussien $Y = [Y_1, \cdots, Y_n]'$ sont indépendantes si et seulement si Σ_Y est diagonale. Dit crûment, dans le cadre vecteur gaussien, indépendance équivaut à décorrélation.

Disons enfin un mot de la densité. Soit $Y \sim \mathcal{N}(\mu, \Sigma_Y)$ un vecteur gaussien. Il admet une densité f sur $\mathbb{R}^n$ si et seulement si sa matrice de dispersion Σ_Y est inversible, auquel cas :

$$f(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma_Y)}} e^{-\frac{1}{2}(y-\mu)'\Sigma_Y^{-1}(y-\mu)}.$$

La non-inversibilité de Σ_Y signifie que le vecteur Y ne prend ses valeurs que dans un sous-espace affine de dimension $n_0 < n$, sur lequel il est distribué comme un vecteur gaussien n_0-dimensionnel. Certaines lois classiques en statistiques sont définies à partir de la loi normale.

Définition 17 (Lois du khi-deux, de Student et de Fisher)
Soit $X_1, \ldots, X_d$ des variables aléatoires iid suivant une loi normale centrée réduite, autrement dit le vecteur $\mathbf{X} = [X_1, \ldots, X_d]'$ est gaussien $\mathcal{N}(0, I_d)$.

- *La loi de la variable $S = \|\mathbf{X}\|^2 = X_1^2 + \cdots + X_d^2$ est dite loi du khi-deux à d degrés de liberté, ce que l'on note $S \sim \chi^2_d$.*
- *Si $Y \sim \mathcal{N}(0,1)$ est indépendante de S, on dit que $T = \frac{Y}{\sqrt{S/d}}$ suit une loi de Student à d degrés de liberté et on note $T \sim \mathcal{T}_d$.*
- *Si $S_1 \sim \chi^2_{d_1}$ est indépendante de $S_2 \sim \chi^2_{d_2}$, on dit que $F = \frac{S_1/d_1}{S_2/d_2}$ suit une loi de Fisher à (d_1, d_2) degrés de liberté, noté $F \sim \mathcal{F}^{d_1}_{d_2}$ ou $F \sim \mathcal{F}(d_1, d_2)$.*

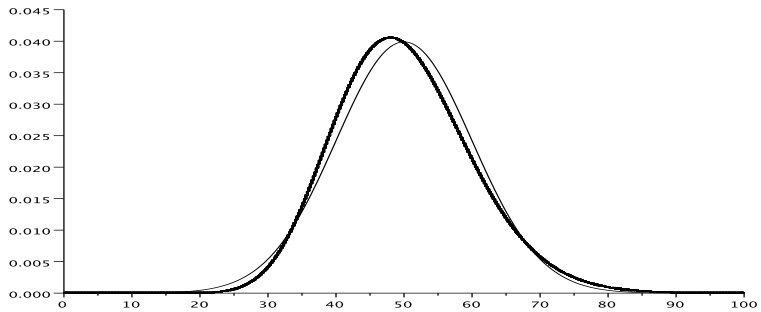

FIGURE 2.5 – Densité d'un χ^2_{50} (trait gras) et densité d'une $\mathcal{N}(50, 100)$ (trait fin).

Rappelons que si $X \sim \mathcal{N}(0,1)$, alors pour tout entier naturel n,

$$\mathbb{E}[X^{2n+1}] = 0 \qquad \text{et} \qquad \mathbb{E}[X^{2n}] = \frac{(2n)!}{2^n n!}$$

d'où l'on déduit que si $S \sim \chi^2_d$ alors

$$\mathbb{E}[S] = d \qquad \text{et} \qquad \mathrm{Var}(S) = 2d.$$

Par ailleurs, lorsque d est grand, on sait par le Théorème Central Limite que S suit approximativement une loi normale de moyenne d et de variance $2d$: $S \approx \mathcal{N}(d, 2d)$. Ainsi, pour d grand, environ 95% des valeurs de S se situent dans l'intervalle $[d - 2\sqrt{2d}, d + 2\sqrt{2d}]$. Ceci est illustré figure 2.5 pour $d = 50$ ddl. Notons enfin le lien avec la loi Gamma : dire que $S \sim \chi^2_d$ est équivalent à dire que $S \sim \Gamma(d/2, 1/2)$, ce qui donne l'expression de sa densité, laquelle ne sera pas utile dans ce qui suit.

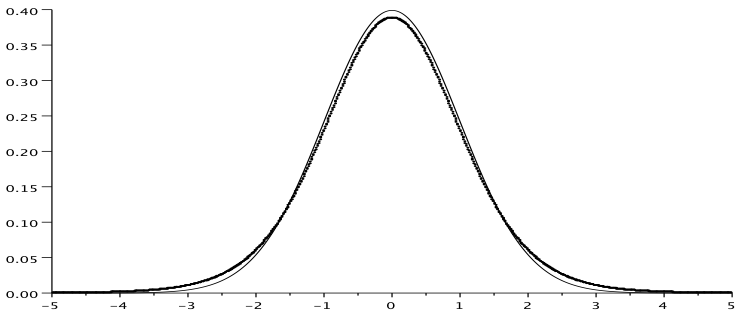

FIGURE 2.6 – Densité d'une $\mathcal{T}_{10}$ (trait gras) et densité d'une $\mathcal{N}(0,1)$ (trait fin).

Concernant la loi de Student : lorsque $d = 1$, T suit une loi de Cauchy et n'a donc pas d'espérance (ni, a fortiori, de variance). Pour $d = 2$, T est centrée mais de variance infinie. Pour $d \geq 3$ (le cas qui nous intéresse), T est centrée et de variance $\frac{d}{d-2}$. D'autre part, lorsque d devient grand, on sait par la Loi des Grands Nombres que le dénominateur tend vers 1. De fait, par le Lemme de Slutsky, lorsque d tend vers l'infini, T tend en loi vers une gaussienne centrée réduite : $T \approx \mathcal{N}(0,1)$. Ceci est illustré figure 2.6 pour $d = 10$ ddl. Par conséquent, lorsque d sera grand, on pourra remplacer les quantiles d'une loi de Student $\mathcal{T}_d$ par ceux d'une loi $\mathcal{N}(0,1)$.

Une remarque enfin sur la loi de Fisher : dans la suite, typiquement, d_2 sera grand, de sorte qu'à nouveau la Loi des Grands Nombres implique que S_2/d_2 tend vers 1. Dans ce cas, F peut se voir comme un khi-deux normalisé par son degré de liberté : $F \approx \chi^2_{d_1}/d_1$. Ceci est illustré figure 2.7 pour $d_1 = 2$ et $d_2 = 10$.

Proposition 10 (Vecteur gaussien et Loi du χ^2)
Soit $Y \sim \mathcal{N}(\mu, \Sigma_Y)$ un vecteur gaussien dans $\mathbb{R}^n$. Si Σ_Y est inversible, alors

$$(Y - \mu)' \Sigma_Y^{-1} (Y - \mu) \sim \chi^2_n$$

loi du khi-deux à n degrés de liberté.

Preuve. Puisque Σ_Y est symétrique définie positive, elle est diagonalisable en base orthonormée, c'est-à-dire sous la forme $\Sigma_Y = Q \Delta Q'$, avec $Q' = Q^{-1}$ et Δ matrice diagonale de coefficients

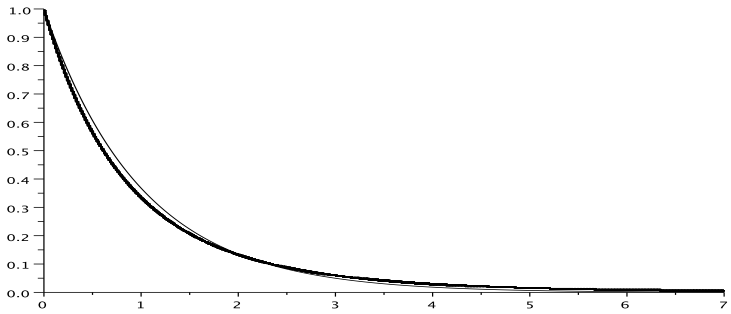

FIGURE 2.7 – Densité d'une $\mathcal{F}_{10}^2$ (trait gras) et densité d'un $\frac{\chi_2^2}{2}$ (trait fin).

diagonaux $\delta_1, \ldots, \delta_n$ tous strictement positifs. Notons $\Delta^{-1/2}$ la matrice diagonale de coefficients diagonaux $1/\sqrt{\delta_1}, \ldots, 1/\sqrt{\delta_n}$. Alors

$$\Sigma_Y = Q\Delta Q' \implies \Sigma_Y^{-1} = Q\Delta^{-1}Q' = (Q\Delta^{-1/2}Q')(Q\Delta^{-1/2}Q') = \Sigma_Y^{-1/2}\Sigma_Y^{-1/2}.$$

Par conséquent

$$(Y-\mu)'\Sigma_Y^{-1}(Y-\mu) = (\Sigma_Y^{-1/2}(Y-\mu))'(\Sigma_Y^{-1/2}(Y-\mu)).$$

Or par stabilité des vecteurs gaussiens par transformations affines, on a

$$Y \sim \mathcal{N}(\mu, \Sigma_Y) \implies \Sigma_Y^{-1/2}(Y-\mu) \sim \mathcal{N}(0, I_n),$$

donc le vecteur $V = [V_1, \ldots, V_n]' = \Sigma_Y^{-1/2}(Y-\mu)$ est gaussien standard et

$$(Y-\mu)'\Sigma_Y^{-1}(Y-\mu) = \|V\|^2 = V_1^2 + \cdots + V_n^2 \sim \chi_n^2,$$

loi du khi-deux à n degrés de liberté.

■

Remarque : dans la preuve précédente, passer du vecteur Y au vecteur $V = \Sigma^{-1/2}(Y-\mu)$ revient à centrer et réduire Y, exactement comme on le fait en dimension 1.

Le Théorème de Cochran, très utile dans la suite, assure que la décomposition d'un vecteur gaussien sur des sous-espaces orthogonaux donne des variables indépendantes dont on peut expliciter les lois. Il peut ainsi être vu comme une version aléatoire du Théorème de Pythagore.

Théorème 9 (Cochran)
Soit $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, $\mathcal{M}$ un sous-espace de $\mathbb{R}^n$ de dimension p, P la matrice de projection orthogonale sur $\mathcal{M}$ et $P_\perp = I_n - P$ la matrice de projection orthogonale sur $\mathcal{M}^\perp$. Nous avons les propriétés suivantes :

(i) $PY \sim \mathcal{N}(P\mu, \sigma^2 P)$ et $P_\perp Y \sim \mathcal{N}(P_\perp \mu, \sigma^2 P_\perp)$;
(ii) les vecteurs PY et $P_\perp Y = (Y - PY)$ sont indépendants ;
(iii) $\frac{\|P(Y-\mu)\|^2}{\sigma^2} \sim \chi_p^2$ et $\frac{\|P_\perp(Y-\mu)\|^2}{\sigma^2} \sim \chi_{n-p}^2$.

Nous allons maintenant voir comment ce résultat s'applique dans notre cadre.

2.2.2 Lois des estimateurs et domaines de confiance

En effet, pour ce qui nous concerne, la gaussianité des résidus implique celle du vecteur Y :

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \implies Y = X\beta + \boldsymbol{\varepsilon} \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

Dès lors, les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ peuvent être vus comme des projections de vecteurs gaussiens sur des sous-espaces orthogonaux.

Propriétés 1 (Lois des estimateurs avec variance connue)
Sous les hypothèses ($\mathcal{H}$), nous avons :

(i) $\hat{\beta}$ est un vecteur gaussien de moyenne β et de variance $\sigma^2(X'X)^{-1}$: $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$;
(ii) $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants ;
(iii) $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

Preuve.

(i) Nous avons vu que

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon}) = \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon},$$

or par hypothèse $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ est un vecteur gaussien. On en déduit que $\hat{\beta}$ est lui aussi un vecteur gaussien, sa loi est donc entièrement caractérisée par la donnée de sa moyenne et de sa matrice de dispersion, lesquelles ont été établies en Proposition 8.
(ii) Comme précédemment, notons $\mathcal{M}_X$ le sous-espace de $\mathbb{R}^n$ engendré par les p colonnes de X et $P_X = X(X'X)^{-1}X'$ la projection orthogonale sur ce sous-espace. On peut noter que :

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X(X'X)^{-1}X')Y = (X'X)^{-1}X'P_XY,$$

donc $\hat{\beta}$ est un vecteur aléatoire fonction de P_XY, tandis que :

$$\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p} = \frac{\|Y - P_XY\|^2}{n-p} = \frac{\|P_{X^\perp}Y\|^2}{n-p}$$

est une variable aléatoire fonction de $P_{X^\perp}Y$. Par le théorème de Cochran, les vecteurs P_XY et $P_{X^\perp}Y$ sont indépendants, il en va donc de même pour toutes fonctions de l'un et de l'autre.
(iii) Puisque $P_{X^\perp}$ est la projection orthogonale sur $\mathcal{M}^\perp(X)$, sous-espace de dimension $(n-p)$ de $\mathbb{R}^n$, on a :

$$\hat{\boldsymbol{\varepsilon}} = (Y - P_XY) = P_{X^\perp}Y = P_{X^\perp}(X\beta + \boldsymbol{\varepsilon}) = P_{X^\perp}\boldsymbol{\varepsilon},$$

où $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Il s'ensuit par le théorème de Cochran que :

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|P_{X^\perp}\boldsymbol{\varepsilon}\|^2}{\sigma^2} = \frac{\|P_{X^\perp}(\boldsymbol{\varepsilon} - \mathbb{E}[\boldsymbol{\varepsilon}])\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

■

Remarque : Le point (iii) et la moyenne du χ^2_{n-p} permettent de retrouver le résultat de la Proposition 9, stipulant que $\hat{\sigma}^2$ est un estimateur non biaisé de σ^2. Mieux, connaissant la variance du χ^2_{n-p}, on en déduit celle de $\hat{\sigma}^2$, donc son erreur quadratique moyenne :

$$\mathrm{Var}\left((n-p)\frac{\hat{\sigma}^2}{\sigma^2}\right) = 2(n-p) \implies \mathrm{Var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n-p} \implies R(\hat{\sigma}^2, \sigma^2) = \frac{2\sigma^4}{n-p}.$$

Par conséquent, pour un modèle donné (i.e. des paramètres $\beta = [\beta_1, \ldots, \beta_p]$ et σ^2 fixés) et une taille n d'échantillon croissante

$$\hat{\sigma}^2 = \hat{\sigma}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2,$$

ce qui est rassurant...

Bien entendu, le premier point de la Proposition 1 n'est pas satisfaisant pour obtenir des régions de confiance sur β car il suppose la variance σ^2 connue, ce qui n'est pas le cas en général. La proposition suivante permet de résoudre le problème.

Proposition 11 (Lois des estimateurs avec variance inconnue)
Sous les hypothèses ($\mathcal{H}$) :

(i) pour $j = 1, \ldots, p$, nous avons

$$T_j := \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}_{n-p}.$$

(ii) On a par ailleurs

$$F := \frac{1}{p\hat{\sigma}^2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \sim \mathcal{F}_{n-p}^p.$$

Preuve :

(i) D'après la proposition précédente, on sait d'une part que $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2(X'X)^{-1}_{jj})$, d'autre part que $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ et enfin que $\hat{\beta}_j$ et $\hat{\sigma}^2$ sont indépendants. Il ne reste plus qu'à écrire T_j sous la forme

$$T_j = \frac{\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X'X)^{-1}_{jj}}}}{\frac{\hat{\sigma}}{\sigma}}$$

pour reconnaître une loi de Student $\mathcal{T}_{n-p}$.
(ii) Puisque $\hat{\beta}$ est un vecteur gaussien de moyenne β et de matrice de covariance $\sigma^2(X'X)^{-1}$, la Proposition 10 assure que

$$\frac{1}{\sigma^2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \sim \chi^2_p.$$

Il reste à remplacer σ^2 par $\hat{\sigma}^2$ en se souvenant que $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ et du fait que $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants. On obtient bien alors la loi de Fisher annoncée.

■

Les variables T_j et F du résultat précédent sont des exemples de variables **pivotales**. Ce ne sont pas des statistiques au sens de la Définition 9 du Chapitre 1, car elles font intervenir le paramètre β du modèle. Néanmoins leur loi est, elle, bel et bien indépendante de ce paramètre. Comme nous le verrons, l'avantage des variables pivotales est de permettre la construction de domaines de confiance. Auparavant, illustrons sur un exemple le second point de la Proposition 11.

Exemple : régression linéaire simple. Considérons le cas $p = 2$, de sorte que

$$(\hat{\beta} - \beta) = \begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{bmatrix}.$$

Si la constante fait partie du modèle, nous sommes dans le cadre d'une régression linéaire simple avec, pour tout $i \in \{1, \ldots, n\}$, $y_i = \beta_1 + \beta_2 x + \varepsilon_i$. Dans ce cas, $\hat{\beta}_1$ et $\hat{\beta}_2$ sont respectivement

l'ordonnée à l'origine et la pente de la droite des moindres carrés. X est la matrice $n \times 2$ dont la première colonne est uniquement composée de 1 et la seconde des x_i, si bien que

$$X'X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix},$$

et le point (ii) de la Proposition 11 s'écrit

$$\frac{1}{2\hat{\sigma}^2}\left(n(\hat{\beta}_1 - \beta_1)^2 + 2n\bar{x}(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) + \sum x_i^2(\hat{\beta}_2 - \beta_2)^2\right) \sim \mathcal{F}^2_{n-2},$$

ce qui nous permettra de construire une ellipse de confiance pour $\beta = (\beta_1, \beta_2)$. Plus généralement, pour $p > 2$, (ii) donnera des hyper-ellipsoïdes de confiance pour β centrés en $\hat{\beta}$. Par ailleurs, ce résultat est à la base de la distance de Cook en validation de modèle.

Les logiciels donnent usuellement des intervalles de confiance pour les paramètres β_j pris séparément. Cependant, ces intervalles de confiance ne tiennent pas compte de la dépendance entre les β_j, laquelle incite plutôt à étudier des domaines de confiance. Nous allons donc traiter les deux cas, en considérant σ^2 inconnue, ce qui est à peu près toujours vrai en pratique.

Théorème 10 (Intervalles et Régions de Confiance)
(i) Pour tout $j \in \{1, \ldots, p\}$, un intervalle de confiance de niveau $(1-\alpha)$ pour β_j est :

$$\left[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}\right],$$

où $t_{n-p}(1-\alpha/2)$ est le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-p}$.
(ii) Un intervalle de confiance de niveau $(1-\alpha)$ pour σ^2 est :

$$\left[\frac{(n-p)\hat{\sigma}^2}{c_{n-p}(1-\alpha/2)}, \frac{(n-p)\hat{\sigma}^2}{c_{n-p}(\alpha/2)}\right],$$

où $c_{n-p}(\alpha/2)$ et $c_{n-p}(1-\alpha/2)$ sont les quantiles d'ordres $\alpha/2$ et $(1-\alpha/2)$ d'une loi χ^2_{n-p}.
(iii) Une région de confiance de niveau $(1-\alpha)$ pour β est l'intérieur de l'hyper-ellipsoïde défini par

$$\left\{\beta \in \mathbb{R}^p : \ \frac{1}{p\hat{\sigma}^2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \leq f^p_{n-p}(1-\alpha)\right\}. \tag{2.3}$$

où $f^p_{n-p}(1-\alpha)$ est le quantile de niveau $(1-\alpha)$ d'une loi de Fisher $\mathcal{F}^q_{n-p}$.

Preuve. Il suffit d'appliquer le point (iii) des Propriétés 1 et les résultats de la Proposition 11. ■

Rappel : soit (x_0, y_0) un point de $\mathbb{R}^2$, $c^2 > 0$ une constante et S une matrice 2×2 symétrique définie positive, alors l'ensemble des points (x, y) du plan tels que

$$[x - x_0, y - y_0]\, S \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \leq c^2 \iff s_{11}(x - x_0)^2 + 2s_{12}(x - x_0)(y - y_0) + s_{22}(y - y_0)^2 \leq c^2$$

est l'intérieur d'une ellipse centrée en (x_0, y_0) dont les axes correspondent aux directions données par les vecteurs propres de S. Il suffit pour s'en convaincre de considérer la diagonalisation $S = Q\Delta Q'$, avec Δ diagonale de coefficients diagonaux δ_1^2 et δ_2^2, et le changement de coordonnées

$$\begin{bmatrix} u \\ v \end{bmatrix} = Q' \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \implies [x - x_0, y - y_0]\, S \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} = \delta_1 u^2 + \delta_2 v^2 \leq c^2.$$

Exemple : reprenons le cas de la régression linéaire simple où $p = q = 2$. Un domaine de confiance de niveau $(1-\alpha)$ pour (β_1, β_2) est défini par l'équation :

$$\left\{(\beta_1,\beta_2) \in \mathbb{R}^2 : \frac{1}{2\hat{\sigma}^2}\left(n(\beta_1-\hat{\beta}_1)^2 + 2n\bar{x}(\beta_1-\hat{\beta}_1)(\beta_2-\hat{\beta}_2) + \sum x_i^2(\beta_2-\hat{\beta}_2)^2\right) \leq f^2_{n-p}(1-\alpha)\right\}.$$

Cette région de confiance est donc l'intérieur d'une ellipse centrée en $(\hat{\beta}_1, \hat{\beta}_2)$ et d'axes donnés par les vecteurs propres de la matrice $X'X$. Considérons maintenant les intervalles de confiance $\hat{I}_1$ et $\hat{I}_2$ de niveau $(1-\alpha)$ pour β_1 et β_2 donnés par le point (i) et le rectangle $\hat{R} = \hat{I}_1 \times \hat{I}_2$. La borne de l'union implique

$$\mathbb{P}((\beta_1,\beta_2) \notin \hat{R}) = \mathbb{P}(\{\beta_1 \notin \hat{I}_1\} \cup \{\beta_2 \notin \hat{I}_2\}) \leq \mathbb{P}(\beta_1 \notin \hat{I}_1) + \mathbb{P}(\beta_2 \notin \hat{I}_2) \leq 2\alpha,$$

et $\hat{R}$ est un domaine de confiance de niveau $(1-2\alpha)$ seulement... Pour obtenir un rectangle de confiance de niveau $(1-\alpha)$, il faut partir d'intervalles de confiance de niveau $(1-\alpha/2)$. La figure 2.8 permet de faire le distinguo entre intervalles de confiance considérés séparément pour β_1 et β_2 et région de confiance simultanée pour (β_1, β_2).

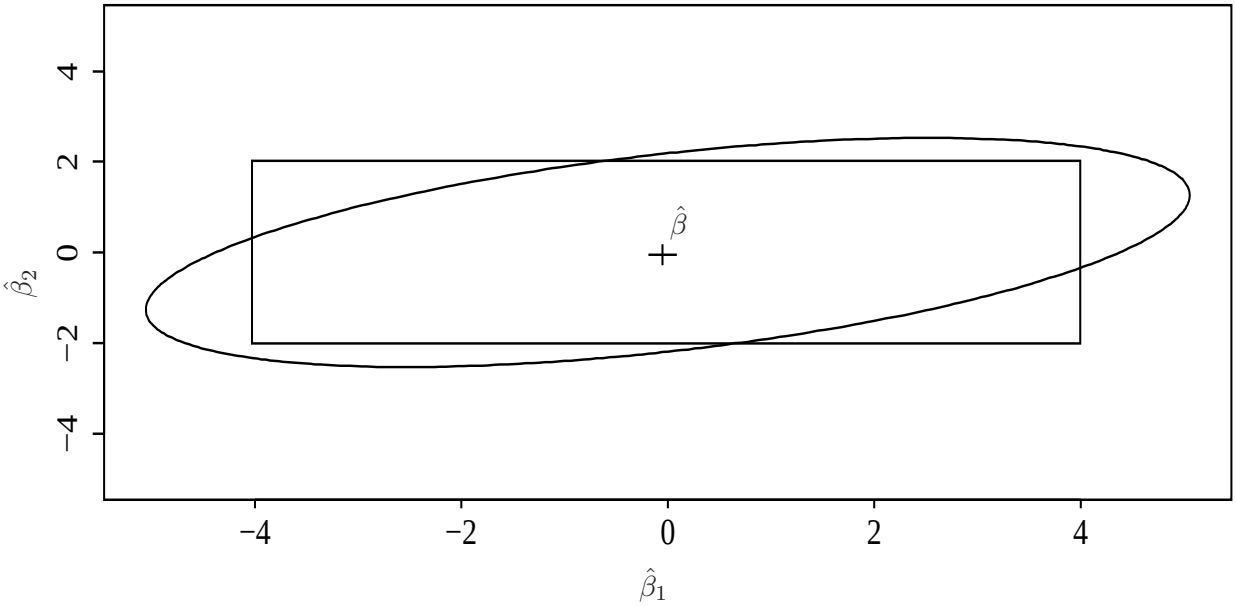

FIGURE 2.8 – Comparaison entre ellipse et rectangle de confiance.

2.2.3 Tests d'hypothèses

Reprenons l'exemple de la prévision de l'ozone vue en début de chapitre. Nous avons décidé de modéliser les pics d'ozone O_3 par la température à midi T, le vent V et la nébulosité à midi N. Il paraît alors raisonnable de se poser par exemple les questions suivantes :

1. Est-ce que la valeur de O_3 est influencée par la variable vent V ?
2. Y a-t-il un effet nébulosité ?
3. Est-ce que la valeur de O_3 est influencée par le vent V ou la température T ?

Rappelons que le modèle utilisé est le suivant :

$$O_{3i} = \beta_1 + \beta_2 T_i + \beta_3 V_i + \beta_4 N_i + \varepsilon_i$$

En termes de tests d'hypothèses, les questions ci-dessus se traduisent comme suit :

1. correspond à $H_0 : \beta_3 = 0$, contre $H_1 : \beta_3 \neq 0$.
2. correspond à $H_0 : \beta_4 = 0$, contre $H_1 : \beta_4 \neq 0$.
3. correspond à $H_0 : \beta_2 = \beta_3 = 0$, contre $H_1 :\ \beta_2 \neq 0$ ou $\beta_3 \neq 0$.

Ces tests d'hypothèses reviennent à tester la nullité d'un ou plusieurs paramètres. Si l'on teste plusieurs paramètres à la fois, on parle de nullité simultanée des coefficients. Ceci signifie que, sous l'hypothèse H_0, certains coefficients sont nuls, donc les variables correspondant à ceux-ci ne sont pas utiles pour la modélisation du phénomène. Ce cas de figure revient à comparer deux modèles emboîtés, l'un étant un cas particulier de l'autre.

Le plan d'expérience privé de ces variables sera noté X_0 et les colonnes de X_0 engendreront un sous-espace noté $\mathcal{M}_0 = \mathcal{M}_{X_0}$. De même, pour alléger les notations, nous noterons $\mathcal{M} = \mathcal{M}_X$ l'espace engendré par les colonnes de X. Le niveau de risque des tests sera fixé de façon classique à α.

Tests entre modèles emboîtés

Rappelons tout d'abord le modèle :

$$Y = X\beta + \boldsymbol{\varepsilon} \quad \text{sous les hypothèses } (\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

En particulier, cela veut dire que $\mathbb{E}[Y] = X\beta \in \mathcal{M}$, sous-espace de dimension p de $\mathbb{R}^n$ engendré par les p colonnes de X. Pour faciliter les notations, on suppose vouloir tester la nullité simultanée des $q = (p - p_0)$ derniers coefficients du modèle (avec $q \leq p$ of course !). Le problème s'écrit alors de la façon suivante :

$$H_0 :\ \beta_{p_0+1} = \cdots = \beta_p = 0 \quad \text{contre} \quad H_1 :\ \exists j \in \{p_0 + 1, \cdots, p\} : \beta_j \neq 0.$$

Que signifie $H_0 :\ \beta_{p_0+1} = \cdots = \beta_p = 0$ en termes de modèle ? Si les q derniers coefficients sont nuls, le modèle devient

$$Y = X_0\beta_0 + \boldsymbol{\varepsilon} \quad \text{sous les hypothèses } (\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X_0) = p_0 \\ (\mathcal{H}_2) : \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

La matrice X_0, de taille $n \times p_0$, est composée des p_0 premières colonnes de X et β_0 est un vecteur colonne de taille p_0. Puisque X est supposée de rang p, X_0 est de rang p_0 donc ses colonnes engendrent un sous-espace $\mathcal{M}_0$ de $\mathbb{R}^n$ de dimension p_0. Ce sous-espace $\mathcal{M}_0$ est bien évidemment aussi un sous-espace de $\mathcal{M}$. Sous H_0, l'espérance de Y, à savoir $\mathbb{E}[Y] = X_0\beta_0$, appartiendra à ce sous-espace $\mathcal{M}_0$. Maintenant que les hypothèses du test sont fixées, il faut proposer une statistique de test. Nous allons voir une approche géométrique et intuitive de l'affaire.

Approche géométrique

Considérons le sous-espace $\mathcal{M}_0$. Nous avons écrit que sous $H_0 : \mathbb{E}[Y] = X_0\beta_0 \in \mathcal{M}_0$. Dans ce cas, la méthode des moindres carrés consiste à projeter Y non plus sur $\mathcal{M}$ et à obtenir $\hat{Y}$ ainsi que les résidus $\hat{\boldsymbol{\varepsilon}} = Y - \hat{Y}$, mais sur $\mathcal{M}_0$ et à obtenir $\hat{Y}_0$ ainsi que les résidus $\hat{\boldsymbol{\varepsilon}}_0 = Y - \hat{Y}_0$. Tout ceci est illustré en Figure 2.9.

L'idée intuitive du test, et donc du choix de conserver ou non H_0, est la suivante : si la projection $\hat{Y}_0$ de Y dans $\mathcal{M}_0$ est "proche" de la projection $\hat{Y}$ de Y dans $\mathcal{M}$, alors il est judicieux d'accepter H_0. En effet, si l'information apportée par les deux modèles est "à peu près la même", il vaut mieux conserver le modèle le plus petit : c'est le principe de parcimonie (rasoir d'Ockham, diraient les philosophes).

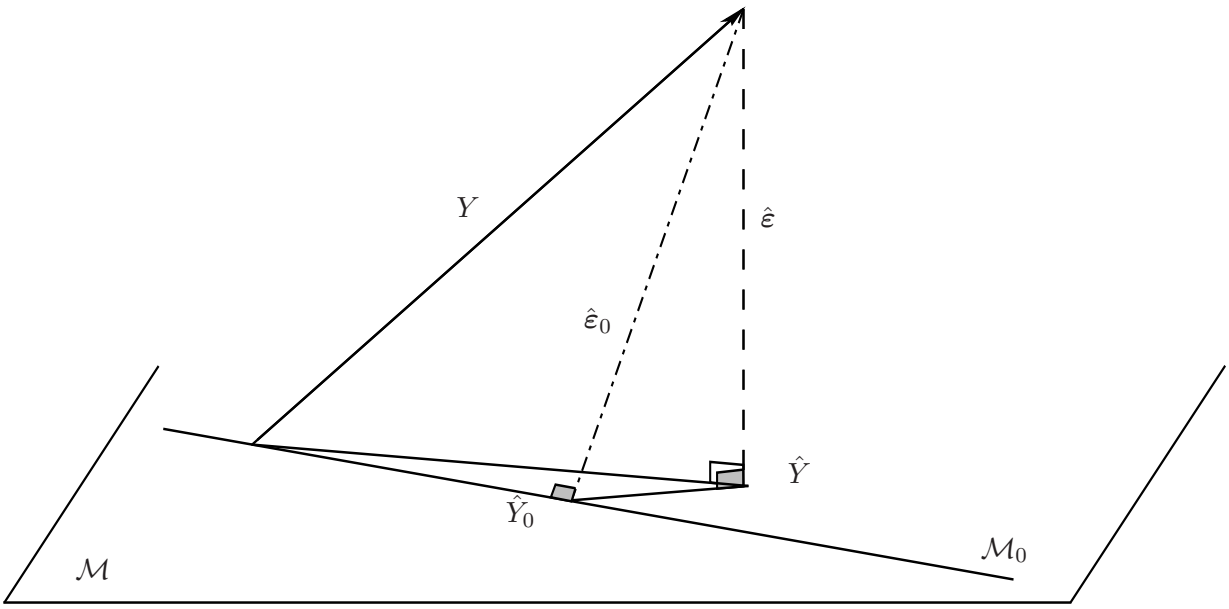

FIGURE 2.9 – Représentation des projections.

Encore faut-il préciser ce qu'on entend par "proche". Pour ce faire, nous pouvons utiliser la distance euclidienne entre $\hat{Y}_0$ et $\hat{Y}$, ou son carré $\|\hat{Y} - \hat{Y}_0\|^2$. Mais cette distance sera variable selon les données et les unités de mesures utilisées. Pour nous affranchir de ce problème d'échelle, nous allons "standardiser" cette distance en la divisant par la norme au carré de l'erreur estimée

$$\|\hat{\boldsymbol{\varepsilon}}\|^2 = \|Y - \hat{Y}\|^2 = (n-p)\hat{\sigma}^2.$$

Les vecteurs aléatoires $(\hat{Y} - \hat{Y}_0)$ et $\hat{\boldsymbol{\varepsilon}}$ n'appartenant pas à des sous-espaces de même dimension, il faut encore diviser chaque terme par son degré de liberté respectif, soit $q = p - p_0$ et $n - p$. Tout ceci nous amène à considérer la quantité suivante :

$$F \quad = \quad \frac{\|\hat{Y} - \hat{Y}_0\|^2/q}{\|Y - \hat{Y}\|^2/(n-p)} = \frac{\|\hat{Y} - \hat{Y}_0\|^2/(p-p_0)}{\|Y - \hat{Y}\|^2/(n-p)}.$$

Pour utiliser cette statistique de test, il faut connaître au moins sa loi sous H_0.

Proposition 12 (Test entre modèles emboîtés)
Sous l'hypothèse H_0, on a la statistique de test suivante

$$F = \frac{n-p}{q} \times \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}\|^2} = \frac{n-p}{q} \times \frac{SCR_0 - SCR}{SCR} \sim \mathcal{F}^q_{n-p},$$

loi de Fisher à $(q, n-p)$ degrés de liberté. Le test consistant à rejeter H_0 si et seulement si $F(\omega) > f^q_{n-p}(1-\alpha)$ est donc de niveau α.

Preuve. Sous H_0, on sait que $Y \sim \mathcal{N}(X_0\beta_0, \sigma^2 I_n)$. La statistique de test correspondant au rapport de deux normes au carré, nous allons déterminer la loi du numérateur, celle du dénominateur et constater leur indépendance. En notant P (resp. P_0) la matrice de projection orthogonale sur $\mathcal{M}$ (resp. $\mathcal{M}_0$), nous savons que :

$$\hat{Y} - \hat{Y}_0 = PY - P_0Y,$$

or $\mathcal{M}_0 \subset \mathcal{M}$ donc $P_0 Y = P_0 P Y$ et :

$$\hat{Y} - \hat{Y}_0 = \quad PY - P_0 PY = (I_n - P_0)PY = P_0^{\perp} PY.$$

Ceci montre que $\hat{Y} - \hat{Y}_0$ est la projection orthogonale de Y sur $\mathcal{M}_0^{\perp} \cap \mathcal{M}$, supplémentaire orthogonal de $\mathcal{M}_0$ dans $\mathcal{M}$. Puisque $\dim(\mathcal{M}) = p$ et $\dim(\mathcal{M}_0) = p_0$, c'est donc une projection sur un sous-espace de dimension $q = p - p_0$. La figure 2.9 permet à nouveau de visualiser ces notions d'orthogonalité. Par ailleurs, on a déjà vu que

$$Y - \hat{Y} = (I_n - P)Y = P^{\perp} Y.$$

Ainsi, sous H_0, les vecteurs aléatoires $(Y - \hat{Y})$ et $(\hat{Y} - \hat{Y}_0)$ sont les projections d'un même vecteur gaussien $Y \sim \mathcal{N}(X_0\beta_0, \sigma^2 I_n)$ sur deux sous-espaces orthogonaux : par Cochran, ils sont donc indépendants.

Le théorème de Cochran nous renseigne par ailleurs sur les lois des numérateur et dénominateur. Le dénominateur a déjà été vu :

$$\frac{1}{\sigma^2}\|Y - \hat{Y}\|^2 \sim \chi^2_{n-p}.$$

Le numérateur est la projection orthogonale de $Y \sim \mathcal{N}(X_0\beta_0, \sigma^2 I_n)$ sur un sous-espace de dimension $q = p - p_0$, donc

$$\frac{1}{\sigma^2}\|P_0^{\perp} P(Y - X_0\beta_0)\|^2 = \frac{1}{\sigma^2}\|P_0^{\perp} PY\|^2 = \frac{1}{\sigma^2}\|\hat{Y} - \hat{Y}_0\|^2 \sim \chi^2_q,$$

car

$$X\beta_0 \in \mathcal{M}_0 \subset \mathcal{M} \implies PX\beta_0 = X\beta_0 \in \mathcal{M}_0 \implies P_0^{\perp} PX_0\beta_0 = 0.$$

Au total, nous avons obtenu la loi de F sous H_0 :

$$F = \frac{n-p}{q} \times \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}\|^2} \sim \mathcal{F}^q_{n-p}.$$

Par ailleurs, la relation $\|\hat{Y} - \hat{Y}_0\|^2 = (SCR_0 - SCR)$ peut se voir facilement en utilisant la figure 2.9, c'est-à-dire en appliquant le théorème de Pythagore au fait que $(Y - \hat{Y})$ et $(\hat{Y} - \hat{Y}_0)$ sont orthogonaux (cf. supra) :

$$\|Y - \hat{Y}_0\|^2 \quad = \quad \|Y - \hat{Y}\|^2 + \|\hat{Y} - \hat{Y}_0\|^2,$$

c'est-à-dire :

$$\|\hat{Y} - \hat{Y}_0\|^2 = \|Y - \hat{Y}_0\|^2 - \|Y - \hat{Y}\|^2 = SCR_0 - SCR.$$

■

Pour conclure, explicitons cette statistique de test dans deux cas particuliers : le premier (dit de Student) est très important ; le second (dit de Fisher global) est anecdotique mais est souvent donné par les logiciels, donc autant savoir à quoi il correspond.

Test de Student de significativité d'un coefficient

Nous voulons tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$, test bilatéral de significativité de β_j. Selon ce qu'on vient de voir, la statistique de test est :

$$F = (n-p) \times \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}\|^2} = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2} \sim \mathcal{F}^1_{n-p}.$$

Nous rejetons H_0 si l'observation de la statistique de test, notée $F(\omega)$, est telle que :

$$F(\omega) > f^1_{n-p}(1-\alpha),$$

où $f^1_{n-p}(1-\alpha)$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à 1 et $(n-p)$ degrés de liberté.

Ce test est en fait équivalent au test de Student déduit de l'intervalle de confiance de la Proposition 10, à savoir que sous H_0

$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} = \frac{\hat{\beta}_j}{\hat{\sigma}_j} \sim \mathcal{T}_{n-p},$$

où

$$\hat{\sigma}_j = \hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$$

est l'écart-type estimé de $\hat{\beta}_j$. On peut en effet montrer (mais ce n'est pas trivial) que $F = T^2$. Nous rejetons H_0 si l'observation de la statistique de test, notée $T(\omega)$, est telle que

$$|T(\omega)| > t_{n-p}(1-\alpha/2),$$

où $t_{n-p}(1-\alpha/2)$ est le quantile d'ordre $(1-\alpha/2)$ d'une loi de Student à $(n-p)$ degrés de liberté. On a bien entendu

$$\mathcal{F}^1_{n-p} \overset{\mathcal{L}}{=} (\mathcal{T}_{n-p})^2 \implies t_{n-p}(1-\alpha/2) = \sqrt{f^1_{n-p}(1-\alpha)}.$$

C'est sous la forme du test de Student que la significativité d'un coefficient apparaît dans tous les logiciels de statistique. Il est donc complètement équivalent au test de Fisher que nous avons proposé lorsqu'on spécialise celui-ci à la nullité d'un seul coefficient.

Test de Fisher global

Si des connaissances a priori du phénomène assurent l'existence d'un terme constant dans la régression, alors pour tester l'influence de tous les autres régresseurs (non constants) sur la réponse Y, on regarde si $\mathbb{E}[Y] = \beta_1$. En d'autres termes, H_0 correspond à la nullité de tous les coefficients sauf la constante. On l'appelle test de Fisher global.

Dans ce cas, en notant $\mathbb{1}$ le vecteur de $\mathbb{R}^n$ uniquement composé de 1 et $\bar{Y} = (Y_1 + \cdots + Y_n)/n$ la moyenne empirique des Y_i, on voit que

$$\hat{Y}_0 = P_0 Y = \mathbb{1}(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1} \times Y = \bar{Y}\mathbb{1}$$

et la statistique de test est la suivante :

$$F = \frac{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2/(p-1)}{\|Y - \hat{Y}\|^2/(n-p)} = \frac{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2/(p-1)}{\hat{\sigma}^2} \sim \mathcal{F}^{p-1}_{n-p}.$$

Ce test est aussi appelé test du R^2 par certains logiciels en raison de son lien avec le coefficient de détermination R^2. Celui-ci peut s'interpréter comme le pourcentage de variance dans les données expliqué par le modèle, et est défini comme suit :

$$R^2 := \frac{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2} \implies F = \frac{n-p}{p-1} \times \frac{R^2}{1-R^2}.$$

Remarque : sauf à considérer un modèle stupide où les variables explicatives n'ont rien à voir avec la variable à expliquer, ce test sera toujours rejeté (p-value epsilonesque). C'est en ce sens qu'il est plutôt anecdotique.

Chapitre 3

Estimation non paramétrique

Introduction

Dans tout ce chapitre, on considère le modèle d'échantillonnage en dimension 1, autrement dit on dispose d'un échantillon $(X_1, \dots, X_n)$ de variables aléatoires réelles iid de loi inconnue P_X. Contrairement au Pile ou Face du Chapitre 1 et au modèle linéaire gaussien du Chapitre 2, la loi P_X n'est plus supposée indexée par un paramètre θ fini-dimensionnel, si bien que l'on se situe dans un cadre non paramétrique. La loi P_X étant caractérisée par la fonction de répartition associée, c'est cette fonction F que l'on va estimer à partir de sa version empirique F_n. Dans ce contexte, des équivalents de la loi des grands nombres et du théorème central limite sont donnés par les Théorèmes de Glivenko-Cantelli et de Kolmogorov-Smirnov.

3.1 Loi et moments empiriques

Commençons par une définition très générale.

Définition 18 (Convergence et normalité asymptotique)
Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires, μ un paramètre réel inconnu et $\hat{\mu}_n = \hat{\mu}_n(X_1, \dots, X_n)$ un estimateur de μ. On dit que la suite $(\hat{\mu}_n)_{n\geq 1}$ est :

— *convergente, ou consistante, si*

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

— *asymptotiquement normale s'il existe $\sigma^2 > 0$ tel que*

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

Remarque : souvent, par abus de langage, on dira simplement que $\hat{\mu}_n$ est un estimateur consistant et asymptotiquement normal de μ.

Rappelons que, d'après la Proposition 6 du Chapitre 1, la normalité asymptotique de $(\hat{\mu}_n)_{n\geq 1}$ implique sa convergence. Par ailleurs, si l'on dispose d'une suite $(\hat{\sigma}_n^2)_{n\geq 1}$ d'estimateurs qui converge vers σ^2, alors le Théorème de Slutsky entraîne que

$$\sqrt{n}\,\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

ce qui permet encore de construire des intervalles de confiance asymptotiques pour μ.

3.1.1 Moyenne et variance empiriques

Partant d'un échantillon $(X_n)_{n\geq 1}$ iid, l'exemple le plus simple d'estimateur de la moyenne $\mu = \mathbb{E}[X_1]$ est celui de la moyenne empirique :

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

Ses propriétés découlent directement de la loi des grands nombres et du théorème central limite.

Proposition 13 (Convergence et normalité asymptotique de la moyenne empirique)
Si les variables $(X_n)_{n\geq 1}$ sont iid et ont un moment d'ordre 2, avec $\mathbb{E}[X_1] = \mu$ et $\mathrm{Var}(X_1) = \sigma^2$, alors la moyenne empirique $\bar{X}_n$ est un estimateur non biaisé, convergent et asymptotiquement normal :

$$\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu \qquad \text{et} \qquad \sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Puisque la variance σ^2 des X_i apparaît dans le résultat de normalité asympotique, il est naturel de chercher à l'estimer à son tour. Ici, les choses se compliquent un peu en raison du biais de la variance empirique.

Lemme 1 (Estimateurs de la variance)
Sous les mêmes hypothèses qu'en Proposition 13, on appelle variance empirique l'estimateur

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2,$$

et estimateur sans biais de la variance

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

lequel vérifie bien $\mathbb{E}[\hat{s}_n^2] = \sigma^2 = \mathrm{Var}(X_1)$.

Attention ! La notation $\hat{s}_n^2$ dans cette définition correspond au $\hat{\sigma}_n^2$ du Chapitre 2.

Preuve. Partons de la seconde expression de la variance empirique, à savoir

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2. \tag{3.1}$$

Puisque $\mathbb{E}[Y^2] = \mathrm{Var}(Y) + \mathbb{E}[Y]^2$ pour toute variable aléatoire de carré intégrable, la moyenne du premier terme est triviale :

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] = \mathbb{E}[X_1^2] = \mathrm{Var}(X_1) + \mathbb{E}[X_1]^2 = \sigma^2 + \mu^2.$$

Le second est à peine plus difficile si l'on tient compte du fait que la variance de la somme de variables indépendantes est égale à la somme des variances :

$$\mathbb{E}[\bar{X}_n^2] = \mathrm{Var}(\bar{X}_n) + \mathbb{E}[\bar{X}_n]^2 = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) + \mathbb{E}[X_1]^2 = \frac{1}{n}\mathrm{Var}(X_1) + \mathbb{E}[X_1]^2 = \frac{\sigma^2}{n} + \mu^2,$$

ce qui mène au résultat annoncé.

■

Les deux estimateurs sont asymptotiquement équivalents puisque

$$\frac{\hat{\sigma}_n^2}{\hat{s}_n^2} = \frac{n-1}{n} = 1 - \frac{1}{n} \xrightarrow[n\to\infty]{\mathbb{P}} 1,$$

et ont les mêmes propriétés de convergence et de normalité asymptotique.

Proposition 14 (Convergence et normalité asymptotique de la variance empirique) *Si les variables $(X_n)_{n\geq 1}$ sont iid et admettent un moment d'ordre 2, avec $\mathrm{Var}(X_1) = \sigma^2$, alors les estimateurs $\hat{\sigma}_n^2$ et $\hat{s}_n^2$ sont convergents :*

$$\hat{\sigma}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2 \qquad \text{et} \qquad \hat{s}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2.$$

Si l'on suppose de plus l'existence d'un moment d'ordre 4 pour les X_i, alors il y a aussi normalité asymptotique :

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, v^2) \qquad \text{et} \qquad \sqrt{n}(\hat{s}_n^2 - \sigma^2) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, v^2),$$

où, en notant $\mu = \mathbb{E}[X_1]$,

$$v^2 = \mathrm{Var}((X_1 - \mu)^2) = \mathbb{E}[(X_1 - \mu)^4] - \sigma^4.$$

Preuve. Pour la convergence, on part de la formule (3.1) à laquelle on applique deux fois la loi des grands nombres :

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \mathrm{Var}(X_1) = \sigma^2.$$

Par la remarque ci-dessus et le Théorème de Slutsky, le même résultat s'applique à $\hat{s}_n^2$. Pour la normalité asymptotique, on bidouille un peu en considérant les variables iid centrées $Y_i = (X_i - \mu)$ et en notant que

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 - \bar{Y}_n^2 = \overline{Y_n^2} - \bar{Y}_n^2.$$

On peut donc écrire

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \sqrt{n}\left(\overline{Y_n^2} - \sigma^2\right) - \sqrt{n}\,\bar{Y}_n^2 = \sqrt{n}\left(\overline{Y_n^2} - \sigma^2\right) - \bar{Y}_n \times \left(\sqrt{n}\,\bar{Y}_n\right),$$

Par la loi des grands nombres, $\bar{Y}_n$ tend en probabilité vers 0. De plus, le TCL appliqué aux variables Y_i de moyenne nulle et de variance σ^2 donne

$$\sqrt{n}\,\bar{Y}_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

d'où par Slutsky

$$\bar{Y}_n \times \left(\sqrt{n}\,\bar{Y}_n\right) \xrightarrow[n\to\infty]{\mathcal{L}} 0.$$

De même, le TCL appliqué aux variables Y_i^2 de moyenne σ^2 et de variance v^2 nous dit que

$$\sqrt{n}\left(\overline{Y_n^2} - \sigma^2\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, v^2).$$

Il reste à appliquer Slutsky pour recoller les morceaux :

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \sqrt{n}\left(\overline{Y_n^2} - \sigma^2\right) - \bar{Y}_n \times \left(\sqrt{n}\ \bar{Y}_n\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, v^2).$$

Quant à l'estimateur sans biais, tout le travail a déjà été fait ou presque, vu que

$$\sqrt{n}(\hat{s}_n^2 - \sigma^2) = \sqrt{n}(\hat{s}_n^2 - \hat{\sigma}_n^2) + \sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \frac{1}{\sqrt{n}}\hat{s}_n^2 + \sqrt{n}(\hat{\sigma}_n^2 - \sigma^2).$$

Il suffit donc d'invoquer la convergence de $\hat{s}_n^2$ et Slutsky pour le premier terme, et la normalité asymptotique de $\hat{\sigma}_n^2$ pour le second.

■

3.1.2 Loi empirique

On parle de moyenne empirique pour $\bar{X}_n$, or moyenne est synonyme d'espérance en probabilités. On peut en fait voir $\bar{X}_n$ comme une espérance, mais par rapport à une mesure de probabilité aléatoire.

Définition 19 (Loi empirique)
Si $x_1, \ldots, x_n$ sont des réels, on appelle loi empirique des x_i la mesure de probabilité $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Autrement dit, si les x_i sont distincts, ν_n est la loi uniforme sur les x_i. De façon générale, pour tout borélien A de $\mathbb{R}$, on a

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A(x_i) = \frac{|\{i \in \{1,\ldots,n\},\ x_i \in A\}|}{n}.$$

Si $X_1, \ldots, X_n$ sont des variables aléatoires iid, on appelle loi empirique de l'échantillon $(X_1, \ldots, X_n)$ la fonction de $\omega \in \Omega$ définie par

$$\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \implies \nu_n(\omega) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}.$$

Ainsi, la loi empirique ν_n de l'échantillon $(X_1, \ldots, X_n)$ est une **probabilité aléatoire** : à chaque $\omega \in \Omega$ sont associées de nouvelles réalisations $(x_1, \ldots, x_n)$, donc une nouvelle mesure de probabilité. On peut voir ν_n comme une application de l'espace mesurable $(\Omega, \mathcal{F})$ dans l'ensemble des mesures de probabilité à support fini [1] sur la droite réelle. Les quantités associées à la mesure ν_n sont donc des variables aléatoires, dites **empiriques**. Par exemple, pour tout borélien A, la quantité

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A(X_i) = \frac{|\{i \in \{1,\ldots,n\},\ X_i \in A\}|}{n}$$

est appelée fréquence empirique de l'ensemble A. En l'occurrence, puisque les X_i sont iid de loi P_X, c'est une variable aléatoire de loi connue :

$$n \times \nu_n(A) \sim \mathcal{B}(n, P_X(A)),$$

loi binomiale de paramètres n et $P_X(A) = \mathbb{P}(X_1 \in A)$. On peut aussi noter que, si la loi P_X des X_i n'a pas d'atome, alors presque sûrement les X_i sont distincts et ν_n n'est rien d'autre que la mesure uniforme sur ces n points aléatoires.

1. de cardinal entre 1 et n, selon le nombre d'égalités parmi les x_i.

Notation. Dans ce qui suit, $\mathbb{E}_\nu[\varphi(Z)]$ correspond à l'espérance de la variable aléatoire $\varphi(Z)$ lorsque Z a pour loi ν. Comme dans les intégrales dans le cours d'analyse, ceci permet de voir Z comme une variable (aléatoire) muette, puisque dans ce cas $\mathbb{E}_\nu[\varphi(Z)] = \mathbb{E}_\nu[\varphi(T)]$. Ceci étant, si on écrit $\mathbb{E}[\varphi(Z)]$ au lieu de $\mathbb{E}_\nu[\varphi(Z)]$, c'est qu'il n'y a aucune ambiguïté sur la loi de Z.

Considérons des réalisations $(x_1, \dots, x_n)$, la loi empirique ν_n des x_i et une variable aléatoire Y de loi ν_n. Ainsi Y prend les valeurs x_i avec les probabilités $1/n$, donc par définition de l'espérance et par le Théorème de Transfert, on a pour toute fonction φ

$$\mathbb{E}_{\nu_n}[\varphi(Y)] = \sum_{i=1}^{n} \varphi(x_i) \times \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i).$$

Pour un échantillon $(X_1, \dots, X_n)$, en notant ν_n la loi empirique de l'échantillon $(X_1, \dots, X_n)$, la quantité

$$\mathbb{E}_{\nu_n}[\varphi(Y)] = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)$$

est donc la variable aléatoire qui, pour tout $\omega \in \Omega$, vaut

$$\mathbb{E}_{\nu_n(\omega)}[\varphi(Y)] = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i(\omega)) = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i).$$

En particulier, la moyenne empirique peut se voir comme une espérance empirique, puisqu'il suffit de prendre $\varphi(y) = y$:

$$\mathbb{E}_{\nu_n}[Y] = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n,$$

tandis que la variance empirique fait aussi intervenir $\varphi(y) = y^2$:

$$\mathbb{E}_{\nu_n}[Y^2] - \mathbb{E}_{\nu_n}[Y]^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = \hat{\sigma}_n^2.$$

La loi des grands nombres et le théorème central limite impliquent alors que, sous réserve d'intégrabilité,

$$\mathbb{E}_{\nu_n}[\varphi(Y)] = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[\varphi(X_1)] = \mathbb{E}_{P_X}[\varphi(Y)],$$

et

$$\sqrt{n}\left(\mathbb{E}_{\nu_n}[\varphi(Y)] - \mathbb{E}_{P_X}[\varphi(Y)]\right)\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}_{P_X}(Y)).$$

Ces deux résultats montrent que la suite de lois empiriques $(\nu_n)_{n\geq 1}$ tend en un certain sens vers la loi P_X. Puisque les lois de probabilité sur $\mathbb{R}$ sont complètement caractérisées par leurs fonctions de répartition, on va s'intéresser à celles-ci d'un point de vue empirique.

3.2 Fonction de répartition et quantiles empiriques

3.2.1 Statistiques d'ordre et fonction de répartition empirique

Avant de définir la fonction de répartition empirique, il convient de mettre de l'ordre dans l'échantillon.

Définition 20 (Statistiques d'ordre)
Partant d'un échantillon $X_1, \dots, X_n$, les n statistiques d'ordre $X_{(1)}, \dots, X_{(n)}$ s'obtiennent en rangeant l'échantillon par ordre croissant, c'est-à-dire qu'elles vérifient

$$X_{(1)} \leq \cdots \leq X_{(n)}.$$

Notation. On rencontre aussi l'écriture suivante pour les statistiques d'ordre :

$$X_{(1,n)} \leq \cdots \leq X_{(n,n)}.$$

Pour tout k entre 1 et n, la variable $X_{(k)}$ est appelée la k-ème statistique d'ordre. Par exemple, la première statistique d'ordre est le minimum de l'échantillon tandis que la n-ème correspond à son maximum.

Achtung ! Même si les X_i sont iid, les $X_{(i)}$ ne le sont clairement plus : à titre d'exemple, la connaissance de $X_{(1)}$ donne de l'information sur $X_{(2)}$, qui ne peut être plus petit.

D'un point de vue algorithmique, ce rangement croissant peut se faire par un algorithme de tri rapide (ou *quicksort*) dont le coût moyen est en $\mathcal{O}(n \log n)$, ce qui n'est pas cher payé. Notons enfin que la définition précédente ne suppose pas les X_i distincts. C'est néanmoins presque sûrement le cas si la fonction de répartition des X_i est continue (cas d'une loi sans atome).

Définition 21 (Fonction de répartition empirique)
La fonction de répartition empirique F_n d'un échantillon $X_1, \dots, X_n$ est la fonction de répartition de la loi empirique ν_n, donc définie pour tout réel x par

$$F_n(x) = \nu_n(]-\infty, x]) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_{(i)}),$$

ou, de façon équivalente,

$$F_n(x) = \frac{|\{i \in \{1, \dots, n\},\ X_i \leq x\}|}{n} = \frac{|\{i \in \{1, \dots, n\},\ X_{(i)} \leq x\}|}{n},$$

c'est-à-dire la proportion de l'échantillon tombant au-dessous de x.

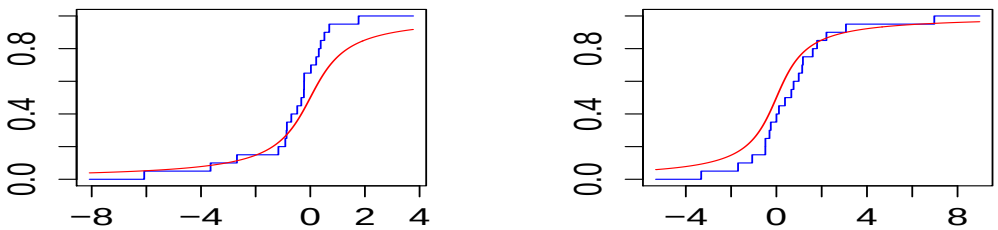

FIGURE 3.1 – En bleu : deux réalisations de F_{20} avec $X_1, \dots, X_{20}$ iid selon une loi de Cauchy. En rouge : fonction de répartition de la loi de Cauchy.

En notant $X_{(n+1)} = +\infty$, cette fonction s'écrit encore

$$F_n(x) = \sum_{i=1}^{n} \frac{i}{n} \mathbb{1}_{[X_{(i)}, X_{(i+1)}[}(x).$$

C'est une fonction (**aléatoire** !) en escalier qui ne présente des sauts qu'aux $X_{(i)}$, ces sauts étant tous égaux à $1/n$ si les X_i sont distincts (cf. Figure 3.1). Dans le cas général, l'amplitude des sauts est toujours un multiple de $1/n$, le multiple en question correspondant au nombre de points de l'échantillon empilés au même endroit.

Proposition 15 (Loi, convergence et normalité asymptotique)
Soit $(X_n)_{n\geq 1}$ des variables iid de fonction de répartition F, alors pour tout réel x, on a :

- *Loi : la variable aléatoire $nF_n(x)$ suit une loi binomiale $\mathcal{B}(n, F(x))$.*
- *Convergence :*
$$F_n(x) \xrightarrow[n\to\infty]{p.s.} F(x).$$
- *Normalité asymptotique :*
$$\sqrt{n}\,(F_n(x) - F(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, F(x)(1-F(x))).$$

Preuve. Dans tous ces résultats, il importe de garder en tête que x est un réel **fixé**. Ainsi $nF_n(x)$ représente tout bonnement le nombre de points de l'échantillon qui tombent à gauche de x :

$$nF_n(x) = \sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i) = \sum_{i=1}^{n} Y_i,$$

où les Y_i sont iid selon une loi de Bernoulli de paramètre

$$p = \mathbb{P}(Y_1 = 1) = \mathbb{P}(X_i \leq x) = F(x),$$

d'où la loi binomiale pour leur somme. De la même façon, la loi des grands nombres appliquée aux variables Y_i assure que

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[Y_1] = F(x),$$

tandis que le TCL donne

$$\sqrt{n}\,(F_n(x) - F(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(Y_1)) = \mathcal{N}(0, F(x)(1-F(x))).$$

■

Ainsi, pour tout réel x, il existe un ensemble $\Omega_0(x)$ de probabilité 1 tel que, pour tout $\omega \in \Omega_0(x)$, pour toute suite de réalisations $x_1 = X_1(\omega), x_2 = X_2(\omega), \ldots$, on a

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(x_i) = \sum_{i=1}^{n} \frac{i}{n} \mathbb{1}_{[x_{(i)},x_{(i+1)}[}(x) \xrightarrow[n\to\infty]{} F(x).$$

A priori, ceci n'assure même pas la convergence simple de F_n vers F de façon presque sûre, car $\Omega_0(x)$ dépend de x, or une intersection non dénombrable d'ensembles de probabilité 1 n'est pas nécessairement de probabilité 1. En fait, de façon presque sûre, il y a bien convergence simple et même mieux : convergence uniforme, comme nous le verrons plus loin avec Glivenko-Cantelli.

3.2.2 Quantiles et quantiles empiriques

Un quantile est défini à partir de la fonction de répartition. Il n'y aucun problème lorsque celle-ci est inversible. Si tel n'est pas le cas, il faut faire un peu attention. Ceci arrive en particulier pour les fonctions de répartition empiriques.

Définition 22 (Inverse généralisée)
Soit F une fonction de répartition. On appelle inverse généralisée de F la fonction définie pour tout $u \in [0,1]$ par

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : \ F(x) \geq u\},$$

avec les conventions $\inf \mathbb{R} = -\infty$ *et* $\inf \emptyset = +\infty$.

Si F est inversible, il est clair que cette fonction quantile coïncide avec l'inverse classique de F (avec les conventions évidentes aux limites). A contrario, considérons une variable aléatoire X discrète à valeurs dans l'ensemble fini $\{x_1 < \cdots < x_m\}$ avec probabilités $(p_1, \ldots, p_m)$. Il est facile de vérifier que pour tout $u \in]0,1[$,

$$F^{-1}(u) = \begin{cases} x_1 & \text{si } 0 < u \leq p_1 \\ x_2 & \text{si } p_1 < u \leq p_1 + p_2 \\ & \vdots \\ x_m & \text{si } p_1 + \cdots + p_{m-1} < u \leq 1 \end{cases}$$

c'est-à-dire

$$F^{-1}(u) = \sum_{k=1}^{m} x_k \mathbb{1}_{p_1+\cdots+p_{k-1} < u \leq p_1+\cdots+p_k}. \tag{3.2}$$

Si l'ensemble des valeurs prises par la variable discrète X n'est pas fini, il suffit de remplacer cette somme par une série. Quoi qu'il en soit, outre que, tout comme F, cette fonction quantile est croissante et en escalier, on notera que, contrairement à F, elle est continue à gauche. Ces propriétés sont en fait toujours vraies.

Convention : dans toute la suite, nous conviendrons que $F(-\infty) = 0$ et $F(+\infty) = 1$ afin de définir sans ambiguïté la fonction composée $F \circ F^{-1}$ sur $[0,1]$.

Propriétés 1
Soit F une fonction de répartition et F^{-1} son inverse généralisée. Alors :

1. *Valeur en 0 : $F^{-1}(0) = -\infty$.*
2. *Monotonie : F^{-1} est croissante.*
3. *Continuité : F^{-1} est continue à gauche.*
4. *Equivalence : $F(x) \geq u \iff x \geq F^{-1}(u)$.*
5. *Inversibilité : $\forall u \in [0,1]$, on a $(F \circ F^{-1})(u) \geq u$. De plus :*
   - *si F est continue mais pas injective, alors $F \circ F^{-1} = Id$, mais il existe u_0 tel que $(F^{-1} \circ F)(u_0) < u_0$;*
   - *si F est injective mais pas continue, alors $F^{-1} \circ F = Id$, mais il existe u_0 tel que $(F \circ F^{-1})(u_0) > u_0$;*
   - *il y a équivalence entre $F \circ F^{-1} = F^{-1} \circ F = Id$ et l'inversibilité de F au sens usuel.*

Exemples : illustrons le dernier point.

1. Si X suit une loi uniforme sur $[0,1]$, alors sa fonction de répartition F est continue mais pas injective. De fait, on a
$$(F^{-1} \circ F)(2) = F^{-1}(1) = 1 < 2.$$
2. Soit $Y \sim \mathcal{N}(0,1)$, $B \sim \mathcal{B}(1/2)$, avec Y et B indépendantes, et $X = 2BY$, alors la fonction de répartition de X présente un saut en 0 puisque $F(0^-) = 1/4$ tandis que $F(0) = 3/4$ (voir Figure 3.2). Elle est injective mais pas continue, et on voit que
$$(F \circ F^{-1})(1/2) = F(0) = \frac{3}{4} > \frac{1}{2}.$$

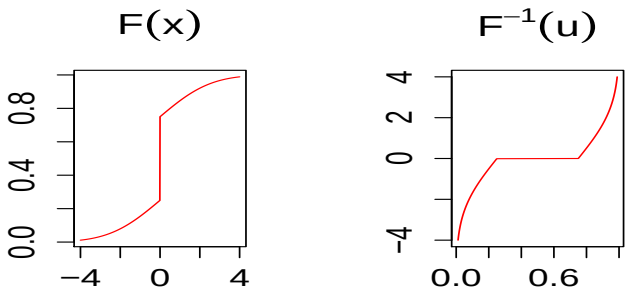

FIGURE 3.2 – Fonction de répartition et fonction de répartition empirique de $X = 2BY$.

Maintenant qu'on a défini l'inverse d'une fonction de répartition en toute généralité, on peut passer aux quantiles.

Définition 23 (Quantiles)
Soit F une fonction de répartition et p un réel de $[0,1]$. On appelle quantile d'ordre p, ou p-quantile, de F

$$x_p = x_p(F) = F^{-1}(p) = \inf\{x \in \mathbb{R} : \ F(x) \geq p\} \in \overline{\mathbb{R}}.$$

On le note aussi q_p (penser aux intervalles de confiance). $x_{1/2}$ est appelé médiane de F, $x_{1/4}$ et $x_{3/4}$ étant ses premier et troisième quartiles.

Remarque. On a toujours $x_0 = -\infty$, tandis que x_1 est la borne supérieure du support (éventuellement $+\infty$).

On peut ainsi définir les quantiles empiriques, lesquels se déduisent des statistiques d'ordre.

Lemme 2 (Quantiles empiriques)
Soit $(X_1, \ldots, X_n)$ un échantillon et F_n la fonction de répartition empirique associée. Pour tout $p \in [0,1]$, on note $x_p(n) = x_p(F_n)$ le quantile empirique (donc aléatoire) associé, c'est-à-dire

$$x_p(n) = F_n^{-1}(p) = \inf\{x \in \mathbb{R} : \ F_n(x) \geq p\}.$$

Avec la convention $X_{(0)} = -\infty$, le quantile $x_p(n)$ coïncide nécessairement avec l'une des statistiques d'ordre :

$$x_p(n) = X_{(i)} \iff \frac{i-1}{n} < p \leq \frac{i}{n} \iff np \leq i < np+1 \iff x_p(n) = X_{(\lceil np \rceil)}$$

où $\lceil x \rceil$ est la partie entière par excès de x, i.e. le plus petit entier supérieur ou égal à x.

Exemple. La médiane empirique dépend de la parité de n : $x_{1/2}(n) = X_{(n/2)}$ si n est pair et $x_{1/2}(n) = X_{((n+1)/2)}$ sinon.

Si $p \in]0,1[$ est fixé, il en va de même pour le p-quantile $x_p = F^{-1}(p)$, que l'on peut chercher à estimer. Disposant d'un échantillon $(X_1, \ldots, X_n)$ iid selon F, que dire du p-quantile empirique $x_p(n)$? Sans prendre de précautions, ça peut mal se passer...

Théorème 11 (Convergence et normalité asymptotique du quantile empirique)
Soit $(X_1, \ldots, X_n)$ iid selon F, $p \in]0,1[$ fixé, x_p le p-quantile de F et $x_p(n)$ le p-quantile empirique.

1. *Convergence : si F est strictement croissante en x_p, alors*

$$x_p(n) \xrightarrow[n\to\infty]{p.s.} x_p.$$

2. *Normalité asymptotique : si F est dérivable en x_p de dérivée $f(x_p) > 0$, alors*

$$\sqrt{n}(x_p(n) - x_p) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{p(1-p)}{f(x_p)^2}\right).$$

Preuve. Pour le premier point, fixons $p \in]0,1[$ et $\varepsilon > 0$. Comme très souvent pour montrer une convergence presque sûre, on va établir une inégalité de concentration du type

$$\mathbb{P}(|x_p(n) - x_p| \geq \varepsilon) \leq \alpha \exp(-\beta_{p,\varepsilon} n),$$

et Borel-Cantelli permettra de conclure. Vu la dissymétrie induite par l'inverse généralisée, on commence par scinder le terme à majorer :

$$\mathbb{P}(|x_p(n) - x_p| \geq \varepsilon) = \mathbb{P}(x_p(n) \leq x_p - \varepsilon) + \mathbb{P}(x_p(n) \geq x_p + \varepsilon). \tag{3.3}$$

Pour le premier, les égalités suivantes sont évidentes :

$$\mathbb{P}(x_p(n) \leq x_p - \varepsilon) = \mathbb{P}(X_{(\lceil np \rceil)} \leq x_p - \varepsilon) = \mathbb{P}(nF_n(x_p - \varepsilon) \geq \lceil np \rceil) = \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{X_i \leq x_p - \varepsilon} \geq \lceil np \rceil\right)$$

où l'on reconnaît une somme de variables de Bernoulli iid :

$$S_n = \sum_{i=1}^n B_i = \sum_{i=1}^n \mathbb{1}_{]-\infty, x_p - \varepsilon]}(X_i) \sim \mathcal{B}(n, F(x_p - \varepsilon)) \implies \mathbb{E}[S_n] = nF(x_p - \varepsilon).$$

Ainsi

$$\mathbb{P}(x_p(n) \leq x_p - \varepsilon) = \mathbb{P}(S_n - \mathbb{E}[S_n] \geq \lceil np \rceil - nF(x_p - \varepsilon)) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \geq n(p - F(x_p - \varepsilon))).$$

Or, par définition de $x_p = \inf\{x,\ F(x) \geq p\}$, on a pour tout $\varepsilon > 0$

$$F(x_p - \varepsilon) < p \implies n(p - F(x_p - \varepsilon)) =: n\delta > 0.$$

A ce stade, Hoeffding s'impose (cf. Chapitre 1 Proposition 4) :

$$\mathbb{P}(x_p(n) \leq x_p - \varepsilon) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \geq n\delta) \leq \exp(-2\delta^2 n),$$

terme général d'une série convergente. Le second terme de l'équation (3.3) se traite de façon comparable :

$$\mathbb{P}(x_p(n) \geq x_p + \varepsilon) = \mathbb{P}(X_{(\lceil np \rceil)} \geq x_p + \varepsilon) = \mathbb{P}(nF_n(x_p + \varepsilon) \leq \lceil np \rceil) = \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{X_i \leq x_p + \varepsilon} \leq \lceil np \rceil\right)$$

où l'on a cette fois

$$S_n = \sum_{i=1}^n \mathbb{1}_{]-\infty, x_p + \varepsilon]}(X_i) \sim \mathcal{B}(n, F(x_p + \varepsilon)) \implies \mathbb{E}[S_n] = nF(x_p + \varepsilon),$$

et

$$\mathbb{P}(x_p(n) \geq x_p + \varepsilon) = \mathbb{P}(S_n - \mathbb{E}[S_n] \leq \lceil np \rceil - nF(x_p + \varepsilon)) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \leq n(p - F(x_p + \varepsilon)) + 1).$$

Or, par hypothèse, on a pour tout $\varepsilon > 0$

$$F(x_p + \varepsilon) > p \implies n(p - F(x_p + \varepsilon)) + 1 < 0,$$

la dernière inégalité étant vraie pour tout $n \geq n_0 = n_0(p, \varepsilon)$. On peut donc à nouveau appliquer Hoeffding pour $n \geq n_0$:

$$\mathbb{P}(x_p(n) \geq x_p + \varepsilon) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \leq n(p - F(x_p + \varepsilon)) + 1) \leq \exp\left(-\frac{2(n(F(x_p + \varepsilon) - p) - 1)^2}{n}\right)$$

ce qui donne encore une série convergente. Le premier point est donc établi.

Le second revient à montrer que pour tout réel x

$$\mathbb{P}(\sqrt{n}(x_p(n) - x_p) \leq x) \xrightarrow[n\to\infty]{} \Phi\left(\frac{f(x_p)}{\sqrt{p(1-p)}}\, x\right),$$

où Φ représente comme d'habitude la fonction de répartition de la gaussienne centrée réduite. Soit donc $p \in]0, 1[$ et x_p le quantile associé. Puisque F est continue en x_p, on a $F(x_p) = p$. Soit maintenant x un réel fixé, alors

$$\mathbb{P}(\sqrt{n}(x_p(n) - x_p) \leq x) = \mathbb{P}\left(x_p(n) \leq x_p + \frac{x}{\sqrt{n}}\right) = \mathbb{P}\left(X_{(\lceil np \rceil)} \leq x_p + \frac{x}{\sqrt{n}}\right),$$

et en tenant compte du fait que les sauts de la fonction de répartition empirique sont d'amplitude au moins $1/n$, ceci s'écrit encore

$$\mathbb{P}(\sqrt{n}(x_p(n) - x_p) \leq x) = \mathbb{P}\left(nF_n(x_p + x/\sqrt{n}) \geq \lceil np \rceil\right) = \mathbb{P}\left(F_n(x_p + x/\sqrt{n}) > \frac{\lceil np \rceil - 1}{n}\right),$$

c'est-à-dire

$$\mathbb{P}(\sqrt{n}(x_p(n) - x_p) \leq x) = 1 - \mathbb{P}\left(F_n(x_p + x/\sqrt{n}) \leq \frac{\lceil np \rceil - 1}{n}\right) = 1 - G_n(y_n),$$

où G_n est la fonction de répartition de la variable aléatoire

$$Y_n = \sqrt{n}\left(F_n(x_p + x/\sqrt{n}) - F(x_p + x/\sqrt{n})\right)$$

et

$$y_n = \sqrt{n}\left(\frac{\lceil np \rceil - 1}{n} - F(x_p + x/\sqrt{n})\right).$$

Par définition de la partie entière par excès et d'après l'hypothèse sur F, il est clair que

$$y_n = \sqrt{n}\left(p + o(1/\sqrt{n}) - \left(F(x_p) + f(x_p)\frac{x}{\sqrt{n}} + o(1/\sqrt{n})\right)\right) \xrightarrow[n\to\infty]{} -f(x_p)x.$$

Concernant la variable Y_n, on a la décomposition $Y_n = Z_n + (Y_n - Z_n)$ avec

$$Z_n = \sqrt{n}(F_n(x_p) - F(x_p)) = \sqrt{n}(F_n(x_p) - p)$$

et la Proposition 15 implique que

$$Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, p(1-p)).$$

Par ailleurs,

$$Y_n - Z_n = \sqrt{n}(F_n(x_p + x/\sqrt{n}) - F_n(x_p)) - \sqrt{n}(F(x_p + x/\sqrt{n}) - F(x_p)),$$

or, comme on l'a vu à plusieurs reprises,

$$n(F_n(x_p + x/\sqrt{n}) - F_n(x_p)) = \sum_{i=1}^{n} \mathbb{1}_{x_p < X_i \leq x_p + x/\sqrt{n}} \sim \mathcal{B}(n, F(x_p + x/\sqrt{n}) - F(x_p)) =: \mathcal{B}(n, \delta_n).$$

Par l'inégalité de Tchebychev, il vient pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(|Y_n - Z_n| \geq c\right) \leq \frac{\delta_n(1-\delta_n)}{\varepsilon^2} \xrightarrow[n\to\infty]{} 0,$$

c'est-à-dire que $(Y_n - Z_n)$ tend en probabilité vers 0. Au total, par le Lemme de Slutsky,

$$Y_n = Z_n + (Y_n - Z_n) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, p(1-p)).$$

Autrement dit, la suite de fonctions de répartition (G_n) converge simplement vers la fonction de répartition de la loi $\mathcal{N}(0, p(1-p)$. D'après la Proposition 2, cette convergence est uniforme. En particulier,

$$\mathbb{P}(\sqrt{n}(x_p(n) - x_p) \leq x) = 1 - G_n(y_n) \xrightarrow[n\to\infty]{} 1 - \Phi\left(-\frac{f(x_p)}{\sqrt{p(1-p)}}\, x\right) = \Phi\left(\frac{f(x_p)}{\sqrt{p(1-p)}}\, x\right),$$

ce qui est le résultat voulu.

■

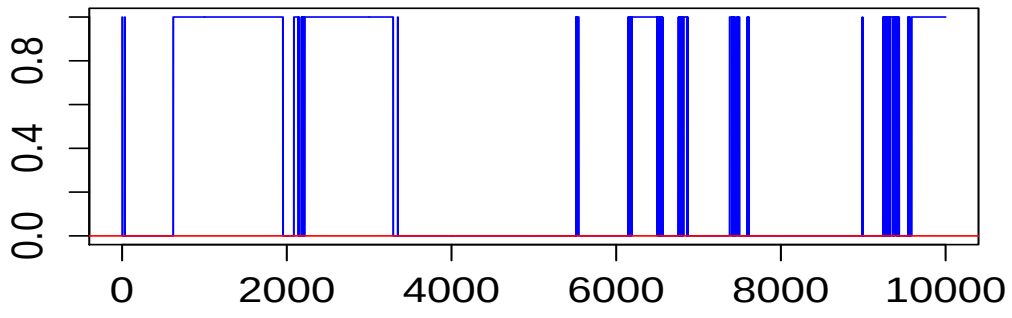

FIGURE 3.3 – Oscillation de la médiane empirique pour des variables de Bernoulli $\mathcal{B}(1/2)$.

Exemples :

1. Si x_p est le quantile d'ordre p de F, on a nécessairement $F(x) < F(x_p)$ si $x < x_p$. La condition de stricte croissance de F en x_p se ramène donc à la condition $F(x) > F(x_p)$ si $x > x_p$. Bref, il ne faut pas que la fonction de répartition soit plate à droite de x_p. Un exemple élémentaire permet de comprendre ce qui se passe : soit X distribué suivant une loi de Bernoulli de paramètre $1/2$. Sa médiane vaut donc 0. Il est néanmoins facile de se convaincre que la médiane empirique $x_{1/2}(n)$ va osciller éternellement (mais pas régulièrement) entre la valeur 0 et la valeur 1 (voir figure 3.3).

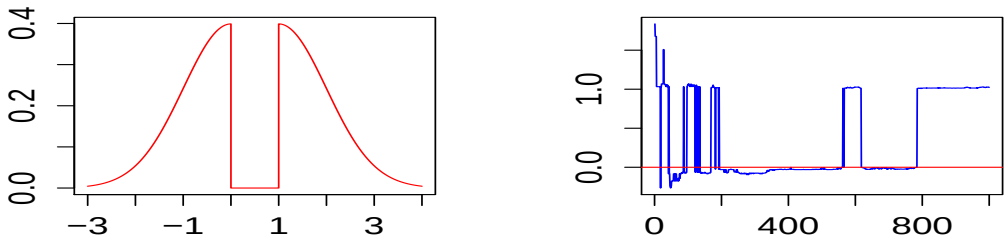

FIGURE 3.4 – Densité de $X = Y\mathbb{1}_{Y<0} + (1+Y)\mathbb{1}_{Y\geq 0}$ et oscillation de la médiane empirique.

2. Le comportement pathologique de la médiane empirique en exemple précédent n'est pas dû au fait que la loi de X est discrète. En effet, on peut très bien avoir le même type de phénomène lorsque X a une densité. Par exemple, soit $Y \sim \mathcal{N}(0,1)$ et la variable X définie comme suit :
$$X = Y\mathbb{1}_{Y<0} + (1+Y)\mathbb{1}_{Y\geq 0}.$$
La densité de X présente donc un trou entre 0 et 1, sa fonction de répartition un plateau sur cet intervalle, et sa médiane vaut $x_{1/2} = 0$ (voir Figure 3.4 à gauche). Ici encore, la médiane empirique $x_{1/2}(n)$ va osciller éternellement entre des valeurs négatives et des valeurs supérieures à 1 (voir Figure 3.4 à droite).
3. Pour comprendre la présence du $f(x_p)$ au dénominateur dans la variance asymptotique, voyons deux exemples. Dans le premier, on considère un mélange équiprobable de deux gaussiennes réduites de moyennes opposées, par exemple 10 et -10. Formellement, en notant X_1 et X_2 les variables gaussiennes en question et B une variable de Bernoulli de paramètre $1/2$, ceci s'écrit (voir Figure 3.5)
$$X = B \times X_1 + (1-B) \times X_2 \implies f(x) = \frac{1}{2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-10)^2}{2}} + \frac{1}{2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+10)^2}{2}}.$$
Par symétrie, la médiane de X est en 0, et par le premier point du théorème on est assuré de la convergence de $x_{1/2}(n)$ vers 0. Néanmoins, cette convergence est très lente : la plupart des points tombant près de l'un ou l'autre des modes, la médiane empirique sera elle-même très longtemps plus proche de l'un ou l'autre des modes que de 0. A contrario, si on considère une brave gaussienne centrée réduite, l'échantillon sera bien concentré autour de 0, donc si on coupe au milieu de celui-ci, la médiane empirique sera proche de 0.
4. On considère $(X_1, \ldots, X_n)$ iid selon la loi de Cauchy de densité
$$f(x) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)}.$$
Sa médiane est clairement le paramètre de translation θ, que l'on estime donc par la médiane empirique $x_{1/2}(n)$. Le résultat précédent nous assure que
$$x_{1/2}(n) \xrightarrow[n\to\infty]{\mathbb{P}} x_{1/2},$$
avec plus précisément
$$\sqrt{n}(x_{1/2}(n) - x_{1/2}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \pi^2/4\right).$$

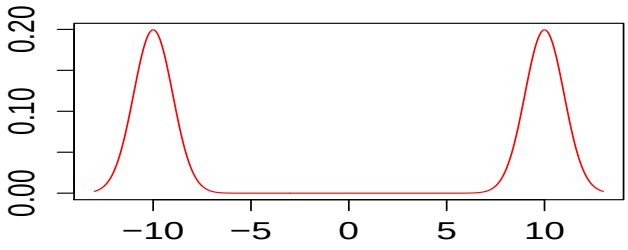

FIGURE 3.5 – Densité d'un mélange équiprobable de gaussiennes.

Remarque. Le résultat de normalité asymptotique du Théorème 11 ne sert généralement à rien si on veut construire des intervalles de confiance puisqu'il fait intervenir la densité f, le plus souvent inconnue. Dit autrement, la loi limite n'est pas pivotale. Alors que faire ?

Astuce : si l'on sait encadrer $F_n(x_p)$, alors il suffira "d'inverser" cet encadrement pour en déduire un intervalle de confiance pour x_p. Or, d'après la Proposition 15, si $F(x_p) = p$, c'est-à-dire si F est continue en x_p, on a

$$\sqrt{n}\,(F_n(x_p) - F(x_p)) = \sqrt{n}\,(F_n(x_p) - p) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, p(1-p)),$$

donc

$$\mathbb{P}\left(p - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n(x_p) < p + \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

On peut alors appliquer le point 4 des Propriétés 1 :

$$F_n(x) \geq u \iff x \geq F_n^{-1}(u) \qquad \text{et} \qquad F_n(x) < v \implies x \leq v$$

pour en déduire un intervalle de confiance de niveau asymptotique $(1-\alpha)$ pour x_p, à savoir :

$$\left[F_n^{-1}\left(p - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right), F_n^{-1}\left(p + \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\right].$$

Noter que cet intervalle s'obtient **très facilement** en pratique : si on définit p^+ et p^- par

$$p^{\pm} = p \pm \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}},$$

l'intervalle de confiance s'écrit tout simplement $[X_{(\lceil np^- \rceil)}, X_{(\lceil np^+ \rceil)}]$, et l'affaire est entendue.

Exemple. Lorsque F est continue en la médiane, un intervalle de confiance à 95% pour celle-ci est, à peu de choses près, complètement défini par les statistiques d'ordres $n/2 - \sqrt{n}$ et $n/2 + \sqrt{n}$. Autrement dit, si $n = 10^4$, il y a environ 95% de chances que la médiane se situe dans l'intervalle $[X_{(4900)}, X_{(5100)}]$.

3.3 Théorèmes limites

Nous avons vu en Proposition 15 que

$$F_n(x) \xrightarrow[n\to\infty]{p.s.} F(x).$$

Dans cette section, nous précisons ce point, d'abord en montrant que la convergence de F_n vers F a même lieu au sens de la norme infinie, ensuite en précisant la vitesse à laquelle cette convergence a lieu. Une idée clé pour prouver ces résultats est de se ramener à la loi uniforme grâce à la propriété suivante.

Lemme 3 (Universalité de la loi uniforme)
Soit U une variable uniforme sur $[0,1]$, F une fonction de répartition et F^{-1} son inverse généralisée. Alors :

1. *la variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F.*
2. *si X a pour fonction de répartition F et si F est continue, alors la variable aléatoire $F(X)$ est de loi uniforme sur $[0,1]$.*

Preuve. Soit $X = F^{-1}(U)$ et x réel fixé, alors d'après le résultat d'équivalence des Propriétés 1, la fonction de répartition de X se calcule facilement :

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

la dernière égalité venant de ce que, pour tout $u \in [0,1]$, $P(U \leq u) = u$. Le premier point est donc établi. On l'applique pour le second : la variable $Y = F^{-1}(U)$ a même loi que X, donc la variable $F(X)$ a même loi que $F(Y) = (F \circ F^{-1})(U)$. Or F est continue, donc par le dernier point des Propriétés 1, $F \circ F^{-1} = Id$, donc $F(Y) = U$ et $F(X)$ est de loi uniforme sur $[0,1]$.

■

A propos du second point, il est clair que si X présente un atome en x_0, la variable $F(X)$ va hériter d'un atome en $F(x_0)$, donc ne sera certainement pas distribuée selon une loi uniforme...

Application : méthode d'inversion en Monte-Carlo. Supposons que l'on dispose d'un générateur aléatoire de variables uniformes[2]. Par exemple, en R, une réalisation est donnée via la commande `u=runif(1)`. Alors, si la fonction de répartition F est facilement inversible, on déduit du résultat précédent une méthode basique pour générer d'une variable de fonction de répartition F.

Exemple : simulation d'une variable exponentielle. On veut générer une variable X selon la loi exponentielle de paramètre 1. Pour tout $x > 0$, $F(x) = 1 - e^{-x}$, il s'ensuit que pour tout $u \in]0,1[$, $F^{-1}(u) = -\log(1-u)$. Ainsi la commande `x=-log(1-runif(1))` donne une réalisation d'une variable exponentielle. Puisque U a la même loi que $1-U$, on peut même aller plus vite par `x=-log(runif(1))`. La fonction `rexp` de R est implémentée de cette façon.

3.3.1 Loi des grands nombres uniforme : Glivenko-Cantelli

Une application typique du Lemme 3 est donnée dès le début de la preuve du théorème suivant et aboutit à l'équation (3.4), qui montre que tout se ramène à l'étude d'un échantillon uniforme.

Théorème 12 (Glivenko-Cantelli)
Soit $(X_n)_{n\geq 1}$ des variables iid de fonction de répartition F, alors

$$\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{p.s.} 0.$$

2. c'est en fait un générateur pseudo-aléatoire.

Preuve. D'après le Lemme 3, si U est uniforme sur $[0,1]$, $F^{-1}(U)$ a même loi que X. Dès lors, considérant une suite (U_n) iid de variables uniformes sur $[0,1]$ et

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{F^{-1}(U_i)\leq x},$$

les **suites** de variables aléatoires $(\|F_n - F\|_\infty)$ et $(\|H_n - F\|_\infty)$ ont même loi. Il en découle l'équivalence suivante :

$$\|F_n - F\|_\infty \xrightarrow[n\to\infty]{p.s.} 0 \iff \|H_n - F\|_\infty \xrightarrow[n\to\infty]{p.s.} 0.$$

D'après le résultat d'équivalence des Propriétés 1, H_n s'écrit encore

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq F(x)}$$

et, puisque $0 \leq F(x) \leq 1$,

$$\|H_n - F\|_\infty = \sup_{x\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq F(x)} - F(x)\right| \leq \sup_{u\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq u} - u\right| =: \|G_n - G\|_\infty. \quad (3.4)$$

Il reste à montrer que ce majorant tend presque sûrement vers 0. Soit N un entier naturel fixé (non nul). On partitionne comme suit :

$$\|G_n - G\|_\infty = \sup_{0\leq j\leq N-1} \sup_{\frac{j}{N}\leq u\leq \frac{j+1}{N}} |G_n(u) - G(u)| = \sup_{0\leq j\leq N-1} \sup_{\frac{j}{N}\leq u\leq \frac{j+1}{N}} |G_n(u) - u|.$$

Puisque G_n est croissante, on a

$$\frac{j}{N} \leq u \leq \frac{j+1}{N} \implies G_n(j/N) - (j+1)/N \leq G_n(u) - u \leq G_n((j+1)/N) - j/N$$

donc, pour tout $u \in [0,1]$,

$$|G_n(u) - u| \leq \frac{1}{N} + \max_{0\leq j\leq N} |G_n(j/N) - j/N| = \frac{1}{N} + \max_{1\leq j\leq N-1} |G_n(j/N) - j/N|,$$

la dernière égalité tenant compte de $G_n(0) = 0$ et $G_n(1) = 1$. Bref, on s'est ramené à la majoration

$$\|G_n - G\|_\infty \leq \frac{1}{N} + \max_{1\leq j\leq N-1} |G_n(j/N) - j/N|.$$

D'après la Proposition 15 dans le cas uniforme, pour tout $j \in \{1,\dots,N-1\}$,

$$G_n(j/N) - j/N \xrightarrow[n\to\infty]{p.s.} 0,$$

donc, par intersection finie d'ensembles de probabilité 1, on a presque sûrement

$$\limsup_{n\to\infty} \|G_n - G\|_\infty \leq \frac{1}{N}.$$

Si Ω_N est l'ensemble de probabilité 1 sur lequel cette inégalité est vérifiée, il reste à prendre l'intersection des Ω_N, laquelle est encore de probabilité 1, pour aboutir au résultat souhaité. ■

Remarque. Les fonctions F_n étant croissantes, la preuve utilise bien sûr des arguments comparables à celle du deuxième théorème de Dini, mais il faut noter qu'ici on n'a même pas supposé la fonction limite F continue ! Il n'y a en fait aucune hypothèse sur celle-ci.

3.3.2 Vitesse uniforme : Kolmogorov-Smirnov et DKWM

Le théorème de Glivenko-Cantelli assure la convergence uniforme de F_n vers F presque sûrement (ne pas oublier que les fonctions F_n sont aléatoires !) :

$$\|F_n - F\|_\infty \xrightarrow[n\to\infty]{p.s.} 0.$$

En gros ceci signifie que, lorsque sa taille croît, l'échantillon $(X_1, \dots, X_n)$ permet de reconstruire la fonction F, donc la loi P_X, ce qui était bien l'objectif annoncé en introduction. On veut maintenant donner un équivalent du Théorème Central Limite, c'est-à-dire préciser la vitesse à laquelle cette convergence a lieu. Nous nous contenterons de donner deux résultats en ce sens, résultats que nous admettrons.

Mentionnons simplement qu'en équation (3.4), si on avait supposé F **continue**, cette inégalité devenait une égalité puisqu'alors $]0,1[\subseteq F(\mathbb{R})$, d'où

$$\|H_n - F\|_\infty = \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i \le F(x)} - F(x) \right| = \sup_{u\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i \le u} - u \right| = \|G_n - G\|_\infty.$$

Puisque $\|H_n - F\|_\infty$ a même loi que $\|F_n - F\|_\infty$, étudier la convergence en loi de $\sqrt{n}\|F_n - F\|_\infty$ revient à étudier celle de $\sqrt{n}\|G_n - G\|_\infty$. Cette idée est à l'œuvre dans la preuve du résultat suivant (admis).

Théorème 13 (Kolmogorov-Smirnov)
Soit $(X_n)_{n\ge1}$ des variables iid de fonction de répartition continue F, alors

$$\sqrt{n}\|F_n - F\|_\infty \xrightarrow[n\to\infty]{\mathcal{L}} K,$$

où la variable K a la loi dite de Kolmogorov-Smirnov, de fonction de répartition

$$F_K(x) = \mathbb{P}(K \le x) = \left(1 - 2\sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2x^2}\right) \mathbb{1}_{x>0}.$$

Une autre façon d'énoncer ce résultat est de dire que, pour tout $c > 0$,

$$\mathbb{P}\left(\sqrt{n}\|F_n - F\|_\infty \ge c\right) \xrightarrow[n\to\infty]{} 2\sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2c^2}.$$

Les quantiles de cette loi sont connus, par exemple $\mathbb{P}(K \ge 1.22) \approx 0.1$ et $\mathbb{P}(K \ge 1.36) \approx 0.05$. Par ailleurs, la fonction F_n étant constante sur les intervalles $]X_{(j-1)}, X_{(j)}[$ et la fonction F croissante et continue, la distance maximale entre F_n et F ne peut être atteinte qu'en l'un des X_j, c'est-à-dire

$$\|F_n - F\|_\infty = \max_{1\le j\le n} \left\{ \max\left(\left| F(X_{(j)}) - \frac{j-1}{n} \right|, \left| F(X_{(j)}) - \frac{j}{n} \right| \right) \right\}. \tag{3.5}$$

En pratique, ceci signifie que, étant donné l'échantillon $(X_1, \dots, X_n)$ et la fonction F, le calcul effectif de $\|F_n - F\|_\infty$ par logiciel est très rapide : il suffit d'ordonner l'échantillon et de prendre le maximum des $2n$ valeurs de la formule (3.5).

Application : test de Kolmogorov-Smirnov. Considérons un échantillon $(X_1, \dots, X_n)$, les X_i étant iid de fonction de répartition inconnue F, et une fonction de répartition continue donnée F_0. On veut tester

$$H_0 \ : \ F = F_0 \quad \text{contre} \quad H_1 \ : \ F \ne F_0.$$

Par l'inégalité triangulaire,

$$\|F - F_0\|_\infty \leq \|F - F_n\|_\infty + \|F_n - F_0\|_\infty.$$

Si H_0 n'est pas vraie, il en découle que, presque sûrement,

$$\liminf_{n\to\infty} \|F_n - F_0\|_\infty > 0 \implies \sqrt{n}\|F_n - F_0\|_\infty \xrightarrow[n\to\infty]{p.s.} +\infty.$$

Ainsi, supposons n "assez" grand et une réalisation $(x_1 = X_1(\omega), \ldots, x_n = X_n(\omega))$ de l'échantillon. si H_0 n'est pas vraie, la statistique de test $\sqrt{n}\|F_n(\omega) - F_0\|_\infty$ prendra des valeurs anormalement grandes par rapport à la loi de Kolmogorov-Smirnov. La procédure de test est donc naturelle : il suffit de fixer par exemple le niveau $\alpha = 5\%$, de calculer la statistique de test grâce à la formule (3.5) et de comparer au quantile de la loi de Kolmogorov-Smirnov :

$$K(\omega) := \sqrt{n}\|F_n(\omega) - F_0\|_\infty \lesseqgtr F_K^{-1}(1-\alpha) = 1.36$$

pour décider si l'on accepte ou rejette H_0. D'après le Théorème de Kolmogorov-Smirnov, ceci donne un test de niveau asymptotique α.

Si on regarde les preuves dans le détail, il ressort que nous avons fait deux hypothèses superflues dans la présentation de ce test. Tout d'abord, il reste de niveau asymptotique α même si on ne suppose pas F_0 continue (par contre, on n'a plus la convergence en loi vers une variable de Kolmogorov-Smirnov). Ensuite, on n'a pas besoin de recourir à l'asymptotique. En effet, on a vu que

$$\sqrt{n}\|F_n - F\|_\infty \overset{\mathcal{L}}{=} \sqrt{n}\|G_n - G\|_\infty =: K_n,$$

où la variable aléatoire K_n s'écrit donc

$$K_n = \sqrt{n}\|G_n - G\|_\infty = \sqrt{n} \sup_{u\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i \leq u} - u \right|.$$

Par conséquent, même dans un cadre non asympotique (e.g. $n < 50$), on peut appliquer exactement la même procédure de test : il suffit d'utiliser les quantiles de K_n et non ceux de K. Même si on n'a plus de forme explicite pour la fonction de répartition de K_n, rien n'empêche d'évaluer numériquement ses quantiles, typiquement par méthode Monte-Carlo. Par exemple, pour $n = 20$, on a $\mathbb{P}(K_{20} \geq 1.18) \approx 0.1$ et $\mathbb{P}(K_{20} \geq 1.31) \approx 0.05$.

Il existe une autre façon, encore plus simple, de construire un test non asymptotique de niveau α sans hypothèse de régularité sur F. Elle est basée sur l'inégalité suivante, aussi facile à énoncer que difficile à prouver.

Théorème 14 (Inégalité de Dvoretzky-Kiefer-Wolfowitz-Massart)
Soit $(X_1, \ldots, X_n)$ un échantillon de variables iid de fonction de répartition F, alors pour tout $c > 0$

$$\mathbb{P}\left(\sqrt{n}\|F_n - F\|_\infty \geq c\right) \leq 2e^{-2c^2}.$$

En 1956, Dvoretzky, Kiefer et Wolfowitz ont montré ce résultat, mais sans préciser la constante devant l'exponentielle. Dès 1958, Birnbaum et McCarty ont conjecturé que la constante optimale valait 2. Finalement, Massart l'a démontré en 1990.

Remarque : lien avec Hoeffding. Supposons x fixé et revenons à l'écriture de $F_n(x)$ comme somme de variables de Bernoulli :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{]-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^n Y_i.$$

Puisque les variables Y_i sont indépendantes et comprises entre 0 et 1, avec $\mathbb{E}[F_n(x)] = F(x)$, l'inégalité de Hoeffding donne

$$\mathbb{P}\left(\sqrt{n}|F_n(x) - F(x)| \geq c\right) \leq 2e^{-2c^2},$$

et ceci étant vrai pour tout réel x, il s'ensuit que

$$\sup_{x\in\mathbb{R}} \mathbb{P}\left(\sqrt{n}|F_n(x) - F(x)| \geq c\right) \leq 2e^{-2c^2}.$$

Le point remarquable de l'inégalité DKWM est que l'on peut en fait passer le supremum à l'intérieur de la probabilité sans changer le majorant !

Revenons au test précédent. Pour un niveau α préconisé, il suffit donc de considérer

$$c_\alpha = \sqrt{\frac{-\log(\alpha/2)}{2}}$$

et de procéder exactement comme avant, c'est-à-dire de comparer

$$\sqrt{n}\|F_n(\omega) - F_0\|_\infty \lesseqgtr c_\alpha$$

pour décider si l'on accepte ou rejette H_0.

Remarque : équivalence des tests. On commence par noter que si $\alpha = 10\%$ (respectivement $\alpha = 5\%$), alors $c_\alpha \approx 1.22$ (respectivement $c_\alpha \approx 1.36$), ce qui correspond justement aux valeurs approchées des quantiles d'ordre 0.9 et 0.95 donnés ci-dessus pour la loi de Kolmogorov-Smirnov. Rien d'étonnant à ça : pour $0 < \alpha < 1$, c_α est défini par $2\exp(-2c_\alpha^2) = \alpha$. La probabilité qu'une variable de Kolmogorov-Smirnov dépasse c_α est alors par définition

$$1 - F_K(c_\alpha) = 2\sum_{k=1}^{+\infty}(-1)^{k+1}e^{-2k^2c_\alpha^2} = 2e^{-2c_\alpha^2} - 2e^{-8c_\alpha^2} + \cdots = \alpha - 2e^{-8c_\alpha^2} + \ldots$$

Par le résultat classique sur les séries alternées, on a donc

$$0 < \alpha - (1 - F_K(c_\alpha)) \leq 2e^{-8c_\alpha^2} = \frac{\alpha^4}{8}.$$

Morale de l'histoire : pour les valeurs de α considérées en pratique, disons $\alpha \leq 10\%$, appliquer le test de Kolmogorov-Smirnov asymptotique ou celui basé sur l'inégalité DKWM ne fait aucune différence.

Chapitre 4

Estimation paramétrique unidimensionnelle

Introduction

Dans tout ce chapitre, on suppose disposer d'un échantillon $(X_1, \dots, X_n)$ de variables aléatoires iid selon une loi P_θ paramétrée par $\theta \in \Theta$, où Θ est un intervalle de $\mathbb{R}$. Autrement dit, nous sommes dans le cadre paramétrique le plus commode qui soit, le paramètre en jeu étant unidimensionnel. De fait, les outils mis en œuvre ici seront bien plus élémentaires que ceux du chapitre précédent.

4.1 Applications de la Delta méthode

Au Chapitre 1, Théorème 7, nous avons énoncé le principe de la Delta méthode dans un contexte très général : si une suite de variables aléatoires convenablement renormalisée converge en loi, alors l'image de cette suite par une fonction dérivable va elle-même converger en loi et on peut spécifier la limite.

En estimation paramétrique, le cadre d'application typique est le suivant : on veut estimer le paramètre θ, sachant qu'à partir des observations on sait construire facilement un estimateur d'une fonction de ce paramètre. Si la fonction en question est assez régulière, il suffit alors d'appliquer la Delta méthode à sa fonction réciproque.

En l'occurrence, une fonction "assez régulière" est un C^1-difféomorphisme, c'est-à-dire une application continûment dérivable, bijective, et dont la fonction réciproque est, elle aussi, continûment dérivable. Au passage, l'exemple $x \mapsto x^3$ montre qu'une fonction peut être bijective de $\mathbb{R}$ vers $\mathbb{R}$ et dérivable sans que sa réciproque soit dérivable partout.

Proposition 16 (Delta méthode et fonction inversible)
Soit $(X_1, \dots, X_n)$ un échantillon de variables aléatoires iid de loi P_θ, avec $\theta \in \Theta$ intervalle de $\mathbb{R}$, et φ un C^1-difféomorphisme de Θ dans $\varphi(\Theta)$. Si $\hat{\varphi}_n = \hat{\varphi}_n(X_1, \dots, X_n)$ est un estimateur convergent de $\varphi(\theta)$ et θ un point intérieur à Θ, alors $\hat{\theta}_n = \varphi^{-1}(\hat{\varphi}_n)$ est défini avec une probabilité qui tend vers 1 lorsque $n \to \infty$ et

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

De plus, s'il existe une suite de réels (v_n) tendant vers l'infini et une variable Z_θ tels que

$$v_n(\hat{\varphi}_n - \varphi(\theta)) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\theta,$$

alors

$$v_n(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{\varphi'(\theta)} Z_\theta.$$

Dans le cas particulier où $v_n = \sqrt{n}$ *et* $Z_\theta \sim \mathcal{N}(0, \sigma_\theta^2)$*, on a donc*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, (\sigma_\theta/\varphi'(\theta))^2).$$

Preuve. Le point θ étant intérieur à Θ et φ continue bijective, le point $\varphi(\theta)$ est intérieur à $\varphi(\Theta)$. Puisque $(\hat{\varphi}_n)$ converge en probabilité vers $\varphi(\theta)$, on en déduit que

$$\mathbb{P}\left(\hat{\varphi}_n \in \varphi(\Theta)\right) \xrightarrow[n\to\infty]{} 1.$$

Ainsi, la fonction φ^{-1} étant définie sur $\varphi(\Theta)$, l'estimateur $\hat{\theta}_n = \varphi^{-1}(\hat{\varphi}_n)$ est bien défini avec une probabilité qui tend vers 1 lorsque n tend vers l'infini. En outre, puisque φ^{-1} est continue, le Théorème 5, dit de Continuité, assure que $(\hat{\theta}_n) = (\varphi^{-1}(\hat{\varphi}_n))$ converge en probabilité vers θ. La Delta méthode appliquée à φ^{-1} donne alors le résultat voulu d'après la fameuse relation

$$(\varphi^{-1})'(y) = \frac{1}{(\varphi' \circ \varphi^{-1})(y)} \implies (\varphi^{-1})'(\varphi(\theta)) = \frac{1}{\varphi'(\theta)}.$$

■

La fonction φ^{-1} étant strictement monotone, si on connaît un intervalle de confiance pour $\hat{\varphi}_n$, alors θ en hérite. Si par exemple φ est croissante :

$$\mathbb{P}(\hat{\varphi}_n - a_n \leq \varphi(\theta) \leq \hat{\varphi}_n + b_n) \geq 1 - \alpha \implies \mathbb{P}(\varphi^{-1}(\hat{\varphi}_n - a_n) \leq \theta \leq \varphi^{-1}(\hat{\varphi}_n + b_n)) \geq 1 - \alpha.$$

Notation. Dans tout ce qui suit, nous noterons $q_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la gaussienne standard. Par exemple, pour des intervalles de confiance à 95%, $\alpha = 0.05$ et $q_{1-\alpha/2} = q_{0.975} \approx 1.96$.

4.1.1 La méthode des moments

Sous ce nom ne se cache rien de plus que le cas particulier où $\varphi(\theta)$ correspond à un moment de P_θ, c'est-à-dire que $\varphi(\theta) = \mathbb{E}_\theta[X^k]$ ou plus généralement $\varphi(\theta) = \mathbb{E}_\theta[h(X)]$. L'exemple le plus connu est celui où l'on estime $\varphi(\theta) = \mathbb{E}_\theta[X]$ par la moyenne empirique $\bar{X}_n$. Nous allons la décliner sur plusieurs exemples.

Lois uniformes

La loi uniforme est la loi du "hasard pur". Rappelons que X suit une loi uniforme sur $[a, b]$, où $-\infty < a < b < +\infty$, si elle a pour densité $f(x) = \mathbb{1}_{[a,b]}(x)/(b-a)$. Sa moyenne vaut $\mathbb{E}[X] = (a+b)/2$ et sa variance $\mathrm{Var}(X) = (b-a)^2/12$.

Considérons le modèle à un paramètre d'une loi uniforme sur $[\theta - 1, \theta + 1]$. On a donc $\mathbb{E}[X] = \theta$ et $\mathrm{Var}(X) = 1/3$. La moyenne empirique $\bar{X}_n$ est donc un estimateur sans biais de θ, son risque quadratique vaut $1/(3n)$ et on a la convergence en loi

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/3).$$

Si on veut des intervalles de confiance pour θ, on a au moins trois méthodes à notre disposition :

— Inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}(|\bar{X}_n - \theta| \geq c) \leq \frac{1}{3nc^2} \implies \mathbb{P}_\theta\left(\bar{X}_n - \frac{1}{\sqrt{3n\alpha}} \leq \theta \leq \bar{X}_n + \frac{1}{\sqrt{3n\alpha}}\right) \geq 1 - \alpha.$$

— Inégalité de Hoeffding :

$$\mathbb{P}(|\bar{X}_n-\theta|\geq c)\leq 2e^{-\frac{c^2n}{2}} \implies \mathbb{P}_\theta\left(\bar{X}_n-\sqrt{\frac{-2\log(\alpha/2)}{n}}\leq\theta\leq\bar{X}_n+\sqrt{\frac{-2\log(\alpha/2)}{n}}\right)\geq 1-\alpha.$$

Noter que l'inégalité de Hoeffding permet aussi de contruire des intervalles de confiance unilatères.

— Normalité asymptotique : on a cette fois des intervalles de confiance asymptotiques

$$\mathbb{P}_\theta\left(\bar{X}_n-\frac{q_{1-\alpha/2}}{\sqrt{3n}}\leq\theta\leq\bar{X}_n+\frac{q_{1-\alpha/2}}{\sqrt{3n}}\right)\xrightarrow[n\to\infty]{}1-\alpha,$$

et on peut construire là encore des intervalles de confiance asymptotiques unilatères.

Comme expliqué au Chapitre 1, on peut déduire de ces intervalles de confiance des tests d'hypothèses.

Lois exponentielles

La loi exponentielle correspond très souvent à la loi d'une durée. Rappelons que la variable X suit une loi exponentielle de paramètre $\lambda>0$, noté $X\sim\mathcal{E}(\lambda)$, si elle a pour densité $f(x)=\lambda e^{-\lambda x}\mathbb{1}_{x\geq 0}$. Sa moyenne vaut $\mathbb{E}[X]=1/\lambda$ et sa variance $\mathrm{Var}(X)=1/\lambda^2$. Le réel λ est un paramètre d'échelle : si $X\sim\mathcal{E}(\lambda)$, alors $Y=\lambda X\sim\mathcal{E}(1)$. Si on considère la moyenne empirique, on a donc

$$\bar{X}_n\xrightarrow[n\to\infty]{p.s.}\frac{1}{\lambda}\qquad\text{et}\qquad\sqrt{n}\left(\bar{X}_n-\frac{1}{\lambda}\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}(0,1/\lambda^2).$$

Si on considère l'estimateur $1/\bar{X}_n=g(\bar{X}_n)$, on sait par le Théorème de Continuité qu'il est convergent et la Delta méthode donne

$$\sqrt{n}\left(\frac{1}{\bar{X}_n}-\lambda\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}(0,\lambda^2).$$

Lois Gamma

En guise de mise en bouche, on rappelle que la fonction Gamma, définie pour tout réel $r>0$ par

$$\Gamma(r)=\int_0^{+\infty}x^{r-1}e^{-x}dx, \tag{4.1}$$

vérifie $\Gamma(1/2)=\sqrt{\pi}$, $\Gamma(1)=1$, $\Gamma(r+1)=r\Gamma(r)$ donc pour tout entier naturel n, $\Gamma(n+1)=n!$. Un changement de variable évident montre ainsi que, pour tout $\lambda>0$, la fonction

$$f(x)=f_{r,\lambda}(x)=\frac{(\lambda x)^{r-1}}{\Gamma(r)}\lambda e^{-\lambda x}\mathbb{1}_{x\geq 0}$$

définit une densité sur $\mathbb{R}^+$. Si la variable aléatoire X a cette densité, on dit que X suit une loi Gamma de paramètres r et λ et on note $X\sim\Gamma(r,\lambda)$.

Propriétés 1 (Loi Gamma)

1. *Lien avec la loi exponentielle :* $\Gamma(1,\lambda)=\mathcal{E}(\lambda)$.
2. *Changement d'échelle : si* $X\sim\Gamma(r,\lambda)$ *et si* $\alpha>0$, *alors* $\alpha X\sim\Gamma(r,\lambda/\alpha)$.
3. *Moments :* $\mathbb{E}[X]=r/\lambda$ *et* $\mathrm{Var}(X)=r/\lambda^2$.
4. *Lien avec la loi du chi-deux : si* $Y\sim\mathcal{N}(0,1)$, *alors* $Y^2\sim\Gamma(1/2,1/2)$, *donc* $\chi^2_1=\Gamma(1/2,1/2)$.

5. *Stabilité : si* $(X_1, \dots, X_n)$ *sont indépendantes de lois respectives* $\Gamma(r_i, \lambda)$*, alors*

$$X_1 + \cdots + X_n \sim \Gamma(r_1 + \cdots + r_n, \lambda).$$

Par conséquent :
— *Si* $(X_1, \dots, X_n)$ *sont iid de loi* $\mathcal{E}(\lambda)$*, alors*

$$\sum_{i=1}^{n} X_i \sim \Gamma(n, \lambda) \qquad et \qquad \bar{X}_n \sim \Gamma(n, n\lambda).$$

— *Si* $(X_1, \dots, X_n)$ *sont iid de loi* $\mathcal{N}(0,1)$*, alors* $\sum_{i=1}^{n} X_i^2 \sim \Gamma(n/2, 1/2)$*, c'est-à-dire que* $\chi^2_n = \Gamma(n/2, 1/2)$.

Lorsque r est grand, la loi $\Gamma(r, \lambda)$ ressemble à une loi normale. Par abus de notation, on écrira parfois "$\Gamma(r, \lambda) \overset{\mathcal{L}}{\approx} \mathcal{N}(r/\lambda, r/\lambda^2)$", en ayant bien conscience de ce que cela signifie, à savoir

$$\frac{\lambda}{\sqrt{r}}\left(X_r - \frac{r}{\lambda}\right) \xrightarrow[r\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \iff \forall x \in \mathbb{R},\ \left|\mathbb{P}\left(\frac{\lambda}{\sqrt{r}}\left(X_r - \frac{r}{\lambda}\right) \le x\right) - \Phi(x)\right| \xrightarrow[r\to\infty]{} 0.$$

Pour l'estimation de paramètres, partant d'un échantillon $(X_1, \dots, X_n)$ iid selon une loi $\Gamma(r, \lambda)$, la moyenne empirique a les propriétés suivantes : $\mathbb{E}[\bar{X}_n] = r/\lambda$, $\mathrm{Var}(\bar{X}_n) = r/(\lambda^2 n)$, donc

$$\sqrt{n}\left(\bar{X}_n - \frac{r}{\lambda}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, r/\lambda^2) \iff \sqrt{\frac{n}{r}}\left(\lambda \bar{X}_n - r\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Supposons que r est connu et que l'on cherche à estimer λ. Un intervalle de confiance asymptotique se déduit donc de la convergence

$$\mathbb{P}\left(\frac{1}{\bar{X}_n}\left(r - \frac{q_{1-\alpha/2}\sqrt{r}}{\sqrt{n}}\right) \le \lambda \le \frac{1}{\bar{X}_n}\left(r + \frac{q_{1-\alpha/2}\sqrt{r}}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} 1 - \alpha.$$

On peut aussi appliquer Tchebychev pour un intervalle non asymptotique. Notons qu'en prenant $r = 1$, tout ceci s'applique en particulier au cas d'une loi exponentielle de paramètre inconnu λ.

Si, réciproquement, λ est connu et que l'on cherche à estimer r, on sait d'une part que $\lambda \bar{X}_n$ est un estimateur convergent de r, d'autre part grâce à la normalité asymptotique ci-dessus et le Théorème de Slutsky que

$$\sqrt{n}\,\frac{\lambda \bar{X}_n - r}{\sqrt{\lambda \bar{X}_n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

ce qui fournit des intervalles de confiance asymptotiques pour r. Là encore, Tchebychev permet d'obtenir des intervalles non asymptotiques, au prix de la résolution d'équations du second degré.

Translation et changement d'échelle

A partir d'une densité f sur $\mathbb{R}$ et considérant un couple $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*$, on peut définir une nouvelle densité $f_{\mu,\sigma}$ par translation et changement d'échelle comme suit :

$$\forall y \in \mathbb{R} \qquad f_{\mu,\sigma}(y) = \frac{1}{\sigma} f((y - \mu)/\sigma).$$

Si X a pour densité $f = f_{0,1}$, la variable aléatoire $Y = \sigma X + \mu$ a pour densité $f_{\mu,\sigma}$. On en trouve des exemples à foison dans la littérature. L'exemple le plus courant est celui où $X \sim \mathcal{N}(0,1)$, auquel cas $Y = \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$. On peut encore citer le cas où $X \sim \mathcal{U}_{[0,1]}$ et $Y = (b-a)X + a \sim \mathcal{U}_{[a,b]}$.

Dans un contexte de statistique inférentielle, supposons que l'on connaisse $\mathbb{E}[X] = m$, $\mathrm{Var}(X) = s^2$ et qu'à partir d'un échantillon $(Y_1, \dots, Y_n)$ iid selon la densité $f_{\mu,\sigma}$, on veuille estimer μ ou σ. On commence par noter que

$$\mathbb{E}[Y] = \sigma m + \mu \qquad \text{et} \qquad \mathrm{Var}(Y) = s^2\sigma^2.$$

Si σ est connu et que l'on veut estimer μ, on propose donc l'estimateur

$$\hat{\mu}_n = \bar{Y}_n - \sigma m = \frac{1}{n}\sum_{i=1}^{n} Y_i - \sigma m.$$

Par les théorèmes classiques, cet estimateur est non biaisé, consistant et obéit à la normalité asymptotique

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2 s^2),$$

ce qui permet de construire des intervalles de confiance asymptotiques. A nouveau, les inégalités de Tchebychev et Hoeffding (dans le cas borné) fournissent des intervalles de confiance non asymptotiques.

Si μ est connu et que l'on veut estimer σ, distinguons deux cas de figure possibles :

— si $m \neq 0$: l'estimateur naturel est alors

$$\hat{\sigma}_n = \frac{1}{m}(\bar{Y}_n - \mu),$$

qui est consistant et vérifie

$$\sqrt{n}(\hat{\sigma}_n - \sigma) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, (\sigma s/m)^2) \iff \sqrt{n}\,\frac{m}{s}\left(\frac{\hat{\sigma}_n}{\sigma} - 1\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

d'où l'on déduit des intervalles de confiance asymptotiques.

— Si $m = 0$, il faut aller à l'ordre 2 : puisque $\mathrm{Var}(Y) = \mathbb{E}[(Y-\mu)^2] = s^2\sigma^2$, l'estimateur est cette fois

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{s}\right)^2,$$

lequel est bien convergent par la loi des grands nombres. Si on suppose de plus l'existence d'un moment d'ordre 4 pour Y (ou, ce qui est équivalent, pour X), alors

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^4 \mathrm{Var}(X^2)/s^4) \iff \sqrt{n}\,\frac{s^2}{\sqrt{\mathrm{Var}(X^2)}}\left(\frac{\hat{\sigma}_n^2}{\sigma^2} - 1\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

et on peut à nouveau obtenir des intervalles de confiance asymptotiques.

4.1.2 Utilisation des quantiles empiriques

Nous avons vu au Chapitre 3 des résultats de consistance et de normalité asymptotique pour le quantile empirique et l'avons illustré sur l'exemple de la médiane d'une loi de Cauchy. Lorsque médiane et moyenne coïncident, on dispose donc de deux estimateurs de celle-ci, moyenne et médiane empiriques, que l'on peut chercher à comparer.

Exemple. Supposons $(X_1, \dots, X_n)$ iid selon la loi normale $\mathcal{N}(\theta, 1)$, alors par le TCL

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

tandis qu'en notant $x_{1/2}(n)$ la médiane empirique, on a

$$\sqrt{n}(x_{1/2}(n) - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \pi/2).$$

Sur ce cas particulier, la médiane empirique correspond donc à un estimateur un peu moins précis que la moyenne empirique. Notons que ça n'est pas toujours le cas, il suffit pour s'en convaincre de considérer une loi de Laplace : l'estimateur de la médiane empirique est asymptotiquement $\sqrt{2}$ fois plus précis que celui de la moyenne empirique.

Même lorsque, comme dans le cas gaussien, l'estimateur de la médiane empirique est théoriquement moins bon, cet estimateur peut être intéressant en raison de sa robustesse. Un exemple très simple permet de comprendre l'idée.

Exemple : donnée aberrante. Supposons $\theta = 0$ dans l'exemple précédent, c'est-à-dire les X_i normales centrées réduites. On dispose de 100 observations, les 99 premières suivant la loi prescrite, tandis que la dernière, pour une raison ou une autre (erreur de manipulation, etc.), est aberrante et vaut 50. Alors, sachant que $X_{100} = 50$, on a pour la moyenne empirique

$$\bar{X}_n = \frac{1}{100}\sum_{i=1}^{99} X_i + \frac{1}{2} \sim \mathcal{N}(1/2, 99/10^4).$$

L'écart-type valant à peu près 1/10, il y a environ 95% de chances que $\bar{X}_n$ se trouve entre 0.3 et 0.7, tandis qu'en l'absence de valeur aberrante, celle-ci se trouverait entre -0.2 et 0.2, d'où le problème : une seule valeur erronée a fait dérailler l'estimateur... A contrario, il est clair que celle-ci n'a quasiment aucune influence sur la médiane empirique. Ainsi la médiane empirique est-elle beaucoup plus stable que la moyenne empirique face aux données aberrantes : on dit qu'elle est **robuste**.

Rappel ! Revenons sur la médiane empirique dans un cadre général. Comme expliqué au chapitre précédent, le résultat de normalité asymptotique

$$\sqrt{n}(x_{1/2}(n) - x_{1/2}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{4f(x_{1/2})^2}\right)$$

est inemployable pour la construction d'intervalles de confiance si on ne connaît pas $f(x_{1/2})$, ce qui est très souvent le cas. Mais on s'en sort quand même grâce à la ruse du passage par $F_n(x_{1/2})$, ce qui donne l'intervalle de confiance asymptotique à 95% (en arrondissant 1.96 à 2) :

$$[X_{(\lceil n/2-\sqrt{n}\rceil)}, X_{(\lceil n/2+\sqrt{n}\rceil)}].$$

4.1.3 Stabilisation de la variance

Lorsqu'on estime un paramètre θ par $\hat{\theta}_n$, un résultat de normalité asymptotique prend typiquement la forme

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(\theta)\right), \tag{4.2}$$

où la variance asymptotique est donc une fonction du paramètre θ. L'idée basique est de remplacer $\sigma^2(\theta)$ par $\sigma^2(\hat{\theta}_n)$. En anglais, cette méthode est connue sous le nom de *plug-in principle*. Si la fonction $\theta \mapsto \sigma(\theta)$ est continue, elle est justifiée par le résultat de Slutsky puisqu'alors

$$\sqrt{n}\,\frac{\hat{\theta}_n - \theta}{\sigma(\hat{\theta}_n)} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

d'ou des intervalles de confiance asymptotiques :

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \leq \frac{q_{1-\alpha/2}\sigma(\hat{\theta}_n)}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Il existe une autre façon de procéder, dite méthode de stabilisation de la variance. Partant de (4.2), si φ est une fonction dérivable, la Delta méthode nous dit que

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

Par conséquent, la variance asymptotique ne dépend plus de θ si la fonction φ est solution de l'équation différentielle $\varphi'(\theta) = 1/\sigma(\theta)$. Puisque $\sigma(\theta) > 0$, ceci implique que φ, en plus d'être continue, est strictement croissante, donc inversible. Puisque

$$\sqrt{n}(\varphi(\hat{\theta}_n) - \varphi(\theta)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

il vient

$$\mathbb{P}\left(\left|\varphi(\hat{\theta}_n) - \varphi(\theta)\right| \leq \frac{q_{1-\alpha/2}}{\sqrt{n}}\right) = \mathbb{P}\left(\varphi^{-1}\left(\varphi(\hat{\theta}_n) - \frac{q_{1-\alpha/2}}{\sqrt{n}}\right) \leq \theta \leq \varphi^{-1}\left(\varphi(\hat{\theta}_n) + \frac{q_{1-\alpha/2}}{\sqrt{n}}\right)\right),$$

qui correspond donc à un intervalle de confiance asymptotique de niveau $(1-\alpha)$. Toute la question se ramène alors à la résolution de l'équation différentielle $\varphi'(\theta) = 1/\sigma(\theta)$.

Exemple : loi de Poisson. Soit $\lambda > 0$ fixé, on dit que la variable aléatoire X suit une loi de Poisson de paramètre λ, noté $X \sim \mathcal{P}(\lambda)$, si elle est à valeurs dans $\mathbb{N}$, avec

$$\forall n \in \mathbb{N} \qquad \mathbb{P}(X = n) = e^{-\lambda}\,\frac{\lambda^n}{n!}.$$

En modélisation, la loi de Poisson sert à modéliser un nombre d'événements, et plus particulièrement des **événements rares**. Ceci s'explique par le lien bien connu avec la loi binomiale : soit (p_n) une suite de réels entre 0 et 1 telle que $\lim_{n\to\infty} np_n = \lambda > 0$ fixé, et (X_n) des variables suivant des lois binomiales $\mathcal{B}(n, p_n)$, alors il n'est pas difficile de montrer que

$$\forall k \in \mathbb{N} \qquad \mathbb{P}(X_n = k) \xrightarrow[n\to\infty]{} e^{-\lambda}\,\frac{\lambda^k}{k!}.$$

Autrement dit, une loi binomiale $\mathcal{B}(n,p)$ avec n grand et p petit (disons $n > 50$ et $np < 5$) "ressemble" à une loi de Poisson $\mathcal{P}(np)$. Lorsqu'on fait un très grand nombre d'essais dont chacun a une probabilité très faible de se produire, le nombre final de succès suit approximativement une loi de Poisson.

Les premiers moments d'une loi de Poisson se retiennent sans trop forcer puisque si $X \sim \mathcal{P}(\lambda)$, alors $\mathbb{E}[X] = \mathrm{Var}(X) = \lambda$. Par ailleurs, les lois de Poisson sont stables par convolution : si $X_1 \sim \mathcal{P}(\lambda_1)$, $X_2 \sim \mathcal{P}(\lambda_2)$ avec X_1 indépendante de X_2, alors $X_1 + X_2 \sim \mathcal{P}(\lambda_1 + \lambda_2)$. De ceci découle en particulier le fait que, lorsque λ est grand, une loi de Poisson ressemble à une loi normale, ce qui s'écrit abusivement "$\mathcal{P}(\lambda) \overset{\mathcal{L}}{\approx} \mathcal{N}(\lambda, \lambda)$" et rigoureusement :

$$\frac{X_\lambda - \lambda}{\sqrt{\lambda}} \xrightarrow[\lambda\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \iff \forall x \in \mathbb{R},\ \left|\mathbb{P}\left(\frac{X_\lambda - \lambda}{\sqrt{\lambda}} \leq x\right) - \Phi(x)\right| \xrightarrow[r\to\infty]{} 0.$$

Si on a un échantillon $(X_1, \ldots, X_n)$ de variables iid suivant une loi de Poisson de paramètre λ inconnu, la moyenne empirique $\bar{X}_n$ est un estimateur sans biais, convergent et

$$\sqrt{n}(\bar{X}_n - \lambda) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \lambda). \tag{4.3}$$

Appliquons la méthode de stabilisation de la variance : on cherche φ telle que

$$\varphi'(\lambda) = \frac{1}{\sqrt{\lambda}} \iff \varphi(\lambda) = 2\sqrt{\lambda},$$

ce qui donne donc

$$2\sqrt{n}(\sqrt{\bar{X}_n} - \sqrt{\lambda}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

d'où les intervalles de confiance asymptotiques :

$$\mathbb{P}\left(\left(\sqrt{\bar{X}_n} - \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right)^2 \leq \lambda \leq \left(\sqrt{\bar{X}_n} - \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right)^2\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Remarque. Le principe du plug-in appliqué à (4.3) donne

$$\sqrt{n}\,\frac{\bar{X}_n - \lambda}{\sqrt{\bar{X}_n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

donc

$$\mathbb{P}\left(\bar{X}_n - \frac{q_{1-\alpha/2}\sqrt{\bar{X}_n}}{\sqrt{n}} \leq \lambda \leq \bar{X}_n + \frac{q_{1-\alpha/2}\sqrt{\bar{X}_n}}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

La méthode de stabilisation de la variance a donc eu pour effet de translater les intervalles de confiance de $q^2_{1-\alpha/2}/(4n)$. Lorsque $1/n$ devient négligeable devant $\sqrt{\lambda/n}$, il n'y a donc plus de différence.

4.2 Le maximum de vraisemblance

4.2.1 Principe et notations

On revient ici au cadre général d'un modèle statistique $(P_\theta)_{\theta\in\Theta}$ dominé par une mesure ν et on note, pour tout $\theta \in \Theta$, $g_\theta = dP_\theta/d\nu$ la densité correspondante. Etant donné une observation $\mathbf{X} = (X_1, \ldots, X_n)$, on peut donc calculer $L_n(\theta) = g_\theta(\mathbf{X})$ et, avec la convention usuelle $\log 0 = -\infty$,

$$\ell_n(\theta) = \log L_n(\theta) = \log g_\theta(\mathbf{X}),$$

respectivement appelées vraisemblance et log-vraisemblance associées à θ.

Définition 24 (Maximum de vraisemblance)
Avec les notations précédentes, un estimateur du maximum de vraisemblance (EMV) de θ est, sous réserve d'existence, un élément $\hat{\theta}$ de Θ qui vérifie

$$L_n(\hat{\theta}) = \sup_{\theta\in\Theta} L_n(\theta) \iff \ell_n(\hat{\theta}) = \sup_{\theta\in\Theta} \ell_n(\theta).$$

Dans le cas où $\mathbf{X} = (X_1, \ldots, X_n)$ avec les X_i iid de densité f_θ, on a donc

$$\ell_n(\hat{\theta}) = \sup_{\theta\in\Theta} \sum_{i=1}^n \log f_\theta(X_i).$$

Interprétation : sous réserve d'existence et d'unicité, l'EMV $\hat{\theta}$ est donc la valeur de θ qui rend le jeu d'observations $X_1, \ldots, X_n$ le plus **vraisemblable**. Dès lors, il est logique que $\hat{\theta}$ soit une variable aléatoire dépendant des X_i.

Lorsque Θ est fini, le modèle identifiable et les X_i iid, on peut montrer qu'il existe un EMV et qu'il est asymptotiquement unique et convergent. Mais, en général, ni l'existence ni l'unicité des EMV ne sont assurées. En fait, à peu près tout peut arriver, comme on pourra s'en rendre compte sur quelques exemples en section suivante.

Supposons que, partant du paramétrage par $\theta \in \Theta$, on considère une bijection $\varphi : \Theta \to \Lambda$. Il est alors équivalent de travailler avec les densités $(g_\theta)_{\theta\in\Theta}$ ou avec les densités $(h_\lambda)_{\lambda\in\Lambda}$ définies par $h_\lambda(\mathbf{x}) = g_{\varphi^{-1}(\lambda)}(\mathbf{x})$. Sous réserve d'existence, un EMV $\hat{\lambda}$ du second paramétrage vérifie alors

$$h_{\hat{\lambda}}(\mathbf{X}) = \sup_{\lambda\in\Lambda} h_\lambda(\mathbf{X}) = \sup_{\lambda\in\Lambda} g_{\varphi^{-1}(\lambda)}(\mathbf{X}) = \sup_{\theta\in\Theta} g_\theta(\mathbf{X}) = g_{\hat{\theta}}(\mathbf{X}),$$

donc il y a correspondance bijective entre EMV pour les deux paramétrages. Il est ainsi équivalent de dire que $\hat{\theta}$ est un EMV de θ ou que $\hat{\lambda} = \varphi(\hat{\theta})$ est un EMV de $\lambda = \varphi(\theta)$. Par convention, on étend ce principe au cas où φ n'est pas bijective.

Définition 25 (Extension de la notion d'EMV)
Si φ est une application définie sur Θ, on dit que $\varphi(\hat{\theta})$ est un estimateur du maximum de vraisemblance de $\varphi(\theta)$ si $\hat{\theta}$ est un estimateur du maximum de vraisemblance de θ.

Exemple. Considérons un modèle gaussien où les variables X_i sont iid de loi $\mathcal{N}(\theta, 1)$. La log-vraisemblance s'écrit (voir aussi Figure 4.1)

$$\ell_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2.$$

On vérifie sans problème que l'unique maximum de cette fonction est en $\hat{\theta} = \bar{X}_n$. L'EMV coïncide donc avec la moyenne empirique. Avec la convention de la définition précédente, nous dirons donc que l'EMV de θ^2 dans ce modèle est $(\bar{X}_n)^2$.

4.2.2 Exemples

Nous présentons maintenant quelques exemples illustrant les différents cas de figures.

Modèle gaussien

On étend l'exemple précédent au cas où $X_i \sim \mathcal{N}(\mu, \sigma^2)$. La log-vraisemblance s'écrit cette fois comme une fonction de deux variables :

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

Si σ est connu, tout se passe comme ci-dessus et l'EMV de μ est $\hat{\mu} = \bar{X}_n$. Si μ est connu et si on cherche l'EMV de σ^2, la dérivation par rapport à σ^2 (et non par rapport à σ !) donne

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Ainsi, dans les deux cas, les EMV correspondent aux estimateurs obtenus par la méthode des moments. Notons que la maximisation de $\ell_n(\mu, \sigma^2)$ par rapport à μ ne dépend pas de la valeur de σ^2 : c'est toujours $\hat{\mu} = \bar{X}_n$. Donc, si les deux paramètres sont inconnus, l'EMV de σ^2 doit maximiser

$$\ell_n(\bar{X}_n, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

qui correspond à la variance empirique.

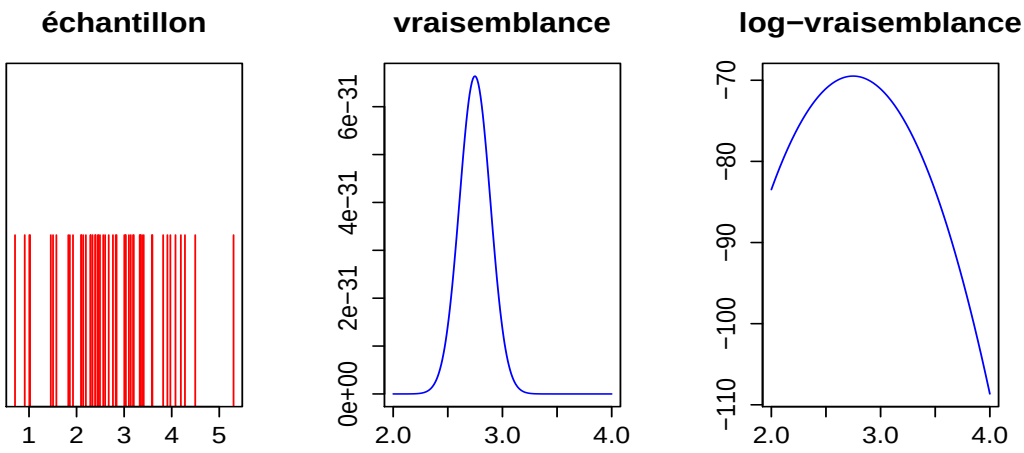

FIGURE 4.1 – Echantillon de 50 variables gaussiennes $\mathcal{N}(3,1)$, vraisemblance et log-vraisemblance.

Loi de Poisson

On passe maintenant à un exemple discret. Si $X \sim \mathcal{P}(\lambda)$, avec $\lambda > 0$, alors $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$ pour tout entier naturel k. La densité de la loi de Poisson par rapport à la mesure de comptage sur $\mathbb{N}$ est ainsi définie par $f_\lambda(x) = e^{-\lambda}\lambda^x/x!$ pour tout entier naturel x. Un échantillon iid $(X_1, \ldots, X_n)$ étant donné, sa log-vraisemblance vaut donc, après quelques bidouillages,

$$\ell_n(\lambda) = -n(\lambda - \bar{X}_n \log \lambda) - \sum_{i=1}^{n} \log(X_i!),$$

laquelle se minimise sans difficulté et aboutit à l'EMV $\hat{\lambda} = \bar{X}_n$ si $\bar{X}_n > 0$. Le cas pathologique où la moyenne empirique est nulle correspond à la nullité de tous les X_i. Dans ce cas $\ell_n(\lambda) = -n\lambda$, qui n'a pas de maximum, la valeur $\lambda = 0$ étant exclue pour une loi de Poisson. Notons cependant que ceci n'arrive qu'avec probabilité $\exp(-n\lambda)$, qui tend exponentiellement vite vers 0 avec n.

Loi uniforme sur $[0, \theta]$

La densité étant égale à $f_\theta(x) = \mathbb{1}_{[0,\theta]}(x)/\theta$, la vraisemblance vaut

$$L_n(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(X_i) = \frac{1}{\theta^n} \mathbb{1}_{[X_{(n)},+\infty[}(\theta),$$

où $X_{(n)} = \max(X_1, \ldots, X_n)$ est la statistique d'ordre n. La maximisation se voit tout de suite : il faut garder l'indicatrice égale à 1 et minimiser θ^n, d'où l'EMV $\hat{\theta} = X_{(n)}$.

On peut dire beaucoup de choses sur cet estimateur, puisque sa fonction de répartition est tout simplement $F_{\hat{\theta}}(t) = \mathbb{P}(X_{(n)} \leq t) = (t/\theta)^n$ pour tout $t \in [0, \theta]$, d'où sa densité et son espérance :

$$f_{\hat{\theta}}(t) = \frac{n}{\theta^n} t^{n-1} \mathbb{1}_{[0,\theta]}(t) \implies \mathbb{E}[\hat{\theta}] = \frac{n}{n+1}\theta,$$

ce qui prouve qu'il est biaisé (biais en $\mathcal{O}(1/n)$). Le moment d'ordre 2 permet de calculer le risque quadratique :

$$\mathbb{E}[\hat{\theta}^2] = \frac{n}{n+2}\theta^2 \implies R(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \frac{2\theta^2}{(n+1)(n+2)}.$$

Grâce à la fonction de répartition, on note que, pour tout $\alpha \in]0,1[$,

$$\mathbb{P}(\hat{\theta} \leq \alpha^{1/n}\theta) \leq \alpha \implies \mathbb{P}(\hat{\theta} \leq \theta \leq \alpha^{-1/n}\hat{\theta}) = 1-\alpha,$$

ce qui fournit un intervalle de confiance (non asymptotique !) de niveau $(1-\alpha)$.

Puisque $\mathbb{E}[\bar{X}_n] = \theta/2$, un estimateur basé sur la méthode des moments serait $\tilde{\theta} = 2\bar{X}_n$, lequel est nettement moins bon en terme de risque quadratique puisque

$$R(\tilde{\theta}, \theta) = \mathrm{Var}(2\bar{X}_n) = \frac{\theta^2}{3n},$$

et ce bien que l'EMV soit biaisé...

Loi uniforme sur $[\theta-1, \theta+1]$

Cette fois la vraisemblance s'écrit

$$L_n(\theta) = \frac{1}{2^n}\prod_{i=1}^{n} \mathbb{1}_{[\theta-1,\theta+1]}(X_i) = \frac{1}{2^n}\mathbb{1}_{[X_{(n)}-1, X_{(1)}+1]}(\theta).$$

Elle ne prend que deux valeurs, 0 et $1/2^n$, de sorte que tout $\theta \in [X_{(n)}-1, X_{(1)}+1]$ est un EMV [1] C'est donc une situation où il n'y a pas unicité de l'EMV. En calculant les fonctions de répartition de $X_{(1)}$ et $X_{(n)}$ à l'instar de ce qui a été fait dans l'exemple précédent, on montre facilement que $X_{(1)}$ tend vers $(\theta-1)$ et $X_{(n)}$ vers $(\theta+1)$. Par conséquent, quel que soit le choix de $\hat{\theta}_n$ dans l'intervalle $[X_{(n)}-1, X_{(1)}+1]$, on aura convergence vers θ. Une possibilité est de couper la poire en deux en choisissant le milieu de l'intervalle, i.e. $\hat{\theta}_n = (X_{(1)} + X_{(n)})/2$.

Loi de Cauchy

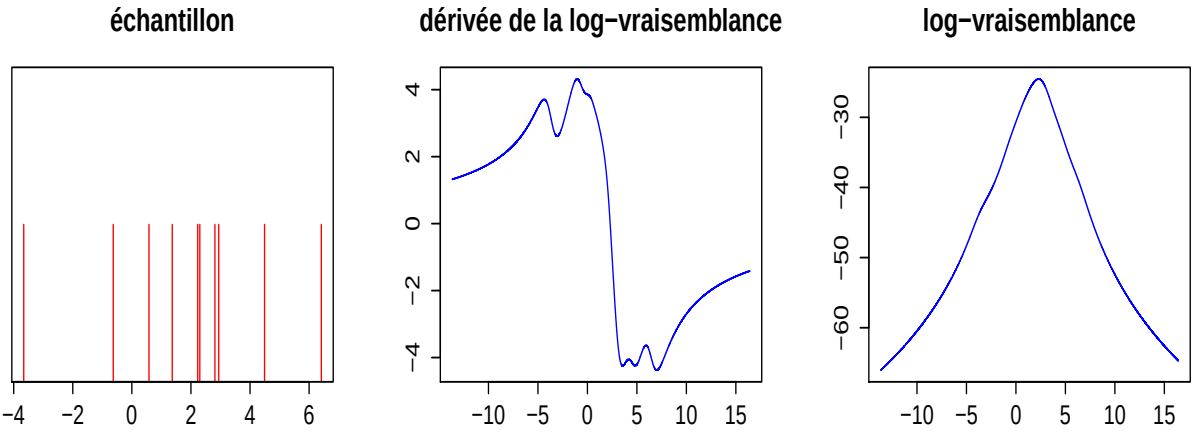

FIGURE 4.2 – 10 variables de Cauchy avec $\theta = 2$, dérivée de la log-vraisemblance et log-vraisemblance.

1. noter que $[X_{(n)}-1, X_{(1)}+1]$ est toujours non vide car $0 < X_{(n)} - X_{(1)} < 2$.

On considère la loi de Cauchy translatée déjà croisée au Chapitre 3, à savoir

$$f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

La log-vraisemblance s'écrit

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

Elle est continue et tend vers $-\infty$ lorsque $\theta \to \pm\infty$, donc elle admet un (ou plusieurs) EMV. Il "suffit" pour le(s) trouver d'annuler la dérivée :

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

Après réduction au même dénominateur, on obtient au numérateur un polynôme non trivial de degré $(2n-1)$. Même en cherchant ses racines de façon numérique, il peut y en avoir jusqu'à $(2n-1)$, ce qui devient prohibitif en temps de calcul en présence d'un échantillon de taille conséquente (voir aussi Figure 4.2). Bref, on préférera de loin l'estimateur $x_{1/2}(n)$ de la médiane empirique vu au Chapitre 3, lequel se calcule en deux coups de cuillère à pot. Il suffit en effet d'ordonner l'échantillon et de prendre le point du milieu : $x_{1/2}(n) = X_{(\lceil n/2\rceil)}$.

Un exemple retors

On part de

$$f(x) = \frac{1}{6}\left(\frac{1}{\sqrt{|x|}}\mathbb{1}_{]0,1]}(|x|) + \frac{1}{x^2}\mathbb{1}_{]1,+\infty[}(|x|)\right).$$

Ceci définit bien une densité, laquelle présente la particularité de ne pas être définie en 0, où elle explose. On considère alors la famille de densités $(f_\theta)_{\theta\in\mathbb{R}}$ obtenues par translation de f, c'est-à-dire pour tout $\theta \in \mathbb{R}$ et tout $x \neq \theta$,

$$f_\theta(x) = f(x-\theta) = \frac{1}{6}\left(\frac{1}{\sqrt{|x-\theta|}}\mathbb{1}_{]0,1]}(|x-\theta|) + \frac{1}{(x-\theta)^2}\mathbb{1}_{]1,+\infty[}(|x-\theta|)\right). \tag{4.4}$$

Pour un n-échantillon $(X_1,\ldots,X_n)$, la log-vraisemblance s'écrit donc

$$\ell_n(\theta) = -n\log 6 - \frac{1}{2}\sum_{i=1}^n \log(|X_i-\theta|)\mathbb{1}_{]0,1]}(|X_i-\theta|) - 2\sum_{i=1}^n \log(|X_i-\theta|)\mathbb{1}_{]1,+\infty[}(|X_i-\theta|).$$

Clairement, cette fonction tend vers $+\infty$ dès que θ tend vers l'un des X_i. Il n'y a donc pas d'estimateur du maximum de vraisemblance (voir Figure 4.4). On peut aussi noter que si X a pour densité f_θ, elle n'admet pas d'espérance, donc la méthode des moments mène elle aussi à une impasse. Pour estimer θ, on peut néanmoins s'en sortir en passant par la médiane empirique. En effet, la fonction de répartition associée à la densité f est

$$F(x) = \begin{cases} -1/(6x) & \text{si } x \le -1 \\ 1/2 - \sqrt{-x}/3 & \text{si } -1 \le x \le 0 \\ 1/2 + \sqrt{x}/3 & \text{si } 0 \le x \le 1 \\ 1 - 1/(6x) & \text{si } x \ge 1 \end{cases}$$

Cette fonction est continue bijective, de médiane 0. Par translation, la médiane de la variable aléatoire X de densité f_θ est donc θ, le paramètre que l'on cherche à estimer. Notant comme d'habitude $x_{1/2}(n) = X_{(\lceil n/2 \rceil)}$ la médiane empirique, le résultat de consistance s'applique :

$$x_{1/2}(n) \xrightarrow[n\to\infty]{p.s.} \theta.$$

Par contre, la normalité asymptotique telle qu'énoncée en Théorème 11 est hors-sujet puisque la densité n'est pas définie en la médiane θ. Il n'en reste pas moins que l'on peut toujours construire des intervalles de confiance grâce à la méthode vue et revue du passage par la fonction de répartition empirique : ainsi, $[X_{(\lceil n/2-\sqrt{n} \rceil)}, X_{(\lceil n/2+\sqrt{n} \rceil)}]$ est un intervalle de confiance asymptotique à 95%.

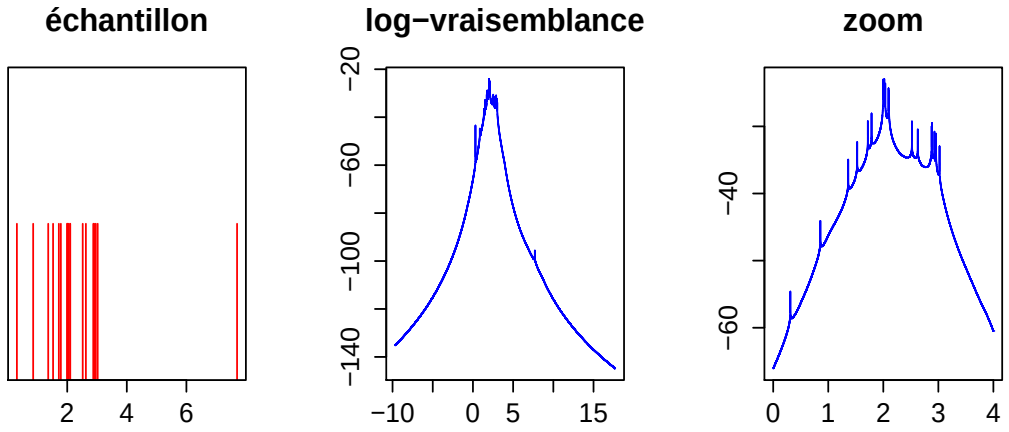

FIGURE 4.3 – Echantillon de 20 variables de loi (4.4) avec $\theta = 2$ et log-vraisemblance "explosive".

4.2.3 Modèle exponentiel

Commençons par un rappel d'analyse. Partons de ν, mesure positive σ-finie sur $\mathbb{R}$: ce sera la mesure de Lebesgue, ou de comptage, ou une mesure absolument continue par rapport à l'une d'entre elles. Pour $T : \mathbb{R} \to \mathbb{R}$ une fonction borélienne donnée et

$$I = \left\{ \eta \in \mathbb{R},\ \int_{\mathbb{R}} \exp(\eta T(x))\ \nu(dx) < +\infty \right\},$$

la convexité de la fonction exponentielle assure que I est un intervalle de $\mathbb{R}$. Dans toute la suite, nous supposerons qu'il est d'intérieur non vide. Le résultat suivant précise alors les dérivées de la fonction $M : I \to \mathbb{R}$ définie par

$$M(\eta) = \int_{\mathbb{R}} \exp(\eta T(x))\ \nu(dx)$$

sur l'intervalle ouvert $\mathring{I}$. Celles-ci s'obtiennent tout bonnement par dérivation sous le signe somme, application classique du Théorème de convergence dominée.

Proposition 17 (Dérivation de la transformée de Laplace)
La fonction M est C^∞ sur $\mathring{I}$ et, pour tout entier naturel k,

$$M^{(k)}(\eta) = \int_{\mathbb{R}} (T(x))^k \exp(\eta T(x))\ \nu(dx).$$

Si η appartient à I, la fonction $g_\eta : \mathbb{R} \to \mathbb{R}$ définie par

$$g_\eta(x) = \exp(\eta T(x) - A(\eta)) \qquad \text{avec} \qquad A(\eta) = \log\left(\int_{\mathbb{R}} \exp(\eta T(x))\ \nu(dx)\right), \tag{4.5}$$

est donc une densité par rapport à la mesure ν. Ou encore : pour tout $\eta \in I$, la mesure $R_\eta = g_\eta \cdot \nu$ définit une mesure de probabilité sur $\mathbb{R}$, et $(R_\eta)_{\eta \in I}$ constitue une famille de lois probabilités sur $\mathbb{R}$. Question naturelle : le modèle associé est-il identifiable ?

Lemme 1 (Condition d'identifiabilité du modèle)
Si T n'est pas ν presque partout constante, c'est-à-dire s'il n'existe pas de constante c et de borélien E tel que $\nu(\mathbb{R} \setminus E) = 0$ et $T(x) = c$ pour tout x dans E, alors le modèle $(R_\eta)_{\eta \in I}$ est identifiable.

Preuve. On procède par contraposition. Si le modèle n'est pas identifiable, il existe η et η' distincts dans I tels que les mesures de probabilités $R_\eta(dx) = g_\eta(x)\nu(dx)$ et $R_{\eta'}(dx) = g_{\eta'}(x)\nu(dx)$ coïncident. Ceci est équivalent à dire que $g_\eta(x) = g_{\eta'}(x)$ pour ν presque tout x, donc

$$(\eta - \eta')T(x) = A(\eta) - A(\eta') \qquad \nu \text{ p.p.},$$

ce qui impliquerait que T est ν presque partout constante.

■

Hypothèse 1 (Identifiabilité du modèle)
Dans toute la suite, nous supposerons que T est non ν presque partout constante, c'est-à-dire que le modèle $(R_\eta)_{\eta \in I}$ est identifiable.

Dans la plupart des cas, la mesure ν étant liée à la mesure de Lebesgue ou à la mesure de comptage, cette vérification est triviale. Noter que même si le modèle défini par $(R_\eta)_{\eta \in I}$ est identifiable, la représentation sous forme exponentielle telle que donnée en (4.5) n'est pas unique pour autant : il suffit pour s'en convaincre de remplacer ν par 2ν et $A(\eta)$ par $A(\eta) + \log 2$.

Définition 26 (Modèle exponentiel canonique)
L'ensemble des lois $(R_\eta)_{\eta \in I}$, avec

$$R_\eta(dx) = g_\eta(x)\nu(dx) = \exp(\eta T(x) - A(\eta))\ \nu(dx),$$

est appelé modèle exponentiel canonique (ou naturel) de dimension 1, η étant le paramètre canonique (ou naturel).

Soit maintenant $\eta \in I$ et X une variable aléatoire de loi R_η. La Proposition 17 permet de calculer les moments de X sans aucune difficulté.

Proposition 18 (Moments de X et dérivées de A)
La fonction de normalisation A est C^∞ sur $\mathring{I}$, avec pour premières dérivées

$$A'(\eta) = \mathbb{E}_\eta[T(X)] \qquad \text{et} \qquad A''(\eta) = \mathrm{Var}_\eta[T(X)],$$

où l'indice η signifie que X suit la loi R_η. De plus, par l'Hypothèse H1, la variance ne peut s'annuler, donc A est strictement convexe.

Nous allons généraliser le modèle exponentiel canonique pour arriver aux formes que prennent la plupart des familles de lois rencontrées en pratique.

Définition 27 (Modèle exponentiel général)
Soit μ une mesure positive σ-finie sur $\mathbb{R}$. On appelle modèle exponentiel général de dimension 1 toute famille de densités par rapport à μ de la forme

$$f_\theta(x) = C(\theta)h(x)\exp(Q(\theta)T(x)),$$

avec $\int_{\mathbb{R}} h(x)\mu(dx) > 0$ et $C(\theta) > 0$ pour tout $\theta \in \Theta$ intervalle d'intérieur non vide de $\mathbb{R}$. La fonction Q est continue et strictement croissante sur Θ, et la fonction T n'est pas presque partout constante pour la mesure $\nu = h \cdot \mu$.

Puisque f_θ est une densité par rapport à μ, $C(\theta)$ n'est rien d'autre qu'une constante de normalisation :

$$C(\theta) = \frac{1}{\displaystyle\int_{\mathbb{R}} h(x)\exp(Q(\theta)T(x))\mu(dx)}.$$

Par ailleurs, les propriétés requises sur Q impliquent que $Q(\Theta)$ est encore un intervalle d'intérieur non vide de $\mathbb{R}$. La connection entre modèle général et modèle canonique se fait comme suit.

Lemme 2 (Passage du général au canonique)
Pour passer du modèle exponentiel général de la Définition 27 au modèle exponentiel canonique de la Définition 26, il suffit de considérer $\nu = h \cdot \mu$, $\eta = Q(\theta)$, $A(\eta) = -\log C(\theta) = -\log(C \circ Q^{-1}(\eta))$ et $Q(\Theta) \subseteq I$.

Les propriétés du modèle général se déduisent donc de celles du modèle canonique. En particulier, puisque Q est injective et T non presque partout constante pour $\nu = h \cdot \mu$, le modèle est identifiable. Par ailleurs, si on note $\mathbb{E}_\theta$ et $\mathbb{E}_\eta$ les moyennes pour le modèle général et pour le modèle canonique, on a pour toute fonction test φ

$$\mathbb{E}_\theta[\varphi(X)] = \int_{\mathbb{R}} \varphi(x)f_\theta(x)\mu(dx) = \int_{\mathbb{R}} \varphi(x)g_\eta(x)\nu(dx) = \mathbb{E}_\eta[\varphi(X)] \text{ si } \eta = Q(\theta).$$

Un grand nombre de lois classiques rentrent dans le modèle exponentiel général. Nous ne mentionnons que deux d'entre elles et donnons par ailleurs un exemple de famille de lois qui ne s'inscrit pas dans ce cadre.

Exemples :

1. Lois de Poisson : comme vu en Section 4.2.2, la densité par rapport à la mesure de comptage sur $\mathbb{N}$ de toute loi de Poisson de paramètre $\lambda > 0$ s'écrit

   $$f_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda} \times \frac{1}{x!} \times \exp(x \log \lambda),$$

   avec toutes les propriétés requises pour $C(\lambda) = e^{-\lambda}$, $h(x) = 1/x!$, $Q(\lambda) = \log \lambda$, $\Lambda =]0, +\infty[$ et $T(x) = x$.
2. Lois normales : la famille de lois $(\mathcal{N}(\theta, 1))_{\theta\in\mathbb{R}}$ est un autre exemple de modèle exponentiel, avec μ la mesure de Lebesgue sur $\mathbb{R}$, $C(\theta) = e^{-\theta^2/2}$, $h(x) = e^{-x^2/2}/\sqrt{2\pi}$, $Q(\theta) = \theta$, $\Theta = \mathbb{R}$ et $T(x) = x$.
3. Lois Gamma : rappelons que si (r, λ) est un couple de réels strictement positifs, la loi $\Gamma(r, \lambda)$ a pour densité par rapport à la mesure de Lebesgue

   $$f(x) = f_{r,\lambda}(x) = \frac{(\lambda x)^{r-1}}{\Gamma(r)}\lambda e^{-\lambda x}\mathbb{1}_{x\geq 0}.$$

Si r est connu et λ inconnu, l'écriture suivante donne cette densité sous forme de modèle exponentiel :

$$f_\lambda(x) = \lambda^r \times \frac{x^{r-1}\mathbb{1}_{x\geq 0}}{\Gamma(r)} \times e^{-\lambda x}.$$

Si λ est connu et r inconnu, on écrira plutôt

$$f_r(x) = \frac{\lambda^r}{\Gamma(r)} \times e^{-\lambda x}\mathbb{1}_{x\geq 0} \times e^{(r-1)\log x},$$

qui donne bien un modèle exponentiel.

4. Lois uniformes : considérons par exemple le modèle à un paramètre d'une loi uniforme sur $[0,\theta]$. Le support $[0,\theta]$ varie avec le paramètre θ, configuration tout à fait exclue par la Définition 27, donc cette famille de lois ne peut s'écrire comme un modèle exponentiel. Cette situation "pathologique" où le support dépend du paramètre à estimer est typique des cas ne rentrant pas dans le modèle exponentiel.

Passons maintenant au problème d'inférence statistique. L'estimateur du maximum de vraisemblance s'étudie très bien dans le cadre des modèles exponentiels. Commençons par le modèle canonique, plus facile d'approche.

Proposition 19 (EMV et modèle canonique)
Soit $(X_1,\dots,X_n)$ un n-échantillon iid de densité g_η par rapport à une mesure ν, où $(g_\eta)_{\eta\in I}$ est un modèle exponentiel canonique :

$$g_\eta(x) = \exp(\eta T(x) - A(\eta)).$$

Si le vrai paramètre η est à l'intérieur de I, alors presque sûrement, pour n assez grand, l'estimateur du maximum de vraisemblance $\hat\eta_n$ existe et est l'unique solution de

$$\mathbb{E}_{\hat\eta_n}[T(X)] = A'(\hat\eta_n) = \bar T_n = \frac{1}{n}\sum_{i=1}^n T(X_i).$$

De plus, cet estimateur est consistant, c'est-à-dire

$$\hat\eta_n \xrightarrow[n\to\infty]{p.s.} \eta,$$

et asymptotiquement normal, avec

$$\sqrt{n}\,(\hat\eta_n - \eta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/A''(\eta)) = \mathcal{N}(0, 1/\mathrm{Var}_\eta(T(X))).$$

Preuve. Partant de la formule (4.5), la log-vraisemblance s'écrit

$$\ell_n(\eta) = \eta\sum_{i=1}^n T(X_i) - nA(\eta) = n\left(\eta\bar T_n - A(\eta)\right) \implies \ell_n''(\eta) = -nA''(\eta).$$

Sur $\mathring{I}$, puisque $A'' > 0$, la fonction ℓ_n est strictement concave et s'il existe $\hat\eta_n$ tel que $\ell_n'(\hat\eta_n) = 0$, ce sera donc l'unique estimateur du maximum de vraisemblance. Ceci est équivalent à dire que

$$A'(\hat\eta_n) = \mathbb{E}_{\hat\eta_n}[T(X)] = \bar T_n. \tag{4.6}$$

Or A' est continue et strictement croissante sur $\mathring{I}$ puisque $A'' > 0$, donc bijective de $\mathring{I}$ sur l'intervalle ouvert $A'(\mathring{I})$. En particulier, si le vrai paramètre η est à l'intérieur de I, alors $A'(\eta)$ est forcément à l'intérieur de $A'(\mathring{I})$. La loi forte des grands nombres implique par ailleurs que

$$\bar T_n \xrightarrow[n\to\infty]{p.s.} \mathbb{E}_\eta[T(X)] = A'(\eta).$$

Ainsi, presque sûrement, pour n assez grand, $\bar{T}_n$ est intérieur à $A'(I)$ et l'équation (4.6) a une unique solution $\hat{\eta}_n = (A')^{-1}(\bar{T}_n)$. Par le théorème de continuité, on en déduit sa consistance :

$$\hat{\eta}_n = (A')^{-1}(\bar{T}_n) \xrightarrow[n\to\infty]{p.s.} (A')^{-1}(A'(\eta)) = \eta.$$

Passons à la vitesse. Le TCL appliqué à $\bar{T}_n$ nous dit que

$$\sqrt{n}\left(\bar{T}_n - \mathbb{E}_\eta[T(X)]\right) = \sqrt{n}\left(\bar{T}_n - A'(\eta)\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}_\eta(T(X))) = \mathcal{N}(0, A''(\eta)).$$

Il suffit pour conclure d'appliquer la Delta méthode avec $(A')^{-1}$, soit

$$\sqrt{n}\,(\hat{\eta}_n - \eta) = \sqrt{n}\left((A')^{-1}(\bar{T}_n) - (A')^{-1}(A'(\eta))\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/A''(\eta)),$$

puisque, comme chacun sait,

$$\left\{\left((A')^{-1}\right)'(A'(\eta))\right\}^2 A''(\eta) = \frac{1}{A''(\eta)}.$$

■

Exemple. Considérons la famille des lois exponentielles de densités, pour tout $\lambda \in \Lambda =]0, +\infty[$,

$$g_\lambda(x) = \lambda e^{-\lambda x} = \exp(-\lambda x + \log(\lambda))$$

par rapport à la mesure $\nu(dx) = \mathbb{1}_{[0,+\infty[}(x)dx$. Cette famille est directement sous forme canonique, avec $T(x) = -x$ et

$$A(\lambda) = -\log(\lambda) \implies A'(\lambda) = -\frac{1}{\lambda} \implies A''(\lambda) = \frac{1}{\lambda^2}.$$

Etant donné un échantillon $(X_1, \ldots, X_n)$ iid de densité g_λ, l'estimateur du maximum de vraisemblance $\hat{\lambda}_n$ vérifie

$$A'(\hat{\lambda}_n) = \bar{T}_n \iff -\frac{1}{\hat{\lambda}_n} = -\frac{1}{n}\sum_{i=1}^n X_i = -\bar{X}_n \iff \hat{\lambda}_n = \frac{1}{\bar{X}_n}.$$

Presque sûrement, l'EMV est donc bien défini, la probabilité que tous les X_i soient nuls étant elle-même nulle. Puisque l'intervalle Λ est ouvert, le vrai paramètre est nécessairement à l'intérieur et le résultat précédent assure à la fois la consistance et la normalité asymptotique, avec plus précisément

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/A''(\eta)) = \mathcal{N}(0, \lambda^2).$$

La passage au modèle exponentiel général est alors très simple : formellement, il consiste juste à appliquer une seconde fois la Delta méthode.

Théorème 15 (EMV et modèle général)
Soit $(X_1, \ldots, X_n)$ un n-échantillon iid de densité f_θ par rapport à la mesure μ, où $(f_\theta)_{\theta\in\Theta}$ est un modèle exponentiel général :

$$f_\theta(x) = C(\theta)h(x)\exp(Q(\theta)T(x)).$$

Si le vrai paramètre θ est à l'intérieur de Θ, alors presque sûrement, pour n assez grand, l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ existe et est l'unique solution de

$$\mathbb{E}_{\hat{\theta}_n}[T(X)] = \bar{T}_n = \frac{1}{n}\sum_{i=1}^n T(X_i).$$

Cet estimateur est consistant, c'est-à-dire

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

Si, de plus, Q est dérivable et $Q'(\theta) \neq 0$, alors

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{(Q'(\theta))^2 \mathrm{Var}_\theta(T(X))}\right).$$

Preuve. Si le vrai paramètre θ est à l'intérieur de Θ alors, dans le modèle canonique, $\eta = Q(\theta)$ est à l'intérieur de $Q(\Theta)$ (puisque Q est continue bijective) donc à l'intérieur de I. Par la Proposition 19, presque sûrement, pour n assez grand, l'EMV $\hat{\eta}_n$ existe, est unique et vérifie

$$\hat{\eta}_n \xrightarrow[n\to\infty]{p.s.} \eta = Q(\theta).$$

En particulier, pour n assez grand, $\hat{\eta}_n$ appartient à $Q(\Theta)$ et on peut définir $\hat{\theta}_n = Q^{-1}(\hat{\eta}_n)$, qui correspond à l'EMV de θ puisque Q^{-1} fournit un changement bijectif de paramètre. Par continuité, on a aussi

$$\hat{\theta}_n = Q^{-1}(\hat{\eta}_n) \xrightarrow[n\to\infty]{p.s.} Q^{-1}(\eta) = \theta.$$

Si Q est dérivable et $Q'(\theta) \neq 0$, la Delta méthode appliquée au résultat de la Proposition 19 nous dit que

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \sqrt{n}\left(Q^{-1}(\hat{\eta}_n) - Q^{-1}(\eta)\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{(Q'(\theta))^2 \mathrm{Var}_\theta(T(X))}\right),$$

puisque, tenant compte de la relation $\eta = Q(\theta)$,

$$(Q^{-1})'(\eta) = \frac{1}{(Q' \circ Q^{-1})(\eta)} = \frac{1}{Q'(\theta)}.$$

■

Remarque. Si $(Q^{-1})'(Q(\theta)) = 0$, c'est-à-dire que Q a une tangente verticale au point θ, alors la Delta méthode s'applique encore en fin de preuve précédente et donne

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

c'est-à-dire que l'EMV tend vers sa cible θ à vitesse plus rapide que $1/\sqrt{n}$.

Exemples :

1. Revenons aux lois de Poisson :

$$f_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda} \times \frac{1}{x!} \times \exp(x \log \lambda) = C(\lambda)h(x)\exp(Q(\lambda)T(x)).$$

D'après le théorème précédent, l'EMV existe presque sûrement pour n assez grand et est solution de

$$\mathbb{E}_{\hat{\theta}_n}[T(X)] = \bar{T}_n = \frac{1}{n}\sum_{i=1}^n T(X_i).$$

Or $T(X) = X$ implique

$$\mathbb{E}_{\hat{\theta}_n}[T(X)] = \mathbb{E}_{\hat{\theta}_n}[X] = \hat{\theta}_n,$$

moyenne d'une variable X suivant une loi de Poisson de paramètre $\hat{\theta}_n$. L'équation précédente se résume donc à $\hat{\theta}_n = \bar{X}_n$, mais ce pour n assez grand uniquement. En effet, comme on l'a

vu précédemment, ceci n'a de sens que si les X_i ne sont pas tous nuls. Puisque $Q(\lambda) = \log \lambda$ et $\mathrm{Var}_\lambda(T(X)) = \mathrm{Var}_\lambda(X) = \lambda$, la variance asymptotique vaut quant à elle

$$\frac{1}{(Q'(\lambda))^2 \mathrm{Var}_\theta(T(X))} = \lambda,$$

ce qui correspond bien à ce qu'on avait trouvé auparavant.

2. Finissons par un exemple illustrant le problème lorsque le vrai paramètre n'est pas intérieur à Θ. On considère $(X_1, \ldots, X_n)$ iid selon une loi gaussienne $\mathcal{N}(\theta, 1)$ avec $\Theta = [0, +\infty[$. Notons pour commencer qu'on a bien un modèle exponentiel général avec $Q(\theta) = \theta$ et $T(x) = x$. Pour trouver l'EMV, revenons à sa définition en maximisant la log-vraisemblance, c'est-à-dire

$$\ell_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2 \implies \operatorname*{argmax}_{\theta \geq 0} \ell_n(\theta) = \operatorname*{argmin}_{\theta \geq 0} \sum_{i=1}^n (X_i - \theta)^2.$$

C'est un brave trinôme en θ, mais il faut tenir compte de ce que $\theta \in [0, +\infty[$. Si $\bar{X}_n \geq 0$, on a donc $\hat{\theta}_n = \bar{X}_n$, sinon $\hat{\theta}_n = 0$. Au total, on conclut que $\hat{\theta}_n = \max(\bar{X}_n, 0)$. Supposons que le vrai paramètre θ est égal à 0. Alors

$$\bar{X}_n \sim \mathcal{N}(0, 1/n) \implies \sqrt{n}\bar{X}_n \sim \mathcal{N}(0,1) \implies \sqrt{n}\hat{\theta}_n \sim \max(\mathcal{N}(0,1), 0).$$

Ainsi, la loi limite n'est pas une gaussienne, mais une "demi-gaussienne", autrement dit une loi mixte (voir Figure 4.4). On n'a donc plus le même comportement asymptotique que pour un point θ intérieur à Θ.

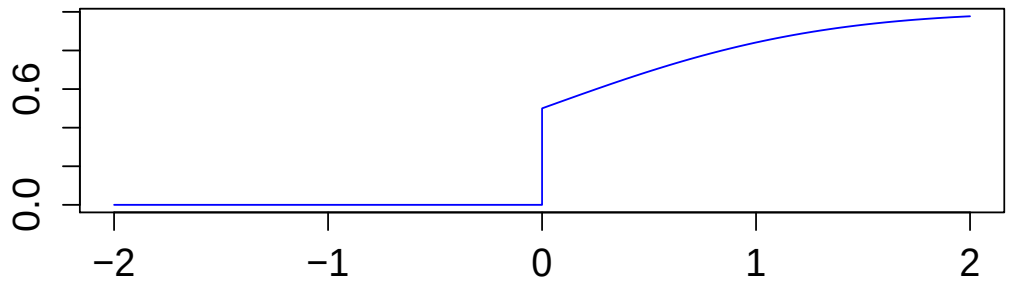

FIGURE 4.4 – Fonction de répartition de la "demi-gaussienne" : $\sqrt{n}\hat{\theta}_n \sim \max(\mathcal{N}(0,1), 0)$.

4.3 Estimateurs de Bayes

4.3.1 Risque bayésien

Pour quantifier la qualité d'une méthode statistique, on peut voir les choses comme suit : on dispose d'un ensemble $\mathcal{D}$ de décisions possibles, par exemple $\mathcal{D} = \Theta$ lorsqu'on construit un estimateur ou

$\mathcal{D} = \{0, 1\}$ pour un test d'hypothèses. Une procédure statistique est alors une application δ de E (typiquement $E = \mathbb{R}^n$) dans $\mathcal{D}$: si l'on observe l'échantillon $\mathbf{X} \in E$, on décide $\delta(\mathbf{X})$. En estimation, $\delta(\mathbf{X}) = \hat{\theta}(\mathbf{X})$ est un estimateur de θ, tandis pour les tests $\delta(\mathbf{X}) = T(\mathbf{X})$ est une statistique de test. L'intérêt d'introduire la notation δ à valeurs dans $\mathcal{D}$ est donc d'englober le tout dans un même cadre théorique.

On se donne en outre une fonction de perte[2], ou de coût, $\ell : \Theta \times \mathcal{D} \to \mathbb{R}_+$. Elle s'interprète comme suit : si θ est le "vrai" paramètre et que l'on décide $d \in \mathcal{D}$, on subit une perte égale à $\ell(\theta, d)$. Pour un problème d'estimation d'un paramètre réel θ, on prend typiquement $\mathcal{D} = \Theta$ et $\ell(\theta, \theta') = (\theta - \theta')^2$: c'est la perte quadratique. En test d'hypothèse, on peut choisir par exemple

$$\ell(\theta, d) = \begin{cases} 0 & \text{si } \theta \in \Theta_0 \text{ et } d = 0 \\ 0 & \text{si } \theta \in \Theta_1 \text{ et } d = 1 \\ C_0 & \text{si } \theta \in \Theta_0 \text{ et } d = 1 \\ C_1 & \text{si } \theta \in \Theta_1 \text{ et } d = 0 \end{cases}$$

Ainsi, la perte est nulle si on ne se trompe pas et vaut C_0 ou C_1 selon la façon dont on se trompe (erreur de première ou de deuxième espèce). En jouant sur ces deux valeurs, on peut donner plus ou moins d'importance aux deux types d'erreurs. Pour un paramètre θ donné, l'erreur moyenne faite par une procédure se calcule alors en intégrant cette perte par rapport à la loi de $\mathbf{X}$, notée comme d'habitude P_θ.

Définition 28 (Risque d'une procédure statistique)
Soit Θ un espace de paramètres, $\mathcal{D}$ un ensemble de décisions possibles, ℓ une fonction de perte, alors le risque d'une procédure statistique $\delta : E \to \mathcal{D}$ est la fonction $R(\cdot, \delta)$ de Θ dans $\mathbb{R}_+$ définie par

$$\forall \theta \in \Theta \qquad R(\theta, \delta) = \mathbb{E}\left[\ell(\theta, \delta(\mathbf{X}))\right] = \mathbb{E}_\theta\left[\ell(\theta, \delta(\mathbf{X}))\right] = \int_E \ell(\theta, \delta(\mathbf{x})) P_\theta(d\mathbf{x}).$$

Exemples :

- Si $\mathcal{D} = \Theta \subseteq \mathbb{R}$ et si ℓ correspond à la perte quadratique, on retrouve le risque quadratique. De façon plus générale, on peut considérer la perte L_p ($p > 0$) définie par $\ell(\theta, \theta') = |\theta - \theta'|^p$, et on obtiendra le risque L_p.
- Pour les tests, avec la fonction de perte définie plus haut, on a

$$R(\theta, \delta) = \begin{cases} C_0\ \mathbb{P}_\theta(T(\mathbf{X}) = 1) & \text{si } \theta \in \Theta_0 \\ C_1\ \mathbb{P}_\theta(T(\mathbf{X}) = 0) & \text{si } \theta \in \Theta_1 \end{cases}$$

Si ce cadre permet d'évaluer les performances d'une procédure δ, il semble difficile de comparer deux procédures δ et δ' entre elles : en général, on n'aura ni $R(\theta, \delta) \leq R(\theta, \delta')$ pour tout $\theta \in \Theta$, ni l'inégalité inverse.

Exemple : dans un cadre d'estimation, pour le modèle gaussien $(\mathcal{N}(\theta, 1))_{\theta \in \mathbb{R}}$ et à partir d'un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ iid de loi $\mathcal{N}(\theta, 1)$, considérons les deux estimateurs suivants : l'EMV $\hat{\theta} = \hat{\theta}(\mathbf{X}) = \bar{X}_n$ et l'estimateur constant $\check{\theta} = \check{\theta}(\mathbf{X}) = 0$. Les risques quadratiques associés sont donc

$$R(\theta, \hat{\theta}) = \frac{1}{n} \qquad \text{et} \qquad R(\theta, \check{\theta}) = \theta^2,$$

d'où

$$R(\theta, \hat{\theta}) \leq R(\theta, \check{\theta}) \iff |\theta| \geq \frac{1}{\sqrt{n}}.$$

2. *loss* en anglais, d'où la notation ℓ, qui n'a donc rien à voir avec la log-vraisemblance.

Autrement dit, à n fixé et puisqu'on ne connaît pas θ, on ne peut pas dire avec certitude que l'EMV est meilleur que le bête estimateur constant.

Une façon de sortir de cette impasse est de considérer un risque moyen : on se donne une mesure de probabilité μ sur Θ et on intègre le risque par rapport à celle-ci.

Définition 29 (Risque bayésien)
Le risque de Bayes de la procédure δ pour la loi a priori μ est défini par

$$R_B(\mu, \delta) = \int_\Theta R(\theta, \delta)\mu(d\theta) = \int_\Theta \left(\int_E \ell(\theta, \delta(\mathbf{x}))P_\theta(d\mathbf{x}) \right) \mu(d\theta).$$

Exemple. Dans le cadre du test indiqué plus haut, on a donc

$$R_B(\mu, \delta) = C_0 \int_{\Theta_0} \mathbb{P}_\theta(T(\mathbf{X}) = 1)\mu(d\theta) + C_1 \int_{\Theta_1} \mathbb{P}_\theta(T(\mathbf{X}) = 0)\mu(d\theta).$$

Ceci est typique de l'approche bayésienne : on considère que le paramètre θ est lui-même une variable aléatoire $\boldsymbol{\theta}$ de loi μ, P_θ étant alors la loi conditionnelle de $\mathbf{X}$ sachant $\boldsymbol{\theta} = \theta$. Dans ce paradigme, le risque bayésien s'écrit

$$R_B(\mu, \delta) = \mathbb{E}\left[\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))\right],$$

où il faut donc moyenniser sur θ et sur $\mathbf{x}$. Quoi qu'il en soit, l'un des intérêts du risque bayésien est de permettre de comparer des procédures entre elles, la qualité d'une procédure étant désormais quantifiée par un nombre positif indépendant de θ. En général, ce risque bayésien ne dépend plus que de n, le nombre d'observations.

Exemple : dans le cadre d'estimation du modèle gaussien, considérons que $\boldsymbol{\theta} \sim \mu = \mathcal{N}(0, 1)$. Les risques bayésiens pour cette loi a priori μ et la perte quadratique sont donc

$$R_B(\mu, \hat{\theta}) = \frac{1}{n} \qquad \text{et} \qquad R_B(\mu, \check{\theta}) = \mathbb{E}[\boldsymbol{\theta}^2] = 1.$$

Ainsi, au sens du risque bayésien, l'EMV est meilleur que l'estimateur constant.

4.3.2 Loi a posteriori et estimateurs de Bayes

Dans tout ce qui suit, nous considérons que les lois en jeu ont des densités par rapport à des mesures de référence. Ainsi λ sera la mesure de référence sur Θ (typiquement la mesure de Lebesgue) et ν la mesure de référence sur E (typiquement la mesure de Lebesgue ou la mesure de comptage). Par ailleurs, nous renvoyons au cours de probabilités pour les définitions des lois, densités et espérances conditionnelles, lesquelles vont jouer un rôle crucial dans la suite. Nous noterons :

- $f(\theta)$ la densité de $\boldsymbol{\theta}$ par rapport à la mesure λ. Elle définit la loi a priori (ou *prior* en anglais).
- $f(\mathbf{x}|\theta)$ la densité conditionnelle de $\mathbf{X}$ sachant $\Theta = \theta$ par rapport à la mesure ν. Elle est appelée vraisemblance (ou *likelihood* en anglais).
- $f(\mathbf{x})$ la densité de $\mathbf{X}$ par rapport à la mesure ν. En général, cette dernière ne sera pas calculée.

Exemple. Supposons que $\boldsymbol{\theta} \sim \mathcal{N}(0, 1)$ et que, sachant $\boldsymbol{\theta} = \theta$, $\mathbf{X} = (X_1, \ldots, X_n)$ soit un échantillon iid de loi $\mathcal{N}(\theta, 1)$. Alors $\Theta = \mathbb{R}$, λ est la mesure de Lebesgue sur $\mathbb{R}$, la loi a priori $f(\theta)$ est la loi normale centrée réduite, $\nu = \lambda^{\otimes n}$ est la mesure de Lebesgue sur $\mathbb{R}^n$ et pour tout $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, la densité conditionnelle s'écrit

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n \tfrac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} = \tfrac{1}{(2\pi)^{n/2}} \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \right\}.$$

Remarque. On rencontre aussi les notations $\pi(\theta)$ ou $g(\theta)$ pour la densité a priori $f(\theta)$, ainsi que $m(\mathbf{x})$ ou $\bar{f}(\mathbf{x})$ pour la densité marginale $f(\mathbf{x})$ et, cerise sur le gâteau, $\ell(\theta|\mathbf{x})$ pour la densité conditionnelle $f(\mathbf{x}|\theta)$.

Si on note $f(\theta|\mathbf{x})$ la densité conditionnelle (par rapport à λ) de $\boldsymbol{\theta}$ sachant $\mathbf{X} = \mathbf{x}$, la formule de Bayes donne (presque sûrement)

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)f(\theta)}{\displaystyle\int_\Theta f(\mathbf{x}|t)f(t)\lambda(dt)}. \tag{4.7}$$

Définition 30 (Loi a posteriori)
La loi conditionnelle de $\boldsymbol{\theta}$ sachant $\mathbf{X} = \mathbf{x}$, de densité $f(\theta|\mathbf{x})$ par rapport à la mesure λ, est appelée loi a posteriori de $\boldsymbol{\theta}$ sachant l'observation $\mathbf{X}$ (ou posterior en anglais).

Exemple. Dans l'exemple précédent, cette densité a posteriori s'écrit donc

$$f(\theta|\mathbf{x}) = \frac{\frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2\right\} \times \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\theta^2}{2}\right\}}{f(\mathbf{x})}$$

Puisqu'on cherche une densité par rapport à la variable θ, tout ce qui ne dépend pas de θ joue le rôle d'une constante de normalisation, d'où l'utilisation du symbole $\propto$ ("proportionnel à"). Ainsi l'écriture

$$f(\theta|\mathbf{x}) \propto \exp\left\{-\tfrac{1}{2}((n+1)\theta^2 - 2n\bar{\mathbf{x}}_n\theta)\right\} \tag{4.8}$$

signifie qu'il existe une constante de normalisation $c(\mathbf{x})$, ne faisant pas intervenir θ, telle que

$$f(\theta|\mathbf{x}) = c(\mathbf{x}) \exp\left\{-\tfrac{1}{2}((n+1)\theta^2 - 2n\bar{\mathbf{x}}_n\theta)\right\} = \frac{\exp\left\{-\frac{1}{2}((n+1)\theta^2 - 2n\bar{\mathbf{x}}_n\theta)\right\}}{\displaystyle\int_{\mathbb{R}} \exp\left\{-\tfrac{1}{2}((n+1)\theta^2 - 2n\bar{\mathbf{x}}_n\theta)\right\} d\theta}.$$

Revenant à l'équation (4.8), il suffit alors de faire apparaître une densité gaussienne pour en déduire la loi a posteriori :

$$f(\theta|\mathbf{x}) \propto \exp\left\{-\tfrac{n+1}{2}\left(\theta - \tfrac{n}{n+1}\bar{\mathbf{x}}_n\right)^2\right\} \implies \mathcal{L}(\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}) = \mathcal{N}\left(\tfrac{n}{n+1}\bar{\mathbf{x}}_n, \tfrac{1}{n+1}\right).$$

Par rapport à la loi a priori (i.e. la gaussienne standard), la loi a posteriori est donc centrée grosso modo autour de la moyenne des données observées et est bien plus concentrée autour de cette moyenne. Cette moyenne empirique se concentrant elle-même autour de la vraie valeur θ, la loi a posteriori se concentre autour de θ. Ceci est illustré Figure 4.5.

Remarque. On rencontre aussi les notations $\pi(\theta|\mathbf{x})$ et $g(\theta|\mathbf{x})$ pour la densité a posteriori $f(\theta|\mathbf{x})$.

Focalisons-nous désormais sur le problème d'estimation, i.e. $\mathcal{D} = \Theta$ et $\delta(\mathbf{X}) = \hat{\theta}(\mathbf{X})$. Avec ces notations, le risque de Bayes s'écrit, grâce au Théorème de Fubini,

$$R_B(\mu,\delta) = \mathbb{E}\left[\ell(\boldsymbol{\theta},\delta(\mathbf{X}))\right] = \mathbb{E}\left[\ell(\boldsymbol{\theta},\hat{\theta}(\mathbf{X}))\right] = \int_E \left(\int_\Theta \ell(\theta,\hat{\theta}(\mathbf{x}))f(\theta|\mathbf{x})\lambda(d\theta)\right) f(\mathbf{x})\nu(d\mathbf{x}).$$

Pour minimiser ce risque, il suffit de mimimiser, pour tout $\mathbf{x}$ fixé, la quantité

$$\int_\Theta \ell(\theta,\hat{\theta}(\mathbf{x}))f(\theta|\mathbf{x})\lambda(d\theta) = \mathbb{E}[\ell(\boldsymbol{\theta},\hat{\theta}(\mathbf{X}))|\mathbf{X}=\mathbf{x}],$$

c'est-à-dire l'espérance de la perte pour la loi a posteriori de $\boldsymbol{\theta}$.

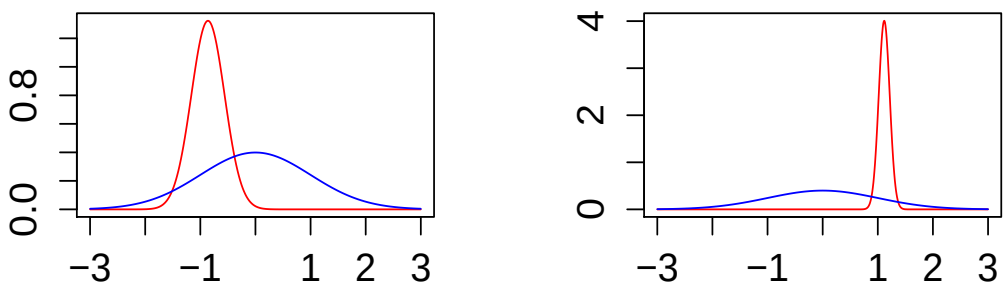

FIGURE 4.5 – Lois a priori $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{N}(0,1)$ (en bleu) et a posteriori $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{N}\left(\frac{n}{n+1}\bar{\mathbf{x}}_n, \frac{1}{n+1}\right)$ (en rouge) : deux exemples pour $n = 10$ (à gauche) et $n = 100$ (à droite).

Proposition 20 (Estimateur de Bayes)
Tout estimateur $\tilde{\theta}(\mathbf{X})$ qui minimise la perte moyenne pour la loi a posteriori de $\boldsymbol{\theta}$, c'est-à-dire tel que pour tout $\mathbf{x}$,

$$\tilde{\theta}(\mathbf{x}) = \operatorname*{argmin}_{t\in\Theta} \mathbb{E}[\ell(\boldsymbol{\theta}, t)|\mathbf{X} = \mathbf{x}] = \operatorname*{argmin}_{t\in\Theta} \int_\Theta \ell(\theta, t) f(\theta|\mathbf{x})\lambda(d\theta),$$

minimise le risque de Bayes. On l'appelle estimateur de Bayes pour la loi a priori μ. En particulier, l'estimateur de Bayes pour la perte quadratique correspond à l'espérance conditionnelle de $\boldsymbol{\theta}$ sachant $\mathbf{X}$, ou moyenne a posteriori de $\boldsymbol{\theta}$:

$$\ell(\theta, t) = (\theta - t)^2 \implies \tilde{\theta}(\mathbf{x}) = \mathbb{E}[\boldsymbol{\theta}|\mathbf{X} = \mathbf{x}] \implies \tilde{\theta}(\mathbf{X}) = \mathbb{E}[\boldsymbol{\theta}|\mathbf{X}].$$

De façon comparable, si $\ell(\theta, t) = |\theta - t|$, le minimum en t est la médiane de $\boldsymbol{\theta}$ sachant $\mathbf{X} = \mathbf{x}$. Ainsi la médiane conditionnelle, ou médiane a posteriori, correspond à l'estimateur de Bayes pour la perte L_1.

Exemple. Dans l'exemple précédent, sachant $\mathbf{X}$, la variable $\boldsymbol{\theta}$ suit une loi gaussienne de moyenne $n\bar{X}_n/(n+1)$ et de variance $1/(n+1)$. L'estimateur de Bayes pour la perte quadratique est donc tout simplement

$$\tilde{\theta}(\mathbf{X}) = \tfrac{n}{n+1}\bar{X}_n.$$

Comparaison. Restons sur cet exemple mais oublions le contexte bayésien et revenons au cadre d'inférence classique (ou fréquentiste). Le modèle statistique est donc celui de la famille de lois $(P_\theta)_{\theta\in\Theta} = (\mathcal{N}(\theta, 1))_{\theta\in\mathbb{R}}$ avec un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ iid de loi $\mathcal{N}(\theta, 1)$. On suppose qu'il existe "un vrai paramètre θ" et que le but est de l'estimer grâce aux X_i à disposition. Dans ce contexte, on a vu que les estimateurs de la méthode des moments et du maximum de vraisemblance coïncident, à savoir $\hat{\theta}(\mathbf{X}) = \bar{X}_n$. Stricto sensu, ils ne correspondent donc pas à l'estimateur de Bayes :

$$\tilde{\theta}(\mathbf{X}) = \tfrac{n}{n+1}\bar{X}_n \neq \hat{\theta}(\mathbf{X}) = \bar{X}_n,$$

néanmoins la différence devient négligeable lorsque la taille n de l'échantillon croît : c'est un phénomène typique. On peut noter que l'EMV est non biaisé, tandis que l'estimateur de Bayes

l'est. Quant à leurs variances respectives, puisque $\bar{X}_n \sim \mathcal{N}(\theta, 1/n)$, on a

$$\mathrm{Var}_\theta(\tilde{\theta}(\mathbf{X})) = \frac{n}{(n+1)^2} < \frac{1}{n} = \mathrm{Var}_\theta(\hat{\theta}(\mathbf{X})).$$

Les risques quadratiques valent donc :

$$R(\theta, \tilde{\theta}(\mathbf{X})) = \frac{n+\theta^2}{(n+1)^2} \qquad \text{et} \qquad R(\theta, \hat{\theta}(\mathbf{X})) = \frac{1}{n}.$$

Leur comparaison dépend de la position de $|\theta|$ par rapport à $\sqrt{2+1/n}$, à nouveau la différence devient négligeable avec n. Si on revient au cadre bayésien en tenant compte de la loi a priori $\mu = \mathcal{N}(0,1)$ sur $\boldsymbol{\theta}$ et en intégrant par rapport à celle-ci, alors on constate sans surprise que l'estimateur de Bayes est (un peu) meilleur :

$$R_B(\mu, \tilde{\theta}(\mathbf{X})) = \int_{\mathbb{R}} R(\theta, \tilde{\theta}(\mathbf{X}))\mu(d\theta) = \frac{1}{n+1} < \frac{1}{n} = \int_{\mathbb{R}} R(\theta, \hat{\theta}(\mathbf{X}))\mu(d\theta) = R_B(\mu, \hat{\theta}(\mathbf{X})).$$

Remarque. Dans le cas général, le problème de minimisation de la Proposition 20 fait intervenir la densité a posteriori $f(\theta|\mathbf{x})$ telle qu'explicitée par l'équation (4.7). En pratique, la densité a priori $f(\theta)$ et la densité conditionnelle $f(\mathbf{x}|\theta)$ sont données. Même si ces lois sont simples et connues, la densité a posteriori n'a, elle, aucune raison de l'être. En particulier, elle fait intervenir la constante de normalisation

$$f(\mathbf{x}) = \int_\Theta f(\mathbf{x}|t) f(t) \lambda(dt),$$

c'est-à-dire une intégrale souvent hors de portée lorsque $\mathbf{X}$ est de grande dimension. Par conséquent, hormis dans quelques cas d'école comme l'exemple précédent et celui qui va suivre, la seule façon de s'en sortir pour simuler/estimer cette loi a posteriori est de faire appel à des méthodes numériques, en particulier des techniques de Monte-Carlo par Chaînes de Markov (MCMC). Plus précisément, puisque le rapport

$$\frac{f(\theta'|\mathbf{x})}{f(\theta|\mathbf{x})} = \frac{f(\mathbf{x}|\theta')f(\theta')}{f(\mathbf{x}|\theta)f(\theta)}$$

ne fait pas intervenir cette normalisation problématique par $f(\mathbf{x})$, l'algorithme de Metropolis-Hastings est l'un des outils privilégiés du domaine.

4.3.3 Un exemple historique

L'exemple qui suit est dû à Bayes lui-même. Une boule de billard est lancée au hasard uniforme sur une ligne de longueur 1, sa position aléatoire étant notée $\boldsymbol{\theta}$ et sa réalisation $\theta \in [0,1]$. Ceci fait, une seconde boule est lancée de la même façon n fois de suite sur cette ligne et $\mathbf{X}$ est le nombre de fois où elle arrive à gauche de la première, donc à gauche de θ.

Avec les notations précédentes, la densité a priori $f(\theta)$ est celle de la loi uniforme sur $[0,1]$, la vraisemblance $f(\mathbf{x}|\theta)$ correspond à une loi binomiale $\mathcal{B}(n,\theta)$, et la densité a posteriori $f(\theta|\mathbf{x})$ est donc par la règle de Bayes rappelée en (4.7) :

$$f(\theta|\mathbf{x}) = \frac{\binom{n}{\mathbf{x}}\theta^{\mathbf{x}}(1-\theta)^{n-\mathbf{x}}}{\displaystyle\int_0^1 \binom{n}{\mathbf{x}} t^{\mathbf{x}}(1-t)^{n-\mathbf{x}} dt} \mathbb{1}_{[0,1]}(\theta) = \frac{\theta^{\mathbf{x}}(1-\theta)^{n-\mathbf{x}}}{\displaystyle\int_0^1 t^{\mathbf{x}}(1-t)^{n-\mathbf{x}} dt} \mathbb{1}_{[0,1]}(\theta).$$

Il ne faut pas se soucier du dénominateur puisqu'il ne dépend pas de la variable θ. Ici, comme toujours dans le calcul de la densité a posteriori, seul ce qui fait intervenir θ importe. Rappelons

que, si $a > 0$ et $b > 0$, la variable aléatoire Y suit une loi Bêta $\beta(a, b)$ si elle a pour densité

$$f(y) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)} \mathbb{1}_{]0,1[}(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\ y^{a-1}(1-y)^{b-1}\mathbb{1}_{]0,1[}(y).$$

Ses moments sont connus, en particulier $\mathbb{E}[Y] = \frac{a}{a+b}$. La loi a posteriori de $\boldsymbol{\theta}$ sachant $\mathbf{X} = \mathbf{x}$ est donc une loi Bêta $\beta(\mathbf{x}+1, n-\mathbf{x}+1)$. La moyenne a posteriori s'en déduit :

$$\mathbb{E}[\boldsymbol{\theta}|\mathbf{X} = \mathbf{x}] = \frac{\mathbf{x}+1}{n+2}.$$

Dès lors, si on adopte la perte quadratique, l'estimateur de Bayes est, d'après la Proposition 20,

$$\tilde{\theta}(\mathbf{X}) = \mathbb{E}[\boldsymbol{\theta}|\mathbf{X}] = \frac{\mathbf{X}+1}{n+2}.$$

Par la méthode des moments, sachant $\boldsymbol{\theta} = \theta$,

$$\mathbf{X} \sim \mathcal{B}(n,\theta) \implies \mathbb{E}[\mathbf{X}] = n\theta \implies \hat{\theta}(\mathbf{X}) = \mathbf{X}/n.$$

Pour le maximum de vraisemblance, il suffit de maximiser en θ la log-vraisemblance

$$\log\binom{n}{\mathbf{x}} + \mathbf{x}\log\theta + (n-\mathbf{x})\log(1-\theta),$$

ce qui donne encore $\hat{\theta}(\mathbf{X}) = \mathbf{X}/n$. Les estimateurs $\tilde{\theta}(\mathbf{X})$ et $\hat{\theta}(\mathbf{X})$ sont donc différents, cette différence s'atténuant avec n. On peut noter que, pour tout $\theta \in [0,1]$, l'estimateur de Bayes est en général biaisé puisque, $\mathbf{X}$ suivant une loi binomiale $\mathcal{B}(n,\theta)$, il vient

$$\mathbb{E}[\tilde{\theta}(\mathbf{X})] = \mathbb{E}\left[\frac{\mathbf{X}+1}{n+2}\right] = \frac{\mathbb{E}[\mathbf{X}]+1}{n+2} = \frac{n\theta+1}{n+2} \neq \theta \text{ si } \theta \neq \frac{1}{2}.$$

Sa variance est

$$\mathrm{Var}(\tilde{\theta}(\mathbf{X})) = \frac{\mathrm{Var}(\mathbf{X})}{(n+2)^2} = \frac{n\theta(1-\theta)}{(n+2)^2}.$$

Ceci donne comme risque quadratique, toujours à θ fixé,

$$R(\theta, \tilde{\theta}(\mathbf{X})) = \Big(\theta - \mathbb{E}[\tilde{\theta}(\mathbf{X})]\Big)^2 + \mathrm{Var}(\tilde{\theta}(\mathbf{X})) = \cdots = \frac{1+(n-4)\theta(1-\theta)}{(n+2)^2}.$$

Conformément à la Définition 29, le risque de Bayes s'obtient alors en intégrant cette quantité par rapport à la loi de $\boldsymbol{\theta}$, c'est-à-dire la loi μ uniforme sur $[0,1]$, ce qui donne :

$$R_B(\mu, \tilde{\theta}(\mathbf{X})) = \int_0^1 R(\theta, \tilde{\theta}(\mathbf{X}))\ d\theta = \int_0^1 \left(\frac{1+(n-4)\theta(1-\theta)}{(n+2)^2}\right) d\theta = \frac{1}{6(n+2)}.$$

Pour l'estimateur du max de vraisemblance, qui est non biaisé, le risque quadratique correspond à la variance

$$R(\theta, \hat{\theta}(\mathbf{X})) = \frac{\theta(1-\theta)}{n},$$

d'où un risque de Bayes égal à

$$R_B(\mu, \hat{\theta}(\mathbf{X})) = \int_0^1 R(\theta, \tilde{\theta}(\mathbf{X}))\ d\theta = \int_0^1 \frac{\theta(1-\theta)}{n}\ d\theta = \frac{1}{6n} > \frac{1}{6(n+2)}.$$

Ceci est bien cohérent avec le fait que l'estimateur de Bayes minimise le risque bayésien.

Remarques :

1. Pour tous paramètres a et b strictement plus grands que 1, le mode a posteriori d'une loi Bêta $\beta(a,b)$ est $(a-1)/(a+b-2)$, ce qui donne ici $\mathbf{X}/n$. L'estimateur du maximum de vraisemblance correspond donc, sur cet exemple, au mode a posteriori. Puisque l'on a considéré la perte quadratique, l'estimateur de Bayes correspond quant à lui à la moyenne a posteriori (Proposition 20).
2. Le fait que $\tilde{\theta}(\mathbf{X})$ soit biaisé n'est pas propre à notre exemple et ne doit pas susciter une émotion excessive. En effet, sauf cas exceptionnel, l'estimateur de Bayes pour la perte quadratique présente un biais. Précisément, dès lors que le risque bayésien est strictement positif, on a

$$0 < R_B(\mu, \tilde{\theta}(\mathbf{X})) = \mathbb{E}\left[(\boldsymbol{\theta} - \tilde{\theta}(\mathbf{X}))^2\right] = \mathbb{E}\left[\boldsymbol{\theta}(\boldsymbol{\theta} - \tilde{\theta}(\mathbf{X}))\right] - \mathbb{E}\left[\tilde{\theta}(\mathbf{X})(\boldsymbol{\theta} - \tilde{\theta}(\mathbf{X}))\right].$$

Or

$$\mathbb{E}\left[\tilde{\theta}(\mathbf{X})(\boldsymbol{\theta} - \tilde{\theta}(\mathbf{X}))\right] = \mathbb{E}\left[\mathbb{E}\left[\tilde{\theta}(\mathbf{X})(\boldsymbol{\theta} - \tilde{\theta}(\mathbf{X}))\middle|\mathbf{X}\right]\right] = \mathbb{E}\left[\tilde{\theta}(\mathbf{X})\mathbb{E}\left[\boldsymbol{\theta} - \tilde{\theta}(\mathbf{X})\middle|\mathbf{X}\right]\right] = 0,$$

puisque $\tilde{\theta}(\mathbf{X}) = \mathbb{E}[\boldsymbol{\theta}|\mathbf{X}]$. En conditionnant le terme restant par rapport à $\boldsymbol{\theta}$, il vient

$$0 < R_B(\mu, \tilde{\theta}(\mathbf{X})) = -\mathbb{E}\left[\boldsymbol{\theta}(\mathbb{E}[\tilde{\theta}(\mathbf{X})|\boldsymbol{\theta}] - \boldsymbol{\theta})\right] = -\int_\Theta \theta(\mathbb{E}_\theta[\tilde{\theta}(\mathbf{X})] - \theta)\mu(d\theta),$$

ce qui prouve que le biais $b(\theta) = \mathbb{E}_\theta[\tilde{\theta}(\mathbf{X})] - \theta$ n'est pas nul μ-presque sûrement.

Chapitre 5

Comparaison d'estimateurs

Introduction

On reste dans le cadre du chapitre précédent, c'est-à-dire celui d'un modèle paramétrique unidimensionnel $(P_\theta)_{\theta\in\Theta}$ où Θ est un intervalle de $\mathbb{R}$. Lorsque plusieurs estimateurs de θ sont disponibles, lequel doit-on choisir ? Plus généralement, existe-t-il un estimateur "optimal", et si oui en quel sens ? Ce chapitre se propose de donner quelques éléments de réponses.

5.1 Principes généraux

5.1.1 Comparaison des risques

Revenons au contexte de théorie de la décision de la Section 4.3 du chapitre précédent, où $\theta \in \Theta$ est un paramètre déterministe, ℓ une fonction de perte et $\mathbf{X}$ l'observation (aléatoire). D'après la Définition 28, on associe à un estimateur $\hat{\theta} = \hat{\theta}(\mathbf{X})$ de θ le risque

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\ell(\theta, \hat{\theta}(\mathbf{X})) \right],$$

où la moyenne se fait par rapport à la loi P_θ de l'observation $\mathbf{X}$. En particulier, pour ce critère, $\hat{\theta}$ sera meilleur que $\tilde{\theta}$ si

$$\forall \theta \in \Theta \qquad R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta}).$$

Ceci est bel et bien, mais comme nous l'avions mentionné, s'il existe θ et θ' tels que $R(\theta, \hat{\theta}) < R(\theta, \tilde{\theta})$ et $R(\theta', \hat{\theta}) > R(\theta', \tilde{\theta})$, on n'est pas plus avancé.

Le point de vue bayésien consiste à mettre une loi μ sur le paramètre θ, dès lors vu comme une variable aléatoire $\boldsymbol{\theta}$, et à comparer les risques de Bayes

$$R_B(\mu, \hat{\theta}(\mathbf{X})) = \mathbb{E}\left[\ell(\boldsymbol{\theta}, \hat{\theta}(\mathbf{X})) \right] \lesseqgtr \mathbb{E}\left[\ell(\boldsymbol{\theta}, \tilde{\theta}(\mathbf{X})) \right] = R_B(\mu, \tilde{\theta}(\mathbf{X})).$$

Cette solution est attrayante, mais elle dépend tout de même de la loi a priori μ sur $\boldsymbol{\theta}$, laquelle peut être sujette à débat...

Oublions le cadre bayésien pour revenir à l'approche fréquentiste et considérons la perte quadratique. Sa décomposition biais carré-variance s'écrit

$$R(\theta, \hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2 \right] = \Big(\mathbb{E}[\hat{\theta}] - \theta \Big)^2 + \mathbb{E}\left[\Big(\hat{\theta} - \mathbb{E}[\hat{\theta}] \Big)^2 \right],$$

et on voit qu'un bon estimateur doit avoir un biais et une variance qui sont **tous deux** petits.

5.1.2 Quelques mots sur le biais

Dans la plupart des cas, nonobstant une idée largement répandue, le non-biais d'un estimateur ne saurait être l'objet d'une attention démesurée. Donnons quelques arguments pour étayer ce point de vue.

Absence d'estimateur non biaisé

Dans certaines situations, ce n'est même pas la peine de se creuser la tête, il n'existe tout bonnement aucun estimateur sans biais. On observe $\mathbf{X}$ suivant une loi binomiale $\mathcal{B}(n, 1/\lambda)$, où n est connu et $\lambda > 1$ est le paramètre que l'on cherche à estimer. Supposons que $\hat{\lambda} = \hat{\lambda}(\mathbf{X})$ soit un estimateur sans biais de λ. Alors, pour tout $\lambda > 1$, on aurait

$$\lambda = \mathbb{E}[\hat{\lambda}(\mathbf{X})] = \sum_{k=0}^{n} \binom{n}{k} \lambda^{-k} \left(1 - \frac{1}{\lambda}\right)^{n-k} \hat{\lambda}(k).$$

Dans cette écriture, les $\hat{\lambda}(k)$ ne sont rien de plus que des coefficients réels dépendant de k et indépendants de λ. L'équation précédente est équivalente à dire que, pour tout $\lambda > 1$,

$$\lambda^{n+1} - \sum_{k=0}^{n} \binom{n}{k} \hat{\lambda}(k)(\lambda - 1)^{n-k} = 0.$$

Un polynôme de degré exactement $(n+1)$ ne pouvant avoir plus de $(n+1)$ racines, ceci est absurde ! Il n'existe donc aucun estimateur sans biais pour ce problème.

Manque de stabilité

Supposons que $\hat{\theta} = \hat{\theta}(\mathbf{X})$ soit un estimateur non biaisé de θ et φ une fonction. Hormis lorsque φ est affine, il n'y a en général aucune raison pour que $\mathbb{E}[\varphi(\hat{\theta})] = \varphi(\mathbb{E}[\hat{\theta}]) = \varphi(\theta)$, donc en général l'absence de biais n'est pas préservé par transformation. Ceci est limpide lorsque φ est strictement convexe (ou concave), car l'inégalité de Jensen impose alors [1]

$$\mathbb{E}[\varphi(\hat{\theta})] > \varphi(\mathbb{E}[\hat{\theta}]) = \varphi(\theta),$$

donc l'estimateur $\varphi(\hat{\theta})$ est biaisé, alors que $\hat{\theta}$ ne l'était pas.

Estimateur bayésien

Comme mentionné en toute fin de chapitre précédent, si on considère un estimateur bayésien $\hat{\theta}$ pour le risque quadratique, alors en général il est biaisé. Ceci ne l'empêche pas d'être optimal au sens du risque bayésien.

L'histoire du débiaisage

Supposons qu'on dispose d'un estimateur biaisé mais que ce biais est facilement rectifiable. Est-ce la meilleure chose à faire pour autant ? Pas forcément... Revenons à l'exemple d'une loi uniforme sur $[0, \theta]$ vu au chapitre précédent, Section 4.2.2. L'estimateur du maximum de vraisemblance était $\hat{\theta} = X_{(n)}$, qui présentait un biais puisque $\mathbb{E}[\hat{\theta}] = (n\theta)/(n+1)$. Par ailleurs nous avions vu que

$$\mathbb{E}[\hat{\theta}^2] = \frac{n}{n+2}\theta^2 \implies R(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \frac{2\theta^2}{(n+1)(n+2)}.$$

1. si $\hat{\theta}$ n'est pas constant, mais cette situation serait sans intérêt.

Considérons l'estimateur débiaisé $\tilde{\theta} = (n+1)X_{(n)}/n$, alors

$$\mathbb{E}[\tilde{\theta}^2] = \frac{(n+1)^2}{n(n+2)}\theta^2 \implies R(\tilde{\theta},\theta) = \mathrm{Var}(\tilde{\theta}) = \mathbb{E}[\tilde{\theta}^2] - \theta^2 = \frac{\theta^2}{n(n+2)}.$$

On en déduit que $R(\tilde{\theta},\theta) \leq R(\hat{\theta},\theta)$, donc le débiaisage a amélioré les choses en terme de risque quadratique. Néanmoins, on peut faire encore mieux. En effet, considérons de façon plus générale un estimateur de la forme $\alpha X_{(n)}$, où α est un réel. Son erreur quadratique s'écrit donc

$$R(\alpha X_{(n)},\theta) = \mathbb{E}\left[(\alpha X_{(n)} - \theta)^2\right] = \theta^2\left(\frac{n}{n+2}\alpha^2 - \frac{2n}{n+1}\alpha + 1\right).$$

Ce trinôme en α est minimal pour $\alpha = (n+2)/(n+1)$. En terme de risque quadratique, l'estimateur biaisé $(n+2)X_{(n)}/(n+1)$ est donc meilleur que l'estimateur non biaisé $\tilde{\theta}$.

Biais et parallélisation

Plaçons-nous du point de vue du risque quadratique. Très souvent [2], les estimateurs que l'on considère sont ou bien non biaisés ou bien biaisés en $\mathcal{O}(1/n)$. Leur variance étant typiquement en $\mathcal{O}(1/n)$, le risque quadratique est lui aussi en $\mathcal{O}(1/n)$. Autrement dit, dès que n est assez grand, même si l'estimateur est biaisé, le biais est "invisible" car masqué par l'écart-type.

Une autre façon de le dire : pour deux estimateurs $\hat{\theta}_n$ et $\tilde{\theta}_n$ avec biais au plus en $\mathcal{O}(1/n)$ et variance en $\mathcal{O}(1/n)$, seules les variances $\sigma_n^2 = \sigma_n^2(\theta)$ et $s_n^2 = s_n^2(\theta)$ importent pour la comparaison. Dès lors, si pour tout $\theta \in \Theta$, $\sigma_n^2(\theta) \leq s_n^2(\theta)$ pour n assez grand, alors on optera pour $\hat{\theta}_n$, au moins asymptotiquement.

Il existe cependant une situation qui peut changer radicalement la donne. Supposons que $\hat{\theta}_n$ présente un biais

$$b_n(\theta) = \mathbb{E}[\hat{\theta}_n] - \theta = \mathcal{O}(1/n),$$

tandis que $\tilde{\theta}_n$ est non biaisé. Supposons que le nombre n de données soit immense mais qu'on dispose aussi d'un très grand nombre de processeurs de façon à pouvoir paralléliser les calculs. Pour simplifier les notations, on va considérer $N = \sqrt{n}$ processeurs, chacun traitant un ensemble de N données. On a donc N estimateurs partiels $\hat{\theta}_N^{(1)}, \ldots, \hat{\theta}_N^{(N)}$ desquels on déduit l'estimateur global par moyennisation

$$\hat{T}_n = \frac{\hat{\theta}_N^{(1)} + \cdots + \hat{\theta}_N^{(N)}}{N}.$$

Les estimateurs partiels étant iid, les propriétés de $\hat{T}_n$ sont immédiates :

$$\mathbb{E}[\hat{T}_n] = b_N(\theta) \text{ et } \mathrm{Var}(\hat{T}_n) = \frac{\sigma_N^2(\theta)}{N} \implies R(\hat{T}_n,\theta) = b_N(\theta)^2 + \frac{\sigma_N^2(\theta)}{N}.$$

Suivant la même démarche, l'estimateur non biaisé $\tilde{\theta}_n$ mène à l'estimateur global

$$\mathbb{E}[\tilde{T}_n] = 0 \text{ et } \mathrm{Var}(\tilde{T}_n) = \frac{s_N^2(\theta)}{N} \implies R(\hat{T}_n,\theta) = \frac{s_N^2(\theta)}{N}.$$

Si $b_N(\theta) = b(\theta)/N$, $\sigma_N^2(\theta) = \sigma^2(\theta)/N$ et $s_N^2(\theta) = s^2(\theta)/N$, alors

$$R(\hat{T}_n,\theta) = \frac{b(\theta)^2 + \sigma^2(\theta)}{n} \qquad \text{et} \qquad R(\tilde{T}_n,\theta) = \frac{s^2(\theta)}{n}.$$

Donc si $b(\theta)^2 + \sigma^2(\theta) > s^2(\theta)$, il faudra désormais privilégier le second estimateur. On voit que la parallélisation des calculs a fait émerger le biais du premier estimateur de façon décisive !

2. mais pas toujours, par exemple l'EMV $X_{(n)}$ pour la loi $\mathcal{U}_{[0,\theta]}$ ne rentre pas dans ce cadre, bref passons.

5.1.3 L'approche asymptotique

Il est souvent plus simple de comparer les choses de façon asymptotique, i.e. lorsque n tend vers l'infini. Le premier critère est bien entendu celui de la vitesse de convergence vers 0. Si $R(\hat{\theta}_n, \theta) = o(R(\tilde{\theta}_n, \theta))$, on optera pour $\hat{\theta}_n$, non pour $\tilde{\theta}_n$.

Exemple. Reprenons l'exemple de la loi uniforme sur $[0, \theta]$, où l'estimateur du maximum de vraisemblance est $\hat{\theta}_n = X_{(n)}$. L'estimateur issu de la méthode des moments est $\tilde{\theta}_n = 2\bar{X}_n$ et a pour risque quadratique $\theta^2/(3n)$. Puisque

$$R(\hat{\theta}_n, \theta) = \frac{2\theta^2}{(n+1)(n+2)} = o\left(\frac{\theta^2}{3n}\right),$$

on préférera l'EMV, et ce malgré son biais.

Cet exemple n'est cependant pas représentatif de la situation typique : en général, les risques quadratiques convergent à vitesse $1/n$ vers 0. Néanmoins, si on dispose pour les estimateurs $\hat{\theta}_n$ et $\tilde{\theta}_n$ de résultats de normalité asymptotique de la forme

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta)) \qquad \text{et} \qquad \sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, s^2(\theta)),$$

avec $\sigma^2(\theta) \leq s^2(\theta)$ pour tout $\theta \in \Theta$, alors on préférera $\hat{\theta}_n$ à $\tilde{\theta}_n$. En effet, en arrondissant 1.96 à 2, on a par exemple

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \leq \frac{2\sigma(\theta)}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 95\% \qquad \text{et} \qquad \mathbb{P}\left(\left|\tilde{\theta}_n - \theta\right| \leq \frac{2s(\theta)}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 95\%$$

donc pour un même niveau de confiance asymptotique, le premier estimateur donne un encadrement plus précis.

A première vue, on n'a fait que reporter le problème, puisque la comparaison des variances asymptotiques soulève les mêmes difficultés que la comparaison des risques quadratiques. On peut en effet très bien imaginer θ et θ' tels que $\sigma^2(\theta) < s^2(\theta)$ et $\sigma^2(\theta') > s^2(\theta')$. Comme nous allons le voir, l'intérêt de la théorie asymptotique est que, sous certaines conditions, il existe une variance asymptotique optimale et des estimateurs atteignant celle-ci [3].

Nota Bene. La normalité asymptotique ne permet pas de contrôler le risque quadratique. Dans le modèle des lois de Poisson $\mathcal{P}(1/\theta)$, $\theta > 0$, l'estimateur $\hat{\theta}_n = 1/\bar{X}_n$ est asymptotiquement normal (Delta méthode), mais de risque quadratique infini puisque $\mathbb{P}(\bar{X}_n = 0) > 0$.

5.2 Exhaustivité

Notre but est d'exhiber de "bons" estimateurs. Avant de chercher à les trouver, on va trouver où les chercher. C'est en ce sens qu'intervient la notion d'exhaustivité.

5.2.1 Statistique exhaustive et factorisation

Revenons au Pile ou Face parcouru en long, en large et en travers au Chapitre 1. Le modèle statistique est le suivant : $E = \{0,1\}^n$ est l'espace des observations $\mathbf{X} = (X_1, \ldots, X_n)$ sur lequel on considère la famille de lois

$$(P_\theta)_{\theta\in\Theta} = \left(\mathcal{B}(\theta)^{\otimes n}\right)_{\theta\in]0,1[}.$$

3. Tuons le suspense : la variance optimale sera l'inverse de l'information de Fisher, asymptotiquement atteinte par l'estimateur du maximum de vraisemblance (sous les hypothèses idoines).

Toute réalisation $\mathbf{x} = (x_1, \dots, x_n)$ de $\mathbf{X}$ a sous P_θ la probabilité

$$P_\theta(\mathbf{x}) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n\bar{x}_n}(1-\theta)^{n-n\bar{x}_n},$$

où $\bar{x}_n = (x_1 + \cdots + x_n)/n$ est la fréquence empirique de Pile dans le n-uplet $\mathbf{x} = (x_1, \dots, x_n)$. Le n-uplet aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ peut en fait être résumé par la moyenne empirique $\bar{X}_n$ sans perte d'information sur le paramètre inconnu θ. En effet, on a pour tout $\mathbf{x} = (x_1, \dots, x_n) \in E$,

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n | \bar{X}_n = \bar{x}_n) = \frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}(\bar{X}_n = \bar{x}_n)},$$

et puisque $n\bar{X}_n$ suit une loi binomiale $\mathcal{B}(n, \theta)$,

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n | \bar{X}_n = \bar{x}_n) = \frac{\theta^{n\bar{x}_n}(1-\theta)^{n-n\bar{x}_n}}{\binom{n}{n\bar{x}_n}\theta^{n\bar{x}_n}(1-\theta)^{n-n\bar{x}_n}} = \frac{1}{\binom{n}{n\bar{x}_n}}.$$

La loi conditionnelle de $\mathbf{X} = (X_1, \dots, X_n)$ sachant $\bar{X}_n$ est donc indépendante du paramètre θ. C'est encore dire que toute l'information sur θ contenue dans l'échantillon $\mathbf{X} = (X_1, \dots, X_n)$ est en fait contenue dans $\bar{X}_n$. On dit que $\bar{X}_n$ est une statistique exhaustive.

Définition 31 (Statistique exhaustive)
La statistique $T(\mathbf{X})$ est dite exhaustive si, pour tout $\theta \in \Theta$, la loi conditionnelle de $\mathbf{X} = (X_1, \dots, X_n)$ sachant $T(\mathbf{X})$ ne dépend pas de θ.

Dit formellement, pour tout ensemble mesurable A de $(E, \mathcal{E})$, il existe une fonction mesurable φ, dépendant de A mais pas de θ, telle que

$$\forall \theta \in \Theta \qquad \mathbb{P}_\theta(\mathbf{X} \in A | T(\mathbf{X})) = \varphi(T(\mathbf{X})) \qquad \mathbb{P}_\theta - p.s.$$

Remarques :

1. L'intérêt d'une statistique exhaustive est clair sur l'exemple du Pile ou Face : on a réduit un n-uplet aléatoire à une seule variable aléatoire. Ainsi, on a grandement résumé l'aléa sans perte d'information sur ce qui nous intéresse.
2. Une statistique exhaustive peut prendre ses valeurs dans $\mathbb{R}^q$ avec $q > 1$. En particulier, il peut s'agir d'un vecteur aléatoire.
3. Par exemple, l'observation $\mathbf{X} = (X_1, \dots, X_n)$ est une statistique exhaustive puisque la loi conditionnelle de $(X_1, \dots, X_n)$ sachant $(X_1, \dots, X_n)$ est tout bonnement la mesure de Dirac en $(X_1, \dots, X_n)$, qui est bel et bien indépendante de θ. Bien entendu, elle n'a aucun intérêt puisqu'elle ne résume rien.

Le résultat suivant, admis, offre une façon simple d'établir l'exhaustivité d'une statistique dans le cas usuel des modèles dominés.

Théorème 16 (Critère de factorisation de Neyman-Fisher)
Soit un modèle statistique dominé défini par la famille de densités $(g_\theta)_{\theta \in \Theta}$ par rapport à la mesure de référence ν sur E[4]*. Alors la statistique $T(\mathbf{X})$ est exhaustive si et seulement si il existe deux fonctions h et ψ indépendantes de θ telles que, pour tout $\theta \in \Theta$,*

$$g_\theta(\mathbf{x}) = h(\mathbf{x})\ \psi(T(\mathbf{x}), \theta) \qquad \nu - p.p.$$

4. Rappel : ceci veut dire que $g_\theta = dP_\theta/d\nu$.

Exemples :

1. Dans le Pile ou Face,

$$T(\mathbf{x}) = T(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n} = \bar{x}_n$$

est une statistique exhaustive puisque

$$g_\theta(\mathbf{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \theta^{n\bar{x}_n}(1-\theta)^{n-n\bar{x}_n}.$$

Sur cet exemple, h est constante égale à 1 et

$$\psi(T(\mathbf{x}), \theta) = \theta^{nT(\mathbf{x})}(1-\theta)^{n-nT(\mathbf{x})}.$$

2. Soit $(X_1, \ldots, X_n)$ iid selon une loi de Poisson de paramètre $\lambda > 0$, alors

$$T(\mathbf{x}) = T(x_1, \ldots, x_n) = x_1 + \cdots + x_n$$

est une statistique exhaustive puisque, avec des notations évidentes,

$$g_\theta(\mathbf{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{1}{x_1! \ldots x_n!} \times \left(\lambda^{T(\mathbf{x})} e^{-n\lambda}\right) = h(\mathbf{x})\ \psi(T(\mathbf{x}), \lambda).$$

Remarques :

1. Dans les deux exemples précédents, on peut prendre indifféremment $T(\mathbf{X}) = X_1 + \cdots + X_n$ ou $T(\mathbf{X}) = \bar{X}_n$ comme statistiques exhaustives. De façon générale, il n'y a pas unicité, on parlera donc d'**une** statistique exhaustive, non de **la** statistique exhaustive.
2. On se souvient que l'estimation au maximum de vraisemblance consiste à maximiser la fonction $L_n(\theta) = g_\theta(\mathbf{X})$ ou son logarithme. Puisque cette optimisation se fait en θ, si on dispose d'une statistique exhaustive, alors le critère de factorisation de Neyman-Fisher assure que

$$\operatorname*{argmax}_{\theta\in\Theta} L_n(\theta) = \operatorname*{argmax}_{\theta\in\Theta}(h(\mathbf{X})\ \psi(T(\mathbf{X}), \theta)) = \operatorname*{argmax}_{\theta\in\Theta} \psi(T(\mathbf{X}), \theta).$$

Ceci signifie que l'estimateur du max de vraisemblance, lorsqu'il existe, ne dépend que de la statistique exhaustive. Ou encore : on a seulement besoin de connaître celle-ci pour estimer θ au max de vraisemblance.

5.2.2 Applications de l'exhaustivité

La notion d'exhaustivité permet d'améliorer la qualité d'un estimateur, au moins en terme de risque quadratique.

Théorème 17 (Rao-Blackwell)
Soit $T(\mathbf{X})$ une statistique exhaustive et $\hat{\theta}(\mathbf{X})$ un estimateur de θ. Alors $\tilde{\theta}(\mathbf{X}) = \mathbb{E}[\hat{\theta}(\mathbf{X})|T(\mathbf{X})]$ est un estimateur de même biais que $\hat{\theta}(\mathbf{X})$ et de risque quadratique inférieur :

$$\forall \theta \in \Theta \qquad R(\theta, \tilde{\theta}(\mathbf{X})) = \mathbb{E}\left[\left(\tilde{\theta}(\mathbf{X}) - \theta\right)^2\right] \leq \mathbb{E}\left[\left(\hat{\theta}(\mathbf{X}) - \theta\right)^2\right] = R(\theta, \hat{\theta}(\mathbf{X})).$$

Preuve. Notons tout d'abord que $\tilde{\theta}(\mathbf{X})$ est bien un estimateur, c'est-à-dire qu'il ne dépend pas de θ : ceci découle de l'exhaustivité de $T(\mathbf{X})$. De plus, par la propriété classique de l'espérance conditionnelle :

$$\mathbb{E}[\tilde{\theta}(\mathbf{X})] = \mathbb{E}[\mathbb{E}[\hat{\theta}(\mathbf{X})|T(\mathbf{X})]] = \mathbb{E}[\hat{\theta}(\mathbf{X})],$$

et les deux estimateurs ont même biais. Venons-en au risque quadratique et commençons par la décomposition :

$$\Big(\hat{\theta}(\mathbf{X}) - \theta\Big)^2 = \Big(\hat{\theta}(\mathbf{X}) - \tilde{\theta}(\mathbf{X})\Big)^2 + 2\Big(\hat{\theta}(\mathbf{X}) - \tilde{\theta}(\mathbf{X})\Big)\Big(\tilde{\theta}(\mathbf{X}) - \theta\Big) + \Big(\tilde{\theta}(\mathbf{X}) - \theta\Big)^2.$$

Or $\tilde{\theta}(\mathbf{X}) = \mathbb{E}[\hat{\theta}(\mathbf{X})|T(\mathbf{X})]$ étant une fonction de $T(\mathbf{X})$, il vient

$$\mathbb{E}\left[\Big(\hat{\theta}(\mathbf{X}) - \tilde{\theta}(\mathbf{X})\Big)\Big(\tilde{\theta}(\mathbf{X}) - \theta\Big)\middle| T(\mathbf{X})\right] = \Big(\tilde{\theta}(\mathbf{X}) - \theta\Big)\,\mathbb{E}\left[\Big(\hat{\theta}(\mathbf{X}) - \tilde{\theta}(\mathbf{X})\Big)\middle| T(\mathbf{X})\right] = 0.$$

Par conséquent

$$\mathbb{E}\left[\Big(\hat{\theta}(\mathbf{X}) - \theta\Big)^2\middle| T(\mathbf{X})\right] = \mathbb{E}\left[\Big(\hat{\theta}(\mathbf{X}) - \tilde{\theta}(\mathbf{X})\Big)^2\middle| T(\mathbf{X})\right] + \mathbb{E}\left[\Big(\tilde{\theta}(\mathbf{X}) - \theta\Big)^2\middle| T(\mathbf{X})\right],$$

et la conclusion suit en prenant l'espérance.

■

Remarque. La preuve montre que l'inégalité ne peut être une égalité que si les deux estimateurs coïncident presque sûrement. Par ailleurs, le résultat de Rao-Blackwell se généralise sans problème à toute fonction de perte convexe[5], par exemple $\ell(\theta, t) = |t - \theta|^p$ avec $p \geq 1$.

Morale de cette histoire : si l'on dispose d'une statistique exhaustive et d'un estimateur, on a tout intérêt, si ce n'est déjà fait, à conditionner celui-ci par celle-là pour diminuer le risque. Ce procédé est parfois appelé Rao-Blackwellisation. On le rencontre sous des formes variées, c'est par exemple une méthode classique de réduction de variance en simulations Monte-Carlo.

Encore faut-il être capable de trouver facilement une statistique exhaustive aussi élémentaire que possible. Pour la plupart des exemples croisés jusqu'à présent, c'est bien le cas. Nous avons vu qu'un grand nombre de lois rentrent dans le modèle exponentiel, ce sera notre premier exemple. Nous avons vu aussi que les lois uniformes n'en font pas partie, ce sera donc notre second exemple.

Exemples :

1. Modèle exponentiel général : considérons donc $(X_1, \ldots, X_n)$ iid de densité $f_\theta(x)$ par rapport à une mesure de référence μ sur $\mathbb{R}$, avec (cf. Définition 27)

   $$f_\theta(x) = C(\theta)h(x)\exp(Q(\theta)T(x)).$$

   La densité jointe s'écrit donc

   $$g_\theta(x_1, \ldots, x_n) = \left(\prod_{i=1}^n h(x_i)\right) \times \left(C(\theta)^n \exp\left(Q(\theta)\sum_{i=1}^n T(x_i)\right)\right).$$

   La critère de factorisation de Neyman-Fisher est ainsi satisfait et

   $$T(\mathbf{X}) = \sum_{i=1}^n T(X_i)$$

   représente une statistique exhaustive. Les exemples de Pile ou Face et de la loi de Poisson vus avant n'en étaient à vrai dire que des cas particuliers.

5. il suffit d'appliquer l'inégalité de Jensen pour les espérances conditionnelles.

2. Modèles uniformes : commençons par la loi uniforme sur $[0,\theta]$. La densité jointe s'écrit

$$g_\theta(x_1,\dots,x_n) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\times \mathbb{1}_{[x_{(n)},+\infty[}(\theta).$$

Ainsi, $T(\mathbf{X}) = X_{(n)}$ est une statistique exhaustive. Si on considère le modèle uniforme sur $[-\theta,\theta]$, on a cette fois

$$g_\theta(x_1,\dots,x_n) = \frac{1}{(2\theta)^n}\prod_{i=1}^n \mathbb{1}_{[-\theta,\theta]}(x_i) = \frac{1}{(2\theta)^n}\times \mathbb{1}_{[\max(-x_{(1)},x_{(n)}),+\infty[}(\theta),$$

et la variable $T(\mathbf{X}) = \max(-X_{(1)}, X_{(n)})$ est une statistique exhaustive. Pour le modèle uniforme sur $[\theta,\theta+1]$,

$$g_\theta(x_1,\dots,x_n) = \prod_{i=1}^n \mathbb{1}_{[\theta,\theta+1]}(x_i) = \mathbb{1}_{]-\infty,x_{(1)}]}(\theta)\times \mathbb{1}_{[x_{(n)}-1,+\infty[}(\theta) = \mathbb{1}_{[x_{(n)}-1,x_{(1)}]}(\theta),$$

et le couple $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ est une statistique exhaustive.

5.3 Information de Fisher

In fine, notre objectif est de préciser ce que l'on peut attendre au mieux d'un estimateur de θ. Un critère d'optimalité est spécifié par l'information de Fisher. Pour préciser cette notion, il faut cependant commencer par circonscrire la classe des modèles sur lesquels on travaille.

Sans même rentrer dans les détails techniques, ceci n'a rien d'étonnant : dans la plupart des exemples croisés jusqu'ici, les estimateurs sont asymptotiquement normaux et de risque quadratique en $1/n$. Un cas très particulier est celui de l'estimateur du maximum de vraisemblance pour le modèle uniforme $(\mathcal{U}_{[0,\theta]})_{\theta>0}$, c'est-à-dire $X_{(n)}$: il n'est pas asymptotiquement normal et son risque quadratique est en $1/n^2$. Bref, il est tout à fait atypique et nous allons préciser en quel sens, à savoir qu'il n'est pas régulier.

5.3.1 Modèles réguliers et information de Fisher

Dans tout ce qui suit, nous considérons sur E un modèle statistique dominé de la forme $(P_\theta)_{\theta\in\Theta} = (g_\theta\cdot\mu)_{\theta\in\Theta}$ où Θ est un intervalle **ouvert**[6] de $\mathbb{R}$ et μ une mesure de référence. Par ailleurs, les symboles de dérivation le seront toujours par rapport au paramètre θ, c'est-à-dire que, sous réserve d'existence, nous noterons pour $\mathbf{x}\in E$ et $\theta\in\Theta$:

$$g'_\theta(\mathbf{x}) = \frac{\partial}{\partial\theta}g_\theta(\mathbf{x}) \qquad \text{et} \qquad g''_\theta(\mathbf{x}) = \frac{\partial^2}{\partial\theta^2}g_\theta(\mathbf{x}).$$

Il existe plusieurs façons de définir un modèle régulier. Celle que nous proposons n'est pas la plus classique, mais présente l'avantage d'être très générale.

Définition 32 (Modèle régulier et information de Fisher)
Le modèle $(P_\theta)_{\theta\in\Theta} = (g_\theta\cdot\mu)_{\theta\in\Theta}$ est dit régulier si :

— *pour μ presque tout $\mathbf{x}$, l'application $\theta\mapsto g_\theta(\mathbf{x})$ est continue sur Θ et C^1 sauf éventuellement en un nombre fini de points ;*

6. On veut éviter les phénomènes de bord de type "demi-gaussienne" comme dans l'exemple de la Section 4.2.3.

— *pour tout $\theta \in \Theta$, l'application*

$$\mathbf{x} \mapsto \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\ \mathbb{1}_{g_\theta(\mathbf{x})>0}$$

est intégrable sur E par rapport à la mesure de référence μ et d'intégrale

$$I(\theta) = \int_E \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\ \mathbb{1}_{g_\theta(\mathbf{x})>0}\ \mu(d\mathbf{x})$$

continue sur Θ.

La quantité $I(\theta)$ est alors appelée information de Fisher du modèle.

Ainsi, pour qu'un modèle soit régulier, la fonction $(\theta, \mathbf{x}) \mapsto g_\theta(\mathbf{x})$ doit respecter une condition de continuité/dérivabilité par rapport à θ, et une condition d'intégrabilité par rapport à $\mathbf{x}$. Par ailleurs, si elle existe, il est clair que l'information de Fisher est toujours supérieure ou égale à 0.

Exemples :

1. Loi exponentielle : considérons $\mathbf{X} \sim \mathcal{E}(\theta)$ avec $\theta \in \Theta =]0, +\infty[$, alors

$$g_\theta(\mathbf{x}) = \theta e^{-\theta \mathbf{x}} \qquad \text{et} \qquad \mu(d\mathbf{x}) = \mathbb{1}_{\mathbf{x}\geq 0}\ d\mathbf{x},$$

donc :
— pour tout $\mathbf{x} \geq 0$, l'application $\theta \mapsto g_\theta(\mathbf{x})$ est C^∞ sur Θ ;
— pour tout $\theta > 0$, l'application

$$\mathbf{x} \mapsto \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\ \mathbb{1}_{g_\theta(\mathbf{x})>0} = \frac{(1-\theta\mathbf{x})^2}{\theta} e^{-\theta\mathbf{x}}$$

est intégrable sur $\mathbb{R}_+$ par rapport à la mesure de Lebesgue, d'intégrale (en se souvenant que $\mathbb{E}[\mathbf{X}^k] = k!/\theta^k$ pour tout $k \in \mathbb{N}$) :

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

continue sur $\Theta =]0, +\infty[$. Ainsi le modèle défini par ces lois exponentielles est bien régulier.

2. Loi de Bernoulli : soit $\mathbf{X} \sim \mathcal{B}(\theta)$ avec $\theta \in \Theta =]0, 1[$, alors $\mu(d\mathbf{x}) = \delta_0 + \delta_1$ est la mesure de comptage sur $\{0, 1\}$, avec

$$g_\theta(0) = 1 - \theta \qquad \text{et} \qquad g_\theta(1) = \theta,$$

donc :
— pour tout $\mathbf{x} \in \{0, 1\}$, l'application $\theta \mapsto g_\theta(\mathbf{x})$ est C^∞ sur Θ ;
— pour tout $\theta \in \Theta$,

$$\frac{g'_\theta(0)^2}{g_\theta(0)} = \frac{1}{1-\theta} \qquad \text{et} \qquad \frac{g'_\theta(1)^2}{g_\theta(1)} = \frac{1}{\theta}.$$

Ainsi

$$I(\theta) = \frac{1}{1-\theta} + \frac{1}{\theta} = \frac{1}{\theta(1-\theta)},$$

qui est continue sur $\Theta =]0, 1[$, donc ce modèle est régulier.

3. Loi uniforme : supposons maintenant que $\mathbf{X} \sim \mathcal{U}_{[0,\theta]}$ avec $\theta \in \Theta =]0, +\infty[$. Pour tout réel $\mathbf{x} \geq 0$ (fixé !), la fonction

$$\theta \mapsto g_\theta(\mathbf{x}) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(\mathbf{x}) = \frac{1}{\theta}\mathbb{1}_{[\mathbf{x},+\infty[}(\theta)$$

est discontinue au point $\mathbf{x}$, donc ce modèle n'est pas régulier. Ceci est en accord avec ce que nous avons annoncé en préambule. Par conséquent, rien de ce qui suit ne s'appliquera à ce modèle.

Puisque

$$I(\theta) = \int_{g_\theta(\mathbf{x})>0} \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})^2}\, g_\theta(\mathbf{x})\mu(d\mathbf{x}) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log g_\theta(\mathbf{X})\right)^2\right],$$

on peut donner une autre formulation de l'information de Fisher.

Définition 33 (Score et information de Fisher)
Dans un modèle régulier, si on note[7] $\ell_\theta(\mathbf{X}) = \log g_\theta(\mathbf{X})$ *la log-vraisemblance, on a*

$$I(\theta) = \mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right], \tag{5.1}$$

où la variable aléatoire

$$\ell'_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\log g_\theta(\mathbf{X}) = \frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}$$

est appelée le score.

On va maintenant donner un résultat de dérivation sous le signe somme. Au préalable, précisons qu'une application $\theta \mapsto \varphi(\theta)$ est localement bornée si

$$\forall\theta_0 \in \Theta,\ \exists\varepsilon = \varepsilon(\theta_0) > 0 \qquad \sup_{\theta_0-\varepsilon<\theta<\theta_0+\varepsilon} |\varphi(\theta)| < +\infty.$$

Clairement, une fonction continue sur Θ est localement bornée. Une fonction bornée sur Θ est a fortiori localement bornée, la réciproque étant fausse : il suffit de considérer $\varphi(\theta) = \theta$ sur $\Theta = \mathbb{R}$. Pour tomber sur une fonction non localement bornée, il faut le faire un peu exprès : c'est par exemple le cas de la fonction définie sur $\mathbb{R}$ par $\varphi(0) = 0$ et $\varphi(\theta) = 1/\theta$ si $\theta \neq 0$, laquelle n'est pas localement bornée à l'origine.

Bref, pour la suite, on retiendra que l'hypothèse "telle fonction est localement bornée" n'est pas bien contraignante. Sa raison d'être est de permettre la dérivation sous le signe somme, comme dans le résultat suivant (admis).

Proposition 21 (Dérivation sous le signe somme)
Soit un modèle régulier sur Θ *et* $T(\mathbf{X})$ *une statistique telle que la fonction* $\theta \mapsto \mathbb{E}_\theta[T(\mathbf{X})^2]$ *soit localement bornée, alors l'application* $\theta \mapsto \mathbb{E}_\theta[T(\mathbf{X})]$ *est* C^1 *de dérivée*

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(\mathbf{X})] = \frac{\partial}{\partial\theta}\int_E T(\mathbf{x})g_\theta(\mathbf{x})\mu(d\mathbf{x}) = \int_E T(\mathbf{x})g'_\theta(\mathbf{x})\mu(d\mathbf{x}) = \mathbb{E}_\theta\left[T(\mathbf{X})\frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right] = \mathbb{E}_\theta[T(\mathbf{X})\ell'_\theta(\mathbf{X})].$$

Autrement dit, on peut dériver sous le signe somme.

Dans la relation précédente, le cas particulier $T(\mathbf{X}) = 1$ donne une nouvelle formule pour l'information de Fisher.

7. Lorsque $\mathbf{X} = (X_1, \ldots, X_n)$, nous notions précédemment $\ell_n(\theta)$ cette log-vraisemblance.

Corollaire 2 (Information de Fisher et variance du score)
Si le modèle est régulier, alors en notant $\ell_\theta(\mathbf{X}) = \log g_\theta(\mathbf{X})$ la log-vraisemblance, on a

$$I(\theta) = \mathrm{Var}_\theta(\ell'_\theta(\mathbf{X})),$$

c'est-à-dire que l'information de Fisher est égale à la variance du score.

Preuve. Prenons $T(\mathbf{X}) = 1$ dans la Proposition 21, alors $\theta \mapsto \mathbb{E}_\theta[T(\mathbf{X})^2] = 1$ est bien localement bornée et

$$0 = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[1] = \mathbb{E}_\theta\left[\frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right] = \mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})\right].$$

D'où l'on déduit, en partant de l'équation (5.1),

$$I(\theta) = \mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] = \mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] - \left(\mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})\right]\right)^2 = \mathrm{Var}_\theta(\ell'_\theta(\mathbf{X})).$$

■

On peut donner une nouvelle formulation de l'information de Fisher, mais elle nécessite des hypothèses supplémentaires. Nous dirons qu'une famille de fonctions $\varphi_\theta(\mathbf{x})$ intégrables par rapport à $\mathbf{x}$ pour la mesure μ est localement dominée dans $L_1(\mu)$ si

$$\forall\theta_0 \in \Theta,\ \exists\varepsilon = \varepsilon(\theta_0) > 0 \qquad \sup_{\theta_0-\varepsilon<\theta<\theta_0+\varepsilon} |\varphi_\theta(\mathbf{x})| \leq \psi(\mathbf{x}) \in L_1(\mu).$$

Ce qu'on a en tête avec ce genre d'hypothèse est clair : pouvoir appliquer les résultats de continuité et de dérivabilité de Lebesgue. Une façon "classique" de définir un modèle régulier est la suivante.

Lemme 1 (Version plus forte de la régularité)
Supposons les hypothèses suivantes :
- *l'ensemble $S = \{\mathbf{x} \in E,\ g_\theta(\mathbf{x}) > 0\}$ est indépendant de θ ;*
- *pour μ presque tout $\mathbf{x}$, l'application $\theta \mapsto g_\theta(\mathbf{x})$ est C^1 sur Θ ;*
- *la famille $(g'_\theta)^2/g_\theta$ est localement dominée dans $L_1(\mu)$.*

Alors le modèle est régulier au sens de la Définition 32.

Preuve. Considérons

$$E' = \{\mathbf{x} \in E,\ g_\theta(\mathbf{x}) > 0\} \cap \{\mathbf{x} \in E, \theta \mapsto g_\theta(\mathbf{x}) \text{ est } C^1 \text{ sur } \Theta\}.$$

Les deux premières hypothèses assurent que, dans la Définition 32, on peut remplacer E par E' et μ par $\mathbb{1}_{E'} \cdot \mu$. Ceci fait de $g_\theta(\mathbf{x})$ une application strictement positive et C^1 en θ. On peut alors appliquer le théorème de continuité de Lebesgue à la fonction

$$I(\theta) = \int_{E'} \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\,\mu(d\mathbf{x}).$$

En tout point θ_0 de Θ, la fonction $\theta \mapsto g'_\theta(\mathbf{x})^2/g_\theta(\mathbf{x})$ est continue. De plus, il existe un voisinage $]\theta_0 - \varepsilon, \theta_0 + \varepsilon[$ tel que

$$0 \leq \sup_{\theta_0-\varepsilon<\theta<\theta_0+\varepsilon} \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})} \leq \psi(\mathbf{x}),$$

avec $\psi \in L_1(\mu)$. Ceci assure que I est continue en θ_0. Celui-ci étant arbitraire, la fonction I est continue sur Θ.

■

En ajoutant une hypothèse du même tonneau, on aboutit à une nouvelle expression pour l'information de Fisher.

Proposition 22 (Information de Fisher et dérivée seconde)
Conservons les hypothèses du Lemme 1 et supposons de plus que :
— *pour μ presque tout $\mathbf{x}$, l'application $\theta \mapsto g_\theta(\mathbf{x})$ est C^2 sur Θ ;*
— *la famille g''_θ est localement dominée dans $L_1(\mu)$.*
Alors l'information de Fisher s'écrit encore

$$I(\theta) = -\mathbb{E}_\theta\left[\ell''_\theta(\mathbf{X})\right].$$

Preuve. On commence par noter que, pour μ presque tout $\mathbf{x}$,

$$\ell''_\theta(\mathbf{x}) = (\log g_\theta(\mathbf{x}))'' = \frac{g''_\theta(\mathbf{x})}{g_\theta(\mathbf{x})} - \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})^2}.$$

Or on a vu en (5.1) que

$$I(\theta) = \mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] = \mathbb{E}_\theta\left[\frac{g'_\theta(\mathbf{X})^2}{g_\theta(\mathbf{X})^2}\right].$$

Pour l'autre terme, il vient

$$\mathbb{E}_\theta\left[\frac{g''_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right] = \int_{E'} g''_\theta(\mathbf{x})\mu(d\mathbf{x}).$$

Soit $\mathbf{x} \in E'$ fixé. En tout point θ_0 de Θ, la fonction $\theta \mapsto \varphi_\theta(\mathbf{x}) = g'_\theta(\mathbf{x})$ est dérivable, de dérivée $g''_\theta(\mathbf{x})$. De plus, par hypothèse, il existe un voisinage $]\theta_0 - \varepsilon, \theta_0 + \varepsilon[$ tel que

$$\sup_{\theta_0-\varepsilon<\theta<\theta_0+\varepsilon} |g''_\theta(\mathbf{x})| \leq \psi(\mathbf{x}),$$

avec $\psi \in L_1(\mu)$. Le théorème de dérivabilité de Lebesgue implique donc que la fonction Φ définie sur Θ par

$$\Phi(\theta) = \int_{E'} \varphi_\theta(\mathbf{x})\mu(d\mathbf{x})$$

est dérivable en θ_0, de dérivée

$$\Phi'(\theta_0) = \int_{E'} g''_\theta(\mathbf{x})\mu(d\mathbf{x}).$$

Ainsi Φ est dérivable sur Θ, de dérivée

$$\Phi'(\theta) = \mathbb{E}_\theta\left[\frac{g''_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right].$$

Or, comme on l'a vu dans la preuve du Corollaire 2, Φ est identiquement nulle sur Θ, donc il en va de même pour sa dérivée.

■

Nous allons maintenant donner quelques propriétés de l'information de Fisher. La première d'entre elles concerne la mesure dominante μ, laquelle n'a aucune importance.

Lemme 2 (Information de Fisher et mesure dominante)
Soit $(P_\theta)_{\theta\in\Theta}$ un modèle dominé. La régularité de ce modèle et la valeur de l'information de Fisher ne dépendent pas de la mesure dominante choisie.

Preuve. Considérons deux mesures dominantes μ et ν, de sorte que

$$g_\theta(\mathbf{x}) = \frac{dP_\theta}{d\mu}(\mathbf{x}) \qquad \text{et} \qquad h_\theta(\mathbf{x}) = \frac{dP_\theta}{d\nu}(\mathbf{x}).$$

La mesure $\lambda = \mu + \nu$ dominant à la fois μ et ν, on peut définir la densité de μ par rapport à λ, que l'on convient de noter

$$\varphi(\mathbf{x}) = \frac{d\mu}{d\lambda}(\mathbf{x}) \implies \frac{dP_\theta}{d\lambda}(\mathbf{x}) = g_\theta(\mathbf{x})\varphi(\mathbf{x}) =: k_\theta(\mathbf{x}).$$

Comme φ ne dépend pas de θ, la régularité en θ de k_θ est la même que celle de g_θ. Quant à l'intégration par rapport à $\mathbf{x}$,

$$\int_E \frac{k'_\theta(\mathbf{x})^2}{k_\theta(\mathbf{x})}\, \mathbb{1}_{k_\theta(\mathbf{x})>0}\, \lambda(d\mathbf{x}) = \int_E \frac{g'_\theta(\mathbf{x})^2\varphi(\mathbf{x})^2}{g_\theta(\mathbf{x})\varphi(\mathbf{x})}\, \mathbb{1}_{k_\theta(\mathbf{x})>0}\, \lambda(d\mathbf{x}) = \int_E \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\, \mathbb{1}_{g_\theta(\mathbf{x})>0}\, \mu(d\mathbf{x}),$$

et l'information de Fisher est la même dans les deux cas. Le raisonnement valant aussi entre ν et λ, le débat est clos.

■

Si l'information de Fisher n'est pas sensible au changement de mesure dominante, elle l'est par contre au changement de paramètre.

Proposition 23 (Information de Fisher et paramétrage)
Soit $(g_\theta)_{\theta\in\Theta}$ un modèle régulier d'information de Fisher $I(\theta)$ et $\eta = \varphi(\theta)$ un changement de paramètre bijectif tel que $\psi = \varphi^{-1}$ soit C^1. Alors le modèle paramétré par η est encore régulier, d'information de Fisher

$$J(\eta) = \psi'(\eta)^2 I(\psi(\eta)).$$

Preuve. Notons $h_\eta(\mathbf{x}) = g_{\psi(\eta)}(\mathbf{x})$. Le modèle initial étant régulier et ψ étant C^1, on en déduit que ϕ est elle-même continue bijective, et que pour μ presque tout $\mathbf{x}$, la fonction $\eta \mapsto h_\eta(\mathbf{x})$ est continue sur l'intervalle ouvert $\varphi(\Theta)$, de dérivée

$$h'_\eta(\mathbf{x}) = \psi'(\eta) g'_{\psi(\eta)}(\mathbf{x}).$$

De cette relation on déduit que, pour tout $\eta \in \varphi(\Theta)$,

$$J(\eta) = \int_E \frac{h'_\eta(\mathbf{x})^2}{h_\eta(\mathbf{x})}\, \mathbb{1}_{h_\eta(\mathbf{x})>0}\, \mu(d\mathbf{x}) = \psi'(\eta)^2 \int_E \frac{g'_{\psi(\eta)}(\mathbf{x})^2}{g_{\psi(\eta)}(\mathbf{x})}\, \mathbb{1}_{g_{\psi(\eta)}(\mathbf{x})>0}\, \mu(d\mathbf{x}) = \psi'(\eta)^2 I(\psi(\eta)),$$

qui correspond à une fonction continue sur $\varphi(\Theta)$ puisque ψ est C^1 et le modèle initial régulier.

■

Voyons ce que ça donne sur les deux exemples les plus classiques de changements de paramètres.

Exemples :

1. Translation : si on pose $\eta = \theta - \theta_0$ avec θ_0 fixé, alors
$$J(\eta) = I(\theta_0 + \eta).$$
2. Changement d'échelle : si on pose $\eta = \theta/\sigma$ avec σ fixé non nul, alors
$$J(\eta) = \sigma^2 I(\sigma\eta).$$

Lorsqu'on dispose d'un échantillon iid, l'information de Fisher croît linéairement avec la taille de l'échantillon. En d'autres termes, l'information apportée par n observations iid est n fois plus grande que l'information apportée par une seule.

Proposition 24 (Information de Fisher d'un échantillon)
Soit $\mathbf{X} = (X_1, \ldots, X_n)$ *un échantillon iid, où* X_i *a pour densité marginale* f_θ *par rapport à la mesure* μ. *Si le modèle* $(f_\theta)_{\theta\in\Theta}$ *est régulier d'information de Fisher* $I(\theta) = I_1(\theta)$, *alors le modèle produit, de densité*

$$g_\theta(\mathbf{x}) = g_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

par rapport à la mesure $\mu^{\otimes n}$, *est encore régulier et d'information de Fisher* $I_n(\theta) = nI_1(\theta)$.

Remarque. Ce résultat est une conséquence du suivant : si $(P_\theta)_{\theta\in\Theta} = (g_\theta \cdot \mu)_{\theta\in\Theta}$ et $(Q_\theta)_{\theta\in\Theta} = (h_\theta \cdot \nu)_{\theta\in\Theta}$ sont deux modèles réguliers d'informations respectives $I_1(\theta)$ et $I_2(\theta)$, alors le modèle produit, de densité

$$k_\theta(\mathbf{x}, \mathbf{y}) = g_\theta(\mathbf{x})h_\theta(\mathbf{y})$$

par rapport à la mesure $\mu\otimes\nu$ sur $E\times F$, est régulier et d'information de Fisher $I(\theta) = I_1(\theta)+I_2(\theta)$. Avec des mots : l'information d'un couple de variables indépendantes est la somme des deux informations.

Preuve. Nous allons démontrer le résultat de la remarque, celui de la proposition s'en déduisant par récurrence. Tout d'abord, on note que la régularité de la fonction

$$\theta \mapsto k_\theta(\mathbf{x}, \mathbf{y}) = g_\theta(\mathbf{x})h_\theta(\mathbf{y})$$

se déduit de celles de $\theta \mapsto g_\theta(\mathbf{x})$ et $\theta \mapsto h_\theta(\mathbf{y})$. Ainsi, pour $\mu \otimes \nu$ presque tout couple $(\mathbf{x}, \mathbf{y})$,

$$k'_\theta(\mathbf{x}, \mathbf{y}) = g'_\theta(\mathbf{x})h_\theta(\mathbf{y}) + g_\theta(\mathbf{x})h'_\theta(\mathbf{y}),$$

d'où

$$k'_\theta(\mathbf{x}, \mathbf{y})^2 = g'_\theta(\mathbf{x})^2 h_\theta(\mathbf{y})^2 + 2(g'_\theta(\mathbf{x})g_\theta(\mathbf{x}))(h'_\theta(\mathbf{y})h_\theta(\mathbf{y})) + g_\theta(\mathbf{x})^2 h'_\theta(\mathbf{y})^2,$$

et sur l'ensemble $S_\theta = \{(\mathbf{x}, \mathbf{y}),\ g_\theta(\mathbf{x})h_\theta(\mathbf{y}) > 0\}$ où l'on calculera l'intégrale d'intérêt, on a donc

$$\frac{k'_\theta(\mathbf{x}, \mathbf{y})^2}{k_\theta(\mathbf{x}, \mathbf{y})} = \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}h_\theta(\mathbf{y}) + 2g'_\theta(\mathbf{x})h'_\theta(\mathbf{y}) + g_\theta(\mathbf{x})\frac{h'_\theta(\mathbf{y})^2}{h_\theta(\mathbf{y})}.$$

De là il ressort que l'intégrale définissant l'information de Fisher

$$I(\theta) = \iint \frac{k'_\theta(\mathbf{x}, \mathbf{y})^2}{k_\theta(\mathbf{x}, \mathbf{y})}\mu(d\mathbf{x})\mu(d\mathbf{y})$$

est la somme de trois termes, le premier et le dernier étant comparables. Le premier s'écrit (l'intégration se faisant sur S_θ)

$$\iint \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}h_\theta(\mathbf{y})\mu(d\mathbf{x})\mu(d\mathbf{y}) = \left(\int \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\mu(d\mathbf{x})\right)\left(\int h_\theta(\mathbf{y})\mu(d\mathbf{y})\right) = I_1(\theta),$$

puisque pour tout θ, $y \mapsto h_\theta(\mathbf{y})$ est une densité, donc d'intégrale 1. De même, le troisième terme vaut $I_2(\theta)$. Reste à montrer que celui du milieu est nul, or

$$\iint g'_\theta(\mathbf{x})h'_\theta(\mathbf{y})\mu(d\mathbf{x})\mu(d\mathbf{y}) = \left(\int g'_\theta(\mathbf{x})\mu(d\mathbf{x})\right)\left(\int h'_\theta(\mathbf{y})\mu(d\mathbf{y})\right) = 0,$$

ces deux intégrales étant nulles via la Proposition 21 : les scores sont des variables centrées. Les fonctions I_1 et I_2 étant toutes deux continues, le résultat est établi.

■

Si l'on admet que le modèle produit est régulier, alors le résultat de la Proposition 24 découle tout simplement du fait que, dans le cas indépendant, la variance de la somme correspond à la somme des variances. Avec un abus de notations :

$$g_\theta(\mathbf{X}) = \prod_{i=1}^n f_\theta(X_i) \Longrightarrow \ell_\theta(\mathbf{X}) = \sum_{i=1}^n \ell_\theta(X_i) \Longrightarrow I_n(\theta) = \mathrm{Var}_\theta(\ell'_\theta(\mathbf{X})) = \sum_{i=1}^n \mathrm{Var}_\theta(\ell'_\theta(X_i)) = nI_1(\theta).$$

5.3.2 Exemples et contre-exemples

La Proposition 24 nous dit que l'information de Fisher d'un échantillon iid se déduit de celle d'une seule variable. C'est pourquoi, dans tout ce qui suit, nous ne noterons plus $\mathbf{x}$ et $\mathbf{X}$, mais x et X qui représentent donc des quantités réelles, discrètes ou continues, et $f_\theta(x)$ au lieu de $g_\theta(\mathbf{x})$ pour les densités.

Modèles exponentiels

Commençons par le modèle exponentiel, lequel comprend un grand nombre de lois classiques. En reprenant les notations de la Section 4.2.3, le modèle général est défini par une densité (par rapport à une mesure de référence μ) de la forme

$$f_\theta(x) = C(\theta)h(x)\exp(Q(\theta)T(x)),$$

la fonction Q étant continue et strictement croissante sur l'intervalle ouvert Θ. Dès que Q est C^1, tout se passe bien.

Proposition 25 (Régularité du modèle exponentiel)
Si, dans le modèle exponentiel général, la fonction Q est C^1 sur Θ, alors le modèle est régulier d'information de Fisher

$$I(\theta) = \mathrm{Var}_\theta(\ell'_\theta(X)) = Q'(\theta)^2\mathrm{Var}_\theta(T(X)).$$

Si la fonction Q est C^2 sur Θ, alors on a aussi

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

où $\ell_\theta(x) = \log f_\theta(x)$.

Preuve. Nous allons montrer que le modèle vérifie les conditions "classiques" du Lemme 1, ce qui assurera donc sa régularité au sens de la Définition 32. Commençons par le modèle naturel, défini par une densité de la forme

$$g_\eta(x) = \exp(\eta T(x) - A(\eta)) \qquad \text{avec} \qquad A(\eta) = \log\left(\int_{\mathbb{R}} \exp(\eta T(x))\ \nu(dx)\right),$$

par rapport à la mesure ν, avec η dans l'intervalle ouvert I. L'ensemble $S = \{x \in \mathbb{R},\ g_\eta(x) > 0\}$ est $\mathbb{R}$ tout entier, donc indépendant de η. Rappelons aussi que, par la Proposition 18 du Chapitre 4, la fonction A est C^∞ sur I. De ce fait, pour tout x, la fonction $\eta \mapsto g_\eta(x)$ est elle-même C^∞ sur I, avec en particulier

$$g'_\eta(x) = (T(x) - A'(\eta))g_\eta(x) \implies \frac{g'_\eta(x)^2}{g_\eta(x)} = (T(x) - A'(\eta))^2 g_\eta(x).$$

Pour $\eta_0 \in I$ fixé et $\varepsilon > 0$ tel que $[\eta_0 - \varepsilon, \eta_0 + \varepsilon] \subset I$, l'inégalité $(a-b)^2 \leq 2(a^2+b^2)$ donne pour tout x la majoration

$$\sup_{\eta_0-\varepsilon<\eta<\eta_0+\varepsilon} \frac{g'_\eta(x)^2}{g_\eta(x)} \leq 2(T^2(x) + C_1^2) \sup_{\eta_0-\varepsilon<\eta<\eta_0+\varepsilon} g_\eta(x),$$

où $C_1 = \sup_{\eta_0-\varepsilon<\eta<\eta_0+\varepsilon} |A'(\eta)|$. Notant encore $C_0 = \sup_{\eta_0-\varepsilon<\eta<\eta_0+\varepsilon} \exp(-A(\eta))$ et revenant à la définition de $g_\eta(x)$, ceci permet d'écrire

$$\sup_{\eta_0-\varepsilon<\eta<\eta_0+\varepsilon} \frac{g'_\eta(x)^2}{g_\eta(x)} \leq 2(T^2(x) + C_1^2)C_0(\exp((\eta_0-\varepsilon)T(x)) + \exp((\eta_0+\varepsilon)T(x))) =: \varphi(x).$$

Puisque $[\eta_0-\varepsilon,\eta_0+\varepsilon]\subset I$, la fonction φ appartient à $L_1(\nu)$ et l'hypothèse de domination locale est satisfaite. Il reste à calculer l'information de Fisher, qui est très simple puisque $A'(\eta)=\mathbb{E}_\eta[T(X)]$:

$$J(\eta)=\int_{\mathbb{R}}\frac{g'_\eta(x)^2}{g_\eta(x)}\nu(dx)=\int_{\mathbb{R}}(T(x)-A'(\eta))^2 g_\eta(x)\nu(dx)=\mathbb{E}_\eta\left[(T(X)-A'(\eta))^2\right]=\mathrm{Var}_\eta(T(X)).$$

Pour le modèle général, on applique le Lemme 2 du Chapitre 4 et la Proposition 23 ci-dessus au changement de paramètre $\theta=Q^{-1}(\eta)$. Puisque $\psi=Q$ est C^1, le modèle général est lui aussi régulier, d'information de Fisher

$$I(\theta)=Q'(\theta)^2 J(Q(\theta))=Q'(\theta)^2\mathrm{Var}_\theta(T(X)).$$

■

Selon la situation, on pourra préférer l'une ou l'autre des formules pour calculer l'information de Fisher du modèle. On peut l'illustrer sur quelques exemples.

Exemples :

1. Loi exponentielle : reprenons l'exemple de la loi exponentielle du début de section. Dans ce cas, on a tout simplement $Q(\theta)=\theta$ et $T(x)=-x$. Puisque Q est C^∞, le modèle exponentiel associé est régulier, d'information de Fisher :
$$I(\theta)=(Q'(\theta))^2\mathrm{Var}_\theta(T(X))=\mathrm{Var}_\theta(-X)=\mathrm{Var}_\theta(X)=\frac{1}{\theta^2},$$
et on retrouve sans souffrir le résultat un peu laborieusement obtenu à l'époque.
2. Loi binomiale : si $X\sim\mathcal{B}(n,\theta)$ avec $0<\theta<1$, alors on a bien un modèle exponentiel avec $T(x)=x$ et $Q(\theta)=\log(\theta/(1-\theta))$ qui est C^2 sur $]0,1[$. On en déduit que ce modèle est régulier, d'information de Fisher
$$I(\theta)=(Q'(\theta))^2\mathrm{Var}_\theta(X)=\frac{n}{\theta(1-\theta)}.$$
3. Loi de Poisson : si $X\sim\mathcal{P}(\lambda)$, alors
$$\ell_\lambda(x)=-\lambda+x\log\lambda-\log(x!)\implies\ell''_\lambda(x)=-\frac{x}{\lambda^2}\implies I(\lambda)=\frac{1}{\lambda^2}\mathbb{E}[X]=\frac{1}{\lambda}.$$
4. Loi gaussienne : si $X\sim\mathcal{N}(\mu,\sigma^2)$, le logarithme de la densité s'écrit
$$\log f(x)=-\frac{1}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}(x-\mu)^2.$$
Si le paramètre est μ (i.e. σ connu), alors la dérivée seconde par rapport à μ est constante
$$\ell''_\mu(x)=-\frac{1}{\sigma^2}\implies I(\mu)=-\mathbb{E}[\ell''_\mu(X)]=\frac{1}{\sigma^2}.$$
Si le paramètre est σ^2 (i.e. μ connu), alors la dérivée seconde par rapport à σ^2 (et non par rapport à σ !) vaut
$$\ell''_{\sigma^2}(x)=\frac{1}{2\sigma^4}-\frac{1}{\sigma^6}(x-\mu)^2\implies I(\sigma^2)=-\mathbb{E}[\ell''_{\sigma^2}(X)]=\frac{1}{2\sigma^4},$$
compte tenu du fait que $\mathbb{E}[(X-\mu)]=\sigma^2$. Noter que si on considère σ comme paramètre, alors la Proposition 23 donne pour information de Fisher $J(\sigma)=2/\sigma^2$.

Interprétation. Revenons sur le modèle gaussien de moyenne μ inconnue. Intuitivement, l'information de Fisher peut s'interpréter comme la quantité d'information apportée par une observation pour estimer le paramètre inconnu. En ce sens, plus l'écart-type σ est petit, plus la variable $X\sim\mathcal{N}(\mu,\sigma^2)$ a des chances de tomber près de la moyenne μ que l'on cherche, donc plus on aura "d'information" sur celle-ci grâce à celle-là : ceci est cohérent avec le fait que $I(\mu)=1/\sigma^2$. Avec cette interprétation, il est tout aussi logique que $I(\mu)$ ne dépende pas de μ : que la moyenne vaille 0 ou 50, l'information sur cette moyenne apportée par une observation est clairement la même.

Modèles de translation

Nous considérons ici une densité $f(x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, indépendante de θ, et le modèle de translation associé

$$(f_\theta(x))_{\theta\in\mathbb{R}} = (f(x-\theta))_{\theta\in\mathbb{R}}.$$

Comme on peut s'y attendre, la régularité de ce modèle ne dépend que de f.

Proposition 26 (Régularité d'un modèle de translation)
Si la densité f est continue sur $\mathbb{R}$, de classe C^1 sauf éventuellement en un nombre fini de points, avec

$$I := \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)} \mathbb{1}_{f(x)>0}\, dx < +\infty,$$

alors le modèle de translation $(f_\theta(x))_{\theta\in\mathbb{R}}$ est régulier, d'information de Fisher constante égale à $I(\theta) = I$ pour tout θ.

Preuve. Pour tout x, la fonction $\theta \mapsto f_\theta(x) = f(x-\theta)$ hérite des propriétés de régularité de f. Puisque f est supposée continue sur $\mathbb{R}$ et de classe C^1 sauf éventuellement en un nombre fini de points, il en va de même pour $f(x-\theta)$ vue comme une fonction de θ. En tout point où elle existe, sa dérivée vaut donc $f'_\theta(x) = -f'(x-\theta)$. L'information de Fisher est alors triviale via le changement de variable $y = x-\theta$:

$$I(\theta) = \int_{\mathbb{R}} \frac{f'_\theta(x)^2}{f_\theta(x)} \mathbb{1}_{f_\theta(x)>0}\, dx = \int_{\mathbb{R}} \frac{f'(x-\theta)^2}{f(x-\theta)} \mathbb{1}_{f(x-\theta)>0}\, dx = \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)} \mathbb{1}_{f(y)>0}\, dy = I.$$

Une application constante étant continue, le modèle est régulier.

■

Remarque. Le modèle gaussien de moyenne inconnue est clairement un cas particulier de ce résultat en prenant pour f la densité d'une gaussienne centrée de variance σ^2.

Donnons quelques exemples pour fixer les idées et voir la différence entre la Définition 32 que nous avons adoptée et celle, plus classique mais plus restrictive, du Lemme 1.

Exemples :

1. Loi de Laplace : partons de la densité $f(x) = \frac{1}{2}e^{-|x|}$. Celle-ci est continue sur $\mathbb{R}$ et, hormis en l'origine, dérivable de dérivée continue (cf. Figure 5.1) :

$$\forall x \neq 0 \qquad f'(x)^2 = \frac{1}{4}e^{-2|x|} \implies \frac{f'(x)^2}{f(x)} = \frac{1}{2}e^{-|x|} \implies \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)} dx = 1.$$

   Le modèle de translation $(f_\theta(x))_{\theta\in\mathbb{R}}$ est donc régulier, d'information de Fisher égale à 1. On remarque au passage que ce modèle ne satisfait pas la condition de régularité requise par le Lemme 1 puisque, quel que soit x, la fonction $\theta \mapsto f_\theta(x) = f(x-\theta)$ n'est pas C^1 sur $\Theta = \mathbb{R}$ (problème en $\theta = x$).

2. Loi exotique : on considère cette fois la densité de classe C^1 (cf. Figure 5.1)

$$f(x) = \frac{1+\cos x}{2\pi}\mathbb{1}_{[-\pi,\pi]}(x) \implies f'(x) = \frac{-\sin x}{2\pi}\mathbb{1}_{[-\pi,\pi]}(x) \implies \frac{f'(x)^2}{f(x)} = \frac{1-\cos x}{2\pi}\mathbb{1}_{[-\pi,\pi]}(x)$$

   donc le modèle de translation associé est régulier et a pour information de Fisher

$$I = \frac{1}{2\pi}\int_{-\pi}^{\pi} (1-\cos x) dx = 1.$$

   Ici, le modèle ne satisfait pas la condition de support du Lemme 1, puisque le support de $f_\theta(x)$ est égal à $[\theta-\pi, \theta+\pi]$, donc dépendant de θ.

3. Contre-exemple de la loi uniforme : si $f(x) = \mathbb{1}_{[0,1]}(x)$, rien de ce qui précède ne s'applique puisque f présente deux discontinuités. Le modèle de translation associé n'est donc pas régulier. On retrouve ici le même problème que pour le modèle $(\mathcal{U}_{[0,\theta]})_{\theta\in\mathbb{R}}$ mentionné en début de section.

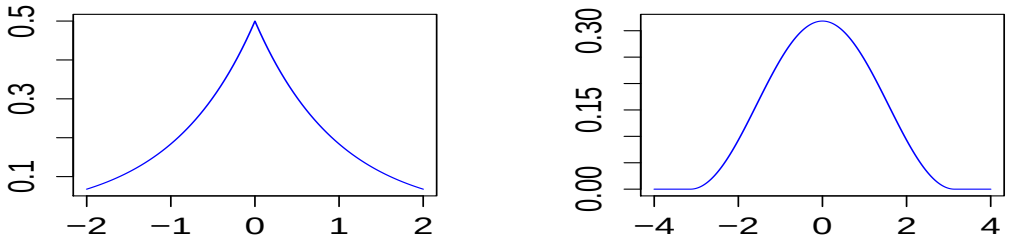

FIGURE 5.1 – Loi de Laplace (à gauche) et loi "exotique" (à droite).

5.4 Inégalités, bornes, efficacités

5.4.1 Inégalité de l'Information et borne de Cramér-Rao

Supposons qu'on veuille estimer θ à partir de l'observation $\mathbf{X}$ dans un modèle régulier. Peut-on avoir une idée du risque quadratique ? Le résultat suivant permet de le minorer. Rappelons que le fait de supposer une fonction localement bornée n'est pas très restrictif.

Proposition 27 (Inégalité de l'Information)
Soit $(f_\theta)_{\theta\in\Theta}$ un modèle régulier, $\hat{\theta}(\mathbf{X})$ un estimateur de θ dont le risque quadratique est localement borné, de biais noté $b(\theta) = \mathbb{E}_\theta[\hat{\theta}(\mathbf{X})] - \theta$. Alors on a la minoration suivante du risque quadratique : si $I(\theta) > 0$,

$$R(\hat{\theta}(\mathbf{X}), \theta) = \mathbb{E}_\theta\left[\left(\hat{\theta}(\mathbf{X}) - \theta\right)^2\right] \geq b(\theta)^2 + \frac{(1 + b'(\theta))^2}{I(\theta)}.$$

Remarque. De façon plus générale, si $\hat{\varphi}(\mathbf{X})$ est un estimateur de $\varphi(\theta)$ de risque quadratique localement borné, avec φ dérivable, de biais $b(\theta) = \mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})] - \varphi(\theta)$, alors si $I(\theta) > 0$, on a

$$\mathbb{E}_\theta\left[(\hat{\varphi}(\mathbf{X}) - \varphi(\theta))^2\right] \geq b(\theta)^2 + \frac{(\varphi'(\theta) + b'(\theta))^2}{I(\theta)}.$$

Preuve. Nous allons démontrer l'inégalité de la remarque, celle de la Proposition 27 s'en déduisant en prenant $\varphi(\theta) = \theta$. Puisque $(a+b)^2 \leq 2(a^2+b^2)$, il vient

$$\hat{\varphi}(\mathbf{X})^2 \leq 2\left\{(\hat{\varphi}(\mathbf{X}) - \varphi(\theta))^2 + \varphi(\theta)^2\right\} \implies \mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})^2] \leq 2\left\{\mathbb{E}_\theta\left[(\hat{\varphi}(\mathbf{X}) - \varphi(\theta))^2\right] + \varphi(\theta)^2\right\}.$$

Les deux membres de droite étant localement bornés, il en va de même pour celui de gauche. On peut donc appliquer la Proposition 21 à la statistique $\hat{\varphi}(\mathbf{X})$, ce qui donne

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})] = \mathbb{E}_\theta\left[\hat{\varphi}(\mathbf{X})\frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right].$$

Or on sait que

$$\mathbb{E}_\theta\left[\frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right] = \mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})\right] = 0,$$

donc l'équation précédente s'écrit encore

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})] = \mathbb{E}_\theta\left[(\hat{\varphi}(\mathbf{X}) - \mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})])\frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right].$$

L'inégalité de Cauchy-Schwarz donne alors

$$\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})]\right)^2 \leq \mathrm{Var}_\theta(\hat{\varphi}(\mathbf{X})) \times I(\theta). \tag{5.2}$$

Il reste à voir que, pour le membre de gauche, $\mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})] = b(\theta)+\varphi(\theta)$. La fonction φ étant supposée dérivable, le biais l'est aussi et

$$\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\varphi}(\mathbf{X})]\right)^2 = (\varphi'(\theta) + b'(\theta))^2.$$

On peut de plus appliquer au membre de droite la décomposition classique du risque quadratique :

$$\mathrm{Var}_\theta(\hat{\varphi}(\mathbf{X})) = \mathbb{E}_\theta\left[(\hat{\varphi}(\mathbf{X}) - \varphi(\theta))^2\right] - b(\theta)^2.$$

On arrive ainsi au résultat souhaité, si tant est que $I(\theta)$ soit strictement positif.

■

Remarque. Dans la preuve précédente, la variance apparaît dans l'inégalité (5.2). On voit que si $I(\theta_0) = 0$, tout s'écroule et on perd toute information sur la variance de $\hat{\varphi}(\mathbf{X})$ en θ_0.

Donnons maintenant la version la plus connue de l'inégalité précédente : elle due à Fréchet, Darmois, Cramér et Rao, mais l'usage n'a conservé que les deux derniers auteurs.

Corollaire 3 (Borne de Cramér-Rao)
Si $\hat{\theta}(\mathbf{X})$ un estimateur sans biais de θ dont le risque quadratique est localement borné et si $I(\theta) > 0$, alors

$$\mathrm{Var}_\theta(\hat{\theta}(\mathbf{X})) \geq \frac{1}{I(\theta)}.$$

Pour un modèle d'échantillonnage régulier où $\mathbf{X} = (X_1, \ldots, X_n)$ et pour un estimateur sans biais $\hat{\theta}_n(\mathbf{X})$, cette borne devient

$$\mathrm{Var}_\theta(\hat{\theta}_n(\mathbf{X})) \geq \frac{1}{nI_1(\theta)}.$$

Un estimateur atteignant cette borne est dit efficace.

Remarque. Pour un estimateur non biaisé $\hat{\varphi}_n(\mathbf{X})$ de $\varphi(\theta)$, la borne de Cramér-Rao s'écrit donc

$$\mathbb{E}_\theta\left[(\hat{\varphi}_n(\mathbf{X}) - \varphi(\theta))^2\right] = \mathrm{Var}_\theta(\hat{\varphi}_n(\mathbf{X})) \geq \frac{\varphi'(\theta)^2}{nI_1(\theta)}.$$

Exemple. Reprenons l'exemple du cas gaussien où la variance $\sigma^2 > 0$ est inconnue, en supposant pour simplifier que la moyenne est nulle (ça ne change rien), c'est-à-dire

$$(X_1, \ldots, X_n) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Ce modèle est régulier, d'information de Fisher $I_1(\sigma^2) = 1/(2\sigma^4)$, d'où $I_n(\sigma^2) = n/(2\sigma^4)$. Considérons l'estimateur du maximum de vraisemblance (cf. Section 4.2.2)

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

Il est clairement non biaisé et de variance [8]

$$\mathrm{Var}(\hat{\sigma}^2) = \frac{\mathrm{Var}(X_1^2)}{n} = \frac{\mathbb{E}[X_1^4] - (\mathbb{E}[X_1^2])^2}{n} = \frac{2\sigma^4}{n},$$

qui est précisément la borne de Cramér-Rao : c'est donc un estimateur efficace.

A ce stade, on serait tenté de dire que la notion d'efficacité est pertinente pour caractériser l'optimalité d'un estimateur. Il se trouve que non. En forçant le trait, on pourrait même dire qu'elle est à peu près sans intérêt et il y a au moins deux raisons à cela. La première est que, comme on l'a vu en Section 5.1.2, les estimateurs sans biais, lorsqu'ils existent, ne sont pas nécessairement les plus intéressants en terme d'erreur quadratique. La seconde vient du résultat suivant (admis).

Proposition 28 (Efficacité et modèles exponentiels)
L'estimateur sans biais et de risque quadratique localement borné $\hat{\theta}(X)$ est un estimateur efficace de θ si et seulement si le modèle est exponentiel et l'estimateur (pour une seule donnée) de la forme $\hat{\theta}(X) = a + bT(X)$.

Ceci signifie qu'un estimateur efficace ne peut exister que dans des conditions très particulières et clairement identifiées. L'exemple gaussien ci-dessus rentre bien dans ce cadre puisque

$$f_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}} \implies T(x) = x^2$$

qui correspond bien à $\hat{\sigma}^2(x)$ lorsqu'on considère un échantillon de taille 1. En fait, la plupart des problèmes d'estimation n'admettent pas d'estimateur efficace, même dans le contexte des modèles exponentiels.

Exemple. Reprenons le cas des lois exponentielles $(\mathcal{E}(\lambda))_{\lambda>0}$. Comme nous l'avons vu, elles rentrent bien dans le cadre du modèle exponentiel avec $Q(\lambda) = \lambda$ et $T(x) = -x$. Pourtant, tout estimateur de la forme $a + bT(X) = a - bX$ est voué à l'échec, ne serait-ce que pour des raisons de biais :

$$\mathbb{E}_\lambda[a + bT(X)] = a - b\mathbb{E}_\lambda[X] = a - \frac{b}{\lambda} \neq \lambda.$$

Le calcul de l'information de Fisher a déjà été fait : $I_1(\lambda) = 1/\lambda^2$. Lorsque

$$\mathbf{X} = (X_1, \ldots, X_n) \overset{iid}{\sim} \mathcal{E}(\lambda),$$

l'estimateur au maximum de vraisemblance (ou de la méthode des moments) est $1/\bar{X}_n$. Il est biaisé : en effet, $n\bar{X}_n \sim \Gamma(n, \lambda)$, or un calcul facile montre que

$$Z \sim \Gamma(n, \lambda) \implies \mathbb{E}[1/Z] = \frac{\lambda}{n-1},$$

d'où l'on déduit que $\mathbb{E}_\lambda[1/\bar{X}_n] = n\lambda/(n-1)$. Considérons alors l'estimateur sans biais

$$\hat{\lambda}_n = \hat{\lambda}_n(\mathbf{X}) = \frac{n-1}{n\bar{X}_n}.$$

8. Rappelons que si $X \sim \mathcal{N}(0,1)$, alors $\mathbb{E}[X^4] = 3$, cas particulier de la formule générale : $\mathbb{E}[X^{2n}] = (2n)!/(2^n n!)$.

Puisqu'un calcul du même type que celui mentionné plus haut assure que

$$Z \sim \Gamma(n,\lambda) \implies \mathbb{E}[1/Z^2] = \frac{\lambda^2}{(n-1)(n-2)},$$

on en déduit que

$$\mathrm{Var}_\lambda(\hat{\lambda}_n) = \frac{\lambda^2}{n-2} > \frac{1}{nI_1(\lambda)} = \frac{\lambda^2}{n}.$$

La borne de Cramér-Rao n'est pas atteinte et cet estimateur n'est pas efficace. Néanmoins, on voit qu'asymptotiquement

$$n\mathrm{Var}_\lambda(\hat{\lambda}_n) \xrightarrow[n\to\infty]{} \frac{1}{\lambda^2} = \frac{1}{I_1(\lambda)}.$$

Ce genre de phénomène, tout à fait typique, incite naturellement à introduire le concept d'efficacité asymptotique.

Remarque. Avant de passer à l'efficacité asymptotique, revenons aux lois exponentielles, que nous définissons cette fois pour tout $\theta > 0$ par [9]

$$f_\theta(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \mathbb{1}_{x\geq 0}.$$

A partir d'un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ iid suivant cette loi, l'estimateur naturel (maximum de vraisemblance ou méthode des moments) est donc maintenant $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}) = \bar{X}_n$. Il est non biaisé et de variance

$$\mathrm{Var}_\theta(\hat{\theta}_n) = \frac{\mathrm{Var}_\theta(X_1)}{n} = \frac{\theta^2}{n}.$$

Or l'information de Fisher vaut, via le changement de paramètre $\lambda = \psi(\theta) = 1/\theta$:

$$J_1(\theta) = \psi'(\theta)^2 I_1(1/\theta) = \frac{1}{\theta^2} \implies J_n(\theta) = \frac{n}{\theta^2} = \frac{1}{\mathrm{Var}_\theta(\hat{\theta}_n)},$$

et on a cette fois un estimateur efficace ! Ceci montre qu'un simple changement de paramètre, aussi régulier soit-il, modifie la propriété d'efficacité. Notons par ailleurs que ceci est cohérent avec la Proposition 28 puisque, dans ce modèle exponentiel, nous avons $T(x) = -x$ et, pour un échantillon de taille 1, l'estimateur $\hat{\theta}(x) = x = -T(x)$.

5.4.2 Efficacité asymptotique

La borne inférieure donnée par l'Inégalité de l'Information vue en Proposition 27 n'est pas satisfaisante en ce sens qu'elle minore le risque quadratique en un seul point. Or on peut trouver un estimateur trivial qui est imbattable à ce jeu-là !

Exemple. En effet, considérons pour simplifier $\Theta = \mathbb{R}$ et l'estimateur constant $\tilde{\theta}(\mathbf{X}) = 0$ pour toute observation $\mathbf{X}$. Le biais et sa dérivée sont élémentaires :

$$b(\theta) = \mathbb{E}_\theta[\tilde{\theta}(\mathbf{X})] - \theta = -\theta \implies b'(\theta) = -1.$$

Par ailleurs, sa variance est nulle, d'où le risque

$$\mathbb{E}_\theta\left[\left(\tilde{\theta}(\mathbf{X}) - \theta\right)^2\right] = \theta^2 = b(\theta)^2 = b(\theta)^2 + \frac{(1+b'(\theta))^2}{I(\theta)},$$

et on a égalité dans l'Inégalité de l'Information. Dirait-on pour autant que cet estimateur est optimal ? Clairement non, il est même désastreux dès que le vrai paramètre θ est loin de l'origine.

9. C'est d'ailleurs la définition donnée dans beaucoup d'ouvrages et considérée par certains logiciels.

Le problème de l'exemple précédent vient de ce qu'on a minimisé le terme de variance (en l'annulant) sans contrôler le terme de biais. Or on sait qu'un bon estimateur doit avoir un biais et une variance qui sont tous deux petits. Pour évacuer ce genre d'estimateur sans intérêt et arriver à nos fins, une idée est de contrôler uniformément le risque quadratique. Le résultat suivant va dans ce sens.

Théorème 18 (Inégalité de l'Information uniforme)
Soit un modèle régulier $(f_\theta)_{\theta\in\Theta}$ d'information de Fisher $I(\theta) = I_1(\theta)$ et J un segment de Θ de longueur $2r$ sur lequel I est majorée par $\bar{I} = \sup_{\theta\in J} I(\theta)$ et ne s'annule pas. Si l'on dispose d'un échantillon $\mathbf{X} = (X_1, \dots, X_n)$ iid selon f_θ, alors pour tout estimateur $\hat{\theta}_n(\mathbf{X})$, on a

$$\sup_{\theta\in J} \mathbb{E}_\theta\left[\left(\hat{\theta}_n(\mathbf{X}) - \theta\right)^2\right] \geq \frac{1}{n\bar{I}} \times \left(\frac{1}{1 + \frac{1}{r\sqrt{n\bar{I}}}}\right)^2.$$

Exemple. Pour l'estimation de la moyenne dans le modèle $(\mathcal{N}(\mu, 1))_{\mu\in\mathbb{R}}$, nous avons vu que l'information est constante égale à $I(\mu) = 1$, donc elle ne s'annule sur aucun intervalle $J = [-r, r]$ et est majorée par 1. L'inégalité précédente nous apprend que, pour tout estimateur $\hat{\mu}_n(\mathbf{X})$,

$$\sup_{-r\leq\mu\leq r} \mathbb{E}_\mu\left[(\hat{\mu}_n(\mathbf{X}) - \mu)^2\right] \geq \frac{1}{n} \times \left(\frac{1}{1 + \frac{1}{r\sqrt{n}}}\right)^2.$$

En particulier, on voit que l'estimateur trivial $\tilde{\mu}(\mathbf{X}) = \tilde{\mu}_n(\mathbf{X}) = 0$ proposé ci-dessus n'est plus du tout optimal puisque

$$\sup_{-r\leq\mu\leq r} \mathbb{E}_\mu\left[(\tilde{\mu}_n(\mathbf{X}) - \mu)^2\right] = r^2,$$

tandis que la borne inférieure tend vers 0 à vitesse $1/n$. Tout ça est rassurant.

Preuve. Afin d'alléger les notations, convenons de noter le risque quadratique

$$R(\theta) = \mathbb{E}_\theta\left[\left(\hat{\theta}_n(\mathbf{X}) - \theta\right)^2\right].$$

Nous cherchons donc à minorer le supremum sur J de $R(\theta)$. S'il n'est pas borné, l'inégalité est évidente. S'il est borné sur un intervalle ouvert contenant J, il est localement borné sur J et on peut appliquer l'Inégalité de l'Information en tout point θ de J, à savoir

$$R(\theta) \geq b(\theta)^2 + \frac{(1 + b'(\theta))^2}{nI(\theta)}.$$

Introduisons un coefficient de réglage $c \in]0, 1[$. Deux cas de figure sont alors envisageables :

— ou bien il existe $\theta_0 \in J$ tel que $|b'(\theta_0)| \leq c$, alors en ce point l'Inégalité de l'Information nous dit que

$$R(\theta_0) \geq b(\theta_0)^2 + \frac{(1 + b'(\theta_0))^2}{nI(\theta_0)} \geq \frac{(1 + b'(\theta_0))^2}{nI(\theta_0)} \geq \frac{(1 - c)^2}{nI(\theta_0)} \geq \frac{(1 - c)^2}{n\bar{I}},$$

et a fortiori

$$\sup_{\theta\in J} R(\theta) \geq R(\theta_0) \geq \frac{(1 - c)^2}{n\bar{I}}.$$

— ou bien $|b'(\theta)| > c$ pour tout $\theta \in J$. Puisqu'elle est continue, b' a donc un signe constant sur J et la variation de b sur J est minorée par $2cr$:

$$\sup_{\theta\in J} b(\theta) - \inf_{\theta\in J} b(\theta) \geq 2cr \implies \sup_{\theta\in J} |b(\theta)| \geq cr$$

et, toujours par l'Inégalité de l'Information,

$$\sup_{\theta\in J} R(\theta) \geq \sup_{\theta\in J} b(\theta)^2 \geq (cr)^2.$$

Quoi qu'il en soit, on a établi que

$$\forall c \in]0,1[\qquad \sup_{\theta\in J} R(\theta) \geq \min\left((cr)^2, \frac{(1-c)^2}{n\bar{I}}\right),$$

d'où, en équilibrant les deux termes,

$$c = \frac{1}{1+r\sqrt{n\bar{I}}} \implies \sup_{\theta\in J} R(\theta) \geq \frac{1}{n\bar{I}} \times \left(\frac{1}{1+\frac{1}{r\sqrt{n\bar{I}}}}\right)^2,$$

ce qui est le résultat voulu. Il reste à voir que si R est borné sur $J = [m-r, m+r]$ mais non localement borné sur un intervalle ouvert contenant J, il suffit d'appliquer ce raisonnement aux intervalles de la forme $[m-r+\varepsilon, m+r-\varepsilon]$ puis de faire tendre ε vers 0. Le résultat passe à la limite grâce à la continuité de I sur Θ, donc sur J.

■

Remarques :

1. L'astuce consistant à choisir c de façon à égaliser les deux termes est un grand classique en statistique : elle revient simplement à équilibrer le biais (au carré) et la variance. En statistique non paramétrique, on la retrouve par exemple pour le choix de la fenêtre dans les estimateurs à noyaux ou le nombre de voisins dans la méthode des plus proches voisins.
2. On peut généraliser l'inégalité de la Proposition 18 à un estimateur $\hat{\varphi}_n(\mathbf{X})$ de $\varphi(\theta)$ tel que φ soit C^1 de dérivée ne s'annulant pas sur Θ. En notant

$$\bar{I}_\varphi = \sup_{\theta\in J} \frac{I(\theta)}{\varphi'(\theta)^2} \qquad \text{et} \qquad \Delta(\varphi) = \sup_{\theta\in J} \varphi(\theta) - \inf_{\theta\in J} \varphi(\theta)$$

on peut en effet montrer que, sous les mêmes hypothèses,

$$\sup_{\theta\in J} \mathbb{E}_\theta\left[(\hat{\varphi}_n(\mathbf{X}) - \varphi(\theta))^2\right] \geq \frac{1}{n\bar{I}_\varphi} \times \left(\frac{1}{1+\frac{2}{\Delta(\varphi)\sqrt{n\bar{I}_\varphi}}}\right)^2.$$

Exemple. Revenons au résultat du Théorème 18. Pour un modèle de translation régulier, on a vu que l'information de Fisher est constante égale à I, donc $\bar{I} = I$. En faisant tendre r vers l'infini, on en déduit que

$$\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta\left[\left(\hat{\theta}_n(\mathbf{X}) - \theta\right)^2\right] \geq \frac{1}{n\bar{I}}.$$

Ce minorant n'est rien d'autre que la borne de Cramér-Rao, mais le point remarquable est qu'elle est valable pour tous les estimateurs de θ, pas uniquement pour les estimateurs sans biais !

De façon générale, on voit que si $r\sqrt{n\bar{I}}$ est grand alors

$$\frac{1}{n\bar{I}} \times \left(\frac{1}{1+\frac{1}{r\sqrt{n\bar{I}}}}\right)^2 \approx \frac{1}{n\bar{I}},$$

et l'inégalité établie est une généralisation de la borne de Cramér-Rao, où l'on n'est plus restreint aux estimateurs sans biais.

Considérons maintenant $J = [\theta_0 - \varepsilon, \theta_0 + \varepsilon]$ avec ε "petit", alors la continuité de la fonction I implique que

$$\bar{I} = \sup_{\theta \in J} I(\theta) = \sup_{\theta_0 - \varepsilon \leq \theta \leq \theta_0 + \varepsilon} I(\theta) \xrightarrow[\varepsilon \to 0]{} I(\theta_0),$$

et en faisant tendre n vers l'infini, la borne inférieure que l'on obtient est en

$$\frac{1}{n\bar{I}} \times \left(\frac{1}{1 + \frac{1}{r\sqrt{n\bar{I}}}} \right)^2 \underset{n \to \infty}{\sim} \frac{1}{n\bar{I}} \xrightarrow[\varepsilon \to 0]{} \frac{1}{nI(\theta_0)}.$$

En généralisant comme toujours via une fonction φ, la borne serait en $\varphi'(\theta_0)^2/(nI(\theta_0))$. Ceci laisse à penser que pour un estimateur $\hat{\varphi}_n(\mathbf{X})$ asymptotiquement normal, c'est-à-dire tel que

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

la plus petite valeur possible pour $\sigma^2(\theta)$ serait $\varphi'(\theta)^2/I(\theta)$. La définition de l'efficacité asymptotique part de ce constat.

Définition 34 (Efficacité asymptotique)
Soit un modèle régulier $(f_\theta)_{\theta \in \Theta}$ d'information de Fisher $I(\theta) = I_1(\theta)$, $\hat{\varphi}_n(\mathbf{X})$ un estimateur de $\varphi(\theta)$ où φ est C^1 de dérivée ne s'annulant pas sur Θ. Cet estimateur est dit asymptotiquement efficace si, lorsque $\mathbf{X} = (X_1, \ldots, X_n)$ est un échantillon iid selon f_θ, on a

$$\sqrt{n}\left(\hat{\varphi}_n(\mathbf{X}) - \varphi(\theta)\right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta)) \qquad \textit{avec} \qquad \sigma^2(\theta) \leq \frac{\varphi'(\theta)^2}{I(\theta)}$$

pour tout θ tel que $I(\theta) > 0$.

En ce sens, sous les hypothèses adéquates, l'information de Fisher permet bien de préciser ce que l'on peut attendre de mieux d'un estimateur. C'est ce que voulait dire, en tout début de Section 5.3, la phrase : "Un critère d'optimalité est spécifié par l'information de Fisher". Avant de donner des exemples d'estimateurs asymptotiquement efficaces, quelques remarques s'imposent.

Remarques :

1. Prenons $\varphi(\theta) = \theta$, qui est bien C^1 de dérivée $\varphi'(\theta) = \theta$ ne s'annulant pas sur Θ. Sous les mêmes hypothèses, un estimateur $\hat{\theta}_n(\mathbf{X})$ de θ est dit **asymptotiquement efficace** si on a

   $$\sqrt{n}\left(\hat{\theta}_n(\mathbf{X}) - \theta\right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta)) \qquad \text{avec} \qquad \sigma^2(\theta) \leq \frac{1}{I(\theta)}$$

   pour tout θ tel que $I(\theta) > 0$.
2. **Conséquence** : si $\hat{\theta}_n(\mathbf{X})$ est un estimateur asymptotiquement efficace de θ et si φ vérifie les hypothèses ci-dessus, alors la Delta méthode assure que $\hat{\varphi}_n(\mathbf{X})$ est un estimateur asymptotiquement efficace de $\varphi(\theta)$.
3. On ne peut avoir inégalité stricte dans la définition ci-dessus que sur un ensemble Θ_0 de mesure de Lebesgue nulle. En général, un estimateur asymptotiquement efficace est donc un estimateur asymptotiquement normal de variance asymptotique $\varphi'(\theta_0)^2/I(\theta_0)$.

Exemples :

1. Revenons au cas des lois exponentielles $(\mathcal{E}(\lambda))_{\lambda>0}$, modèle régulier d'information de Fisher $I_1(\lambda) = 1/\lambda^2$ strictement positive pour tout $\lambda > 0$. Nous avons vu que l'estimateur naturel $\tilde{\lambda}_n(\mathbf{X}) = 1/\bar{X}_n$ n'est pas efficace : d'une part il est biaisé, d'autre part même si on le débiaise on n'atteint pas la borne de Cramér-Rao. Néanmoins, quel que soit $\lambda > 0$, si $\mathbf{X} = (X_1, \dots, X_n)$ est un échantillon iid selon f_λ, le Théorème Central Limite nous dit que

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/\lambda^2).$$

La Delta méthode donne alors

$$\sqrt{n}\left(\tilde{\lambda}_n(\mathbf{X}) - \lambda\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \lambda^2) \qquad \text{avec} \qquad \lambda^2 = \frac{1}{I(\lambda)},$$

donc cet estimateur est asymptotiquement efficace.

2. Le modèle $(\mathcal{N}(\mu, 1))_{\mu\in\mathbb{R}}$ est lui aussi régulier, d'information de Fisher constante $I(\mu) = 1$. Si $\mathbf{X} = (X_1, \dots, X_n)$ est un échantillon iid selon f_μ, alors en vertu du Théorème Central Limite l'estimateur $\hat{\mu}_n(\mathbf{X}) = \bar{X}_n$ vérifie

$$\sqrt{n}\,(\hat{\mu}_n(\mathbf{X}) - \mu) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1) \qquad \text{avec} \qquad 1 = \frac{1}{I(\mu)},$$

donc c'est un estimateur asymptotiquement efficace. Etant donné que $\mathbb{E}_\mu[\hat{\mu}_n(\mathbf{X})] = \mu$ et $\mathrm{Var}_\mu(\hat{\mu}_n(\mathbf{X})) = 1/n$, il est également efficace.

3. Voyons ce qui peut se passer lorsque l'information de Fisher s'annule. On effectue un changement de paramètre dans l'exemple précédent en considérant le modèle $(\mathcal{N}(\theta^3, 1))_{\theta\in\mathbb{R}}$, où θ est le paramètre inconnu que l'on cherche à estimer. Avec les notations de la Proposition 23, on a la bijection $\mu = \theta^3 = \psi(\mu)$ avec ψ de classe C^1. Ce modèle est donc régulier, d'information de Fisher

$$J(\theta) = \psi'(\theta)^2 I(\psi(\theta)) = 9\theta^4,$$

qui est strictement positive si et seulement si θ est non nul. L'estimateur naturel (moments ou EMV) est

$$\hat{\theta}_n(\mathbf{X}) = \bar{X}_n^{1/3} = \left(\frac{X_1 + \cdots + X_n}{n}\right)^{1/3}.$$

On sait que

$$\sqrt{n}\left(\bar{X}_n - \theta^3\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

donc la Delta méthode telle qu'énoncée en Proposition 16 assure que, si $\theta \neq 0$,

$$\sqrt{n}\left(\hat{\theta}_n(\mathbf{X}) - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/(9\theta^4)) \qquad \text{avec} \qquad \frac{1}{9\theta^4} = \frac{1}{J(\theta)},$$

ce qui prouve l'efficacité asymptotique. Si $\theta = 0$, alors $\sqrt{n}\bar{X}_n \sim \mathcal{N}(0, 1)$. Soit Z une variable aléatoire distribuée selon une loi normale centrée réduite, alors la fonction $x \mapsto x^{1/3}$ étant continue, le résultat de continuité de la Proposition 5 affirme que

$$\sqrt{n}\bar{X}_n \xrightarrow[n\to\infty]{\mathcal{L}} Z \implies n^{1/6}\hat{\theta}_n(\mathbf{X}) \xrightarrow[n\to\infty]{\mathcal{L}} Z^{1/3}.$$

Ou encore, de façon équivalente : notons Y une variable telle que $Y^3 \sim \mathcal{N}(0, 1)$, alors

$$n^{1/6}\left(\hat{\theta}_n(\mathbf{X}) - 0\right) \xrightarrow[n\to\infty]{\mathcal{L}} Y.$$

La variable Y n'est pas gaussienne : elle est bimodale (voir figure 5.2), sa densité $f(y)$ pouvant se calculer comme suit

$$F(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Y^3 \leq y^3) = \Phi(y^3) \implies f(y) = \frac{3}{\sqrt{2\pi}}\, y^2 e^{-\frac{y^6}{2}}.$$

Bref, on a toujours convergence, mais la limite n'est plus gaussienne et la vitesse de convergence n'est plus en $n^{-1/2}$, mais en $n^{-1/6}$, donc bien plus lente.

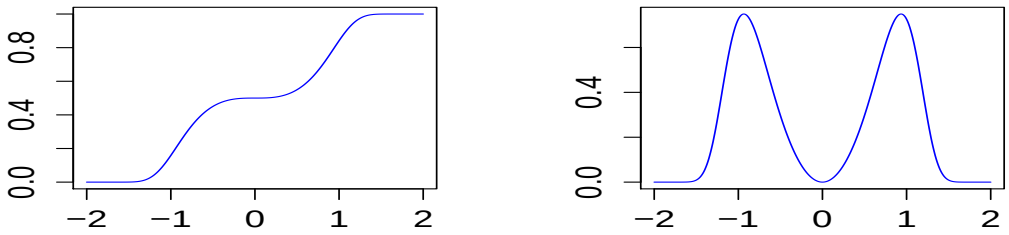

FIGURE 5.2 – Fonction de répartition et densité de la variable Y telle que $Y^3 \sim \mathcal{N}(0,1)$.

Un contexte important dans lequel l'efficacité asymptotique est vérifiée est celui des modèles exponentiels. Grâce à ce qui a déjà été dit, le résultat tombe maintenant comme un fruit mûr.

Proposition 29 (Modèles exponentiels et efficacité asymptotique)
Dans un modèle exponentiel général, si Θ est un intervalle ouvert sur lequel Q est C^1 et Q' ne s'annule pas, alors l'estimateur du maximum de vraisemblance est asymptotiquement efficace.

Preuve. D'une part, puisque Θ est ouvert, nous sommes dans le cadre d'application du Théorème 15 : pour tout paramètre θ, presque sûrement, pour n assez grand, l'EMV $\hat{\theta}_n(\mathbf{X})$ existe et est unique, vérifiant

$$\sqrt{n}\left(\hat{\theta}_n(\mathbf{X}) - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{Q'(\theta)^2 \mathrm{Var}_\theta(T(X))}\right),$$

où X dénote une variable générique de même loi que $X_1, \ldots, X_n$. D'autre part, d'après la Proposition 25, un tel modèle est régulier d'information de Fisher

$$I(\theta) = \mathrm{Var}_\theta(\ell'_\theta(X)) = Q'(\theta)^2 \mathrm{Var}_\theta(T(X)).$$

C'est exactement dire que $\hat{\theta}_n(\mathbf{X})$ est asymptotiquement efficace.

■

Mieux encore, on peut montrer (mais nous l'admettrons) un résultat général assurant l'efficacité asymptotique de l'estimateur du maximum de vraisemblance dans un modèle régulier.

Théorème 19 (EMV et efficacité asymptotique)
Soit un modèle régulier $(f_\theta)_{\theta\in\Theta}$ d'information de Fisher $I(\theta)$, soit $\theta_0 \in \Theta$ vérifiant $I(\theta_0) > 0$ et $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon iid selon f_{θ_0}. S'il existe une suite $(\hat{\theta}_n(\mathbf{X}))_{n\geq n_0}$ d'estimateurs du maximum de vraisemblance qui converge vers θ_0, alors

$$\sqrt{n}\left(\hat{\theta}_n(\mathbf{X}) - \theta_0\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/I(\theta_0)),$$

c'est-à-dire qu'on a efficacité asymptotique.

Pour reprendre la question posée en début de chapitre : "Existe-t-il un estimateur optimal, et si oui en quel sens ?" on peut dire que, du point de vue asymptotique dans le cadre des modèles réguliers, c'est l'estimateur du maximum de vraisemblance qui répond au problème (sous les réserves qui s'imposent : existence de l'EMV, non-nullité de l'information de Fisher, etc.). Encore faut-il pouvoir le calculer, ce qui n'est pas toujours chose facile...

Notant θ_0 la vraie valeur du paramètre, le Théorème 19 signifie que plus l'information de Fisher en ce point est grande, plus on peut estimer précisément θ_0, en particulier par l'estimateur au max de vraisemblance. Dit autrement, plus $I(\theta_0)$ est grande, plus l'information moyenne apportée par une donnée est importante : on peut par exemple écrire

$$\mathbb{P}\left(\hat{\theta}_n - \frac{2}{\sqrt{nI(\theta_0)}} \leq \theta_0 \leq \hat{\theta}_n + \frac{2}{\sqrt{nI(\theta_0)}}\right) \xrightarrow[n\to\infty]{} 0.95.$$

Noter que ceci ne correspond pas à un intervalle de confiance asymptotique à 95% : puisqu'on ne connaît pas θ_0, en général on ne connaît pas non plus $I(\theta_0)$.

Par ailleurs, on peut donner une interprétation graphique de l'information de Fisher grâce au lien avec la théorie de l'information[10]. On se contente d'en donner l'idée en considérant que tous les objets sont bien définis et suffisamment réguliers. Si f et g sont deux densités, on appelle divergence de Kullback-Leibler, ou entropie relative, de g par rapport à f la quantité

$$D(f \parallel g) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x)\, dx.$$

L'inégalité de Jensen assure que celle-ci est toujours positive, et nulle si et seulement si f et g sont égales presque partout. Stricto sensu, cette divergence ne peut cependant s'interpréter comme une distance puisque ni la symétrie ni l'inégalité triangulaire ne sont en général vérifiées. En terme d'inférence, supposons que θ_0 soit la vraie valeur du paramètre, alors pour une autre valeur θ, la divergence de f_θ à f_{θ_0} peut encore s'écrire

$$D(f_{\theta_0} \parallel f_\theta) = \mathbb{E}_{\theta_0}[\ell_{\theta_0}(X)] - \mathbb{E}_{\theta_0}[\ell_\theta(X)].$$

FIGURE 5.3 – Divergence et information de Fisher, avec $I(\theta_0)$ plus grande à droite qu'à gauche.

Sous les hypothèses de régularité ad hoc, on a donc au voisinage de θ_0

$$\ell_\theta(X) \approx \ell_{\theta_0}(X) + \ell'_{\theta_0}(X)(\theta - \theta_0) + \tfrac{1}{2}\ell''_{\theta_0}(\theta - \theta_0)^2.$$

Passant à l'espérance, puisque le score est centré, on en déduit que

$$D(f_{\theta_0} \parallel f_\theta) \approx \tfrac{1}{2} I(\theta_0)(\theta - \theta_0)^2.$$

10. voir [4] pour une introduction très lisible à ce domaine.

Autrement dit, l'information de Fisher en θ_0 correspond à la courbure de la divergence de Kullback-Leibler au voisinage de θ_0. Plus cette courbure est importante, plus il est facile de discriminer entre la vraie valeur θ_0 et une valeur voisine, et inversement. La figure 5.3 illustre ce point de vue.

L'interprétation précédente permet également de comprendre pourquoi l'estimation au maximum de vraisemblance apparaît de façon naturelle dans ce cadre. Le but est en effet de trouver la valeur de θ qui minimise la divergence $D(f_{\theta_0} \parallel f_\theta)$, c'est-à-dire qui maximise la fonction $\theta \mapsto \mathbb{E}_{\theta_0}[\ell_\theta(X)]$, dite fonction de contraste. Celle-ci étant hors d'atteinte, l'idée est de maximiser sa version empirique : en effet, par la Loi des Grands Nombres, si les X_i sont iid de densité f_{θ_0}, alors

$$\frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i) \xrightarrow[n\to\infty]{p.s.} \mathbb{E}_{\theta_0}[\ell_\theta(X)].$$

Or maximiser le terme de gauche, c'est justement ce que fait l'estimateur au maximum de vraisemblance.

Annexe A

Annales

Université Pierre et Marie Curie
Master Mathématiques et Applications
Arnaud Guyader

Mardi 5 mai 2015
Aucun document autorisé
Durée : 3 heures

Statistique mathématique

Exercice 1 (2 points)
On observe n variables aléatoires $X_1, \dots, X_n$ iid de densité

$$\lambda x^{\lambda-1} \mathbb{1}_{]0,1[}(x),$$

où $\lambda > 0$ est inconnu. On suppose que la loi a priori de λ est exponentielle de paramètre 1.

1. Montrer que la loi a posteriori de λ est une loi Gamma (de paramètres à préciser en fonction de n et de $x_1, \dots, x_n$).
2. En déduire l'estimateur de Bayes $\tilde{\lambda}_n$ de λ pour le risque quadratique.

Exercice 2 (3 points)
Soit $\theta \in]-1, 1[$ un paramètre inconnu et $X_1, \dots, X_n$ un échantillon iid de densité

$$f_\theta(x) = \frac{1}{2}(1 + \theta x)\, \mathbb{1}_{]-1,1[}(x).$$

1. Donner un estimateur $\hat{\theta}_n$ de θ par la méthode des moments.
2. Donner le biais de $\hat{\theta}_n$ ainsi que son risque quadratique.
3. Déterminer la loi limite de $\hat{\theta}_n$ (correctement centré et normalisé).
4. En déduire un intervalle de confiance bilatère asymptotique de niveau $(1 - \alpha)$.
5. On veut tester $H_0 : \theta \geq 0$ contre $H_1 : \theta < 0$. Proposer un test de niveau asymptotique α.

Exercice 3 (4 points)
On considère le modèle linéaire

$$Y_i = \beta_1 w_i + \beta_2 z_i + \varepsilon_i \qquad i = 1, \dots, n,$$

où les w_i et les z_i sont des réels fixés connus, les ε_i des variables aléatoires iid de loi $\mathcal{N}(0,\sigma^2)$, et β_1, β_2, σ^2 des paramètres réels inconnus. On note W, Y et Z les vecteurs colonnes de $\mathbb{R}^n$ de coordonnées respectives (w_i), (Y_i) et (z_i). On note $\langle\cdot,\cdot\rangle$ le produit scalaire usuel et $\|\cdot\|$ la norme euclidienne de $\mathbb{R}^n$. Dans tout l'exercice, on suppose que $\langle W,Z\rangle^2 \neq \|W\|^2\|Z\|^2$.

1. (a) Donner l'estimateur $\hat\beta = [\hat\beta_1,\hat\beta_2]'$ des moindres carrés de $\beta = [\beta_1,\beta_2]'$ en fonction des produits scalaires et des normes des vecteurs W, Y et Z.
   (b) Quelle est la loi de $\hat\beta = [\hat\beta_1,\hat\beta_2]'$?
   (c) Que dire de la dépendance entre $\hat\beta_1$ et $\hat\beta_2$ si W et Z sont orthogonaux ?
2. Dans ce qui suit, on ne suppose pas l'orthogonalité de W et Z. On suppose que $\|W\| = \|Z\| = 1$ et on note θ l'angle formé par les deux vecteurs, c'est-à-dire que $\langle W,Z\rangle = \cos\theta$.
   (a) Donner la loi de la variable aléatoire $U = |\sin\theta|(\hat\beta_2 - \beta_2)/\hat\sigma$, où $\hat\sigma^2$ est l'estimateur habituel (sans biais) de σ^2 dans le modèle de régression.
   (b) En déduire un intervalle de confiance (exact et bilatère) de β_2 de niveau $(1-\alpha)$.
   (c) Soit $L^2(\theta)$ la variable aléatoire égale au carré de la longueur de l'intervalle de confiance trouvé en question précédente. Etudier la fonction $g(\theta) = \mathbb{E}[L^2(\theta)]$ pour $\theta \in]0,\pi/2[$. Donner le comportement de g en 0 et $\pi/2$ et interpréter ces résultats.

Exercice 4 (6 points)
On considère la densité suivante par rapport à la mesure de Lebesgue :

$$f_\theta(x) = \theta^2 x e^{-\theta x}\mathbb{1}_{[0,+\infty[}(x),$$

où $\theta > 0$ est un paramètre que l'on cherche à estimer.

1. Est-ce un modèle exponentiel ?
2. Ce modèle est-il régulier ? Si oui, calculer l'information de Fisher $I_1(\theta)$ du modèle à une observation.
3. On observe $X_1,\dots,X_n$ iid de densité f_θ. Donner une statistique exhaustive du modèle.
4. Déterminer l'estimateur du maximum de vraisemblance $\hat\theta_n$. Est-il asymptotiquement efficace ?
5. Montrer que si Y suit une loi Gamma de paramètres (r,λ), avec $\lambda > 0$ et $r > 2$, alors $\mathbb{E}[1/Y] = \lambda/(r-1)$ et $\mathbb{E}[1/Y^2] = \lambda^2/((r-1)(r-2))$.
6. Donner la loi de $\bar X_n = (\sum_{i=1}^n X_i)/n$. Pour $n \geq 2$, que valent $\mathbb{E}[\hat\theta_n]$ et $\mathrm{Var}(\hat\theta_n)$?
7. Déduire de la question précédente un estimateur $\tilde\theta_n$ sans biais pour θ. Est-il efficace ? Est-il asymptotiquement efficace ?

Exercice 5 (5 points)
Soit F une fonction de répartition sur $\mathbb{R}$ et soit $\theta \in \mathbb{R}_+$ un paramètre inconnu. On dispose d'un échantillon iid $(X_1,\dots,X_n)$ de fonction de répartition $\mathbb{P}(X_i \leq x) = F_\theta(x) = F(x-\theta)$. On considère la variable aléatoire

$$W_n = \sum_{i=1}^n \mathbb{1}_{\{X_i>0\}}.$$

1. On suppose F connue.
   (a) Pour $n \in \mathbb{N}^\star$ fixé, donner la loi de W_n.
   (b) Pour une constante p à préciser, montrer que la loi limite de $(W_n - np)/\sqrt{n}$, lorsque $n \to \infty$, est gaussienne. Préciser la moyenne et la variance de cette loi limite.
2. On suppose désormais F inconnue. On sait cependant qu'elle est continue et symétrique, c'est-à-dire que $F(-x) = 1 - F(x)$ pour tout réel x.

(a) Que vaut $F(0)$? On souhaite tester l'hypothèse $H_0 : \theta = 0$ contre $H_1 : \theta > 0$. Proposer un test de niveau asymptotique α basé sur W_n.

(b) Quelle est la p-value de ce test asymptotique si $n = 100$ et $W_n = -10$ (on rappelle que si $Z \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(Z > 3) \approx 0.001$) ?

3. On suppose toujours F inconnue, continue et symétrique. On suppose de plus qu'elle est dérivable, de dérivée f strictement positive sur $\mathbb{R}$ (inconnue elle aussi !).

(a) Quelle est la médiane de X_1 ? En déduire un estimateur $\hat{\theta}_n$ de θ, ainsi que la propriété de normalité asymptotique qu'il vérifie.

(b) Peut-on déduire de ce résultat un intervalle de confiance asymptotique pour θ ?

(c) Notons $F_\theta^n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$ pour tout réel x. Rappeler le résultat de normalité asymptotique vérifié par $F_\theta^n(x)$. En déduire un intervalle de confiance asymptotique bilatère de niveau 95% pour θ (rappel : si $Z \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(|Z| \leq 2) \approx 0.95$).

Université Pierre et Marie Curie
Master Mathématiques et Applications
Arnaud Guyader

Mardi 5 mai 2015
Aucun document autorisé
Durée : 3 heures

Statistique mathématique

Corrigé

Exercice 1
On observe n variables aléatoires $X_1, \dots, X_n$ iid de densité

$$\lambda x^{\lambda-1} \mathbb{1}_{]0,1[}(x),$$

où $\lambda > 0$ est inconnu. On suppose que la loi a priori de λ est exponentielle de paramètre 1.

1. Les réalisations $x_1, \dots, x_n$ de $]0,1[$ étant connues, la densité a posteriori s'écrit

$$f(\lambda|x_1, \dots, x_n) = \frac{f(\lambda) f(x_1, \dots, x_n|\lambda)}{f(x_1, \dots, x_n)} \propto f(\lambda) \prod_{i=1}^n f(x_i|\lambda),$$

où le symbole $\propto$, "proportionnel à", signifie qu'il existe une constante C indépendante de λ (mais dépendant des x_i) telle que

$$f(\lambda|x_1, \dots, x_n) = C \times f(\lambda) \prod_{i=1}^n f(x_i|\lambda).$$

On en déduit que

$$f(\lambda|x_1, \dots, x_n) \propto e^{-\lambda} \left(\prod_{i=1}^n \lambda x_i^{\lambda-1} \right) \mathbb{1}_{\lambda>0} \propto \lambda^n \exp\left(- \left(1 - \sum_{i=1}^n \log x_i \right) \lambda \right) \mathbb{1}_{\lambda>0},$$

d'où l'on déduit que la loi a posteriori de λ est une loi $\Gamma(n+1, 1 - \sum_{i=1}^n \log x_i)$.

2. L'estimateur de Bayes pour le risque quadratique correspond à la moyenne a posteriori, c'est-à-dire à l'espérance conditionnelle de λ sachant $X_1, \dots, X_n$, donc à l'espérance d'une loi Gamma :

$$\tilde{\lambda}_n = \frac{n+1}{1 - \sum_{i=1}^n \log X_i}.$$

Exercice 2
Soit $\theta \in]-1,1[$ un paramètre inconnu et $X_1, \dots, X_n$ un échantillon iid de densité

$$f_\theta(x) = \frac{1}{2}(1 + \theta x) \mathbb{1}_{]-1,1[}(x).$$

1. Un calcul facile montrant que $\mathbb{E}[X_1] = \theta/3$, on considère donc l'estimateur $\hat{\theta}_n = 3\bar{X}_n$.
2. Clairement $\hat{\theta}_n$ est sans biais. Le risque quadratique vaut donc

$$R(\hat{\theta}_n, \theta) = \mathrm{Var}(\hat{\theta}_n) = \frac{9}{n}\mathrm{Var}(X_1) = \frac{9}{n}(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2) = \frac{9}{n}\left(\frac{1}{3} - \frac{\theta^2}{9}\right) = \frac{3-\theta^2}{n}.$$

3. Le TCL implique

$$\sqrt{n}\left(\bar{X}_n - \mathbb{E}[X_1]\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(X_1)) \iff \sqrt{n}\left(\bar{X}_n - \frac{\theta}{3}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, (3-\theta^2)/9).$$

Puisque $\hat{\theta}_n = 3\bar{X}_n$, il suffit alors d'appliquer la Delta-méthode avec $\varphi(x) = 3x$ pour aboutir à

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 3-\theta^2).$$

4. La convergence en loi précédente assure en particulier que $\hat{\theta}_n$ converge en probabilité vers θ et on déduit alors du Théorème de Slutsky que

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{3-\hat{\theta}_n^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

En notant $q_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la loi normale centrée réduite, on a donc

$$\mathbb{P}\left(\hat{\theta}_n - \frac{q_{1-\alpha/2}\sqrt{3-\hat{\theta}_n^2}}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + \frac{q_{1-\alpha/2}\sqrt{3-\hat{\theta}_n^2}}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha,$$

ce qui est exactement dire qu'un intervalle de confiance bilatère asymptotique de niveau $(1-\alpha)$ est donné par

$$\left[\hat{\theta}_n - \frac{q_{1-\alpha/2}\sqrt{3-\hat{\theta}_n^2}}{\sqrt{n}}; \hat{\theta}_n + \frac{q_{1-\alpha/2}\sqrt{3-\hat{\theta}_n^2}}{\sqrt{n}}\right].$$

Remarque : si on tient à tout prix à se compliquer la vie, on peut aussi appliquer la méthode de stabilisation de la variance.

5. Le test $H_0 : \theta \geq 0$ contre $H_1 : \theta < 0$ part de la même idée, en considérant un intervalle de confiance asymptotique unilatère. Soit $q_{1-\alpha} = \Phi^{-1}(1-\alpha)$ le quantile d'ordre $(1-\alpha)$ de la loi normale centrée réduite. Pour obtenir un test de niveau asymptotique α, il suffit de rejeter H_0 si

$$\hat{\theta}_n + \frac{q_{1-\alpha}\sqrt{3-\hat{\theta}_n^2}}{\sqrt{n}} < 0.$$

Exercice 3
On considère le modèle linéaire

$$Y_i = \beta_1 w_i + \beta_2 z_i + \varepsilon_i \qquad i = 1, \ldots, n,$$

où les w_i et les z_i sont des réels fixés connus, les ε_i des variables aléatoires iid de loi $\mathcal{N}(0, \sigma^2)$, et β_1, β_2, σ^2 des paramètres réels inconnus. On note W, Y et Z les vecteurs colonnes de $\mathbb{R}^n$ de coordonnées respectives (w_i), (Y_i) et (z_i). On note $\langle \cdot, \cdot \rangle$ le produit scalaire usuel et $\| \cdot \|$ la norme euclidienne de $\mathbb{R}^n$. Dans tout l'exercice, on suppose que $\langle W, Z\rangle^2 \neq \|W\|^2\|Z\|^2$.

1. (a) En notant $X = [W|Z]$ la matrice $n\times 2$ de colonnes W et Z, on sait que $\hat{\beta} = (X'X)^{-1}X'Y$, c'est-à-dire

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \frac{1}{\|W\|^2\|Z\|^2 - \langle W, Z\rangle^2} \begin{bmatrix} \langle W, Y\rangle\|Z\|^2 - \langle W, Z\rangle\langle Y, Z\rangle \\ \langle Z, Y\rangle\|W\|^2 - \langle Z, W\rangle\langle Y, W\rangle \end{bmatrix}.$$

(b) On sait que $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2]' \sim \mathcal{N}(0, \sigma^2 (X'X)^{-1})$, avec

$$(X'X)^{-1} = \frac{1}{\|W\|^2\|Z\|^2 - \langle W, Z\rangle^2} \begin{bmatrix} \|Z\|^2 & -\langle W, Z\rangle \\ -\langle Z, W\rangle & \|W\|^2 \end{bmatrix}.$$

(c) Si W et Z sont orthogonaux, c'est-à-dire si $\langle W, Z\rangle = 0$, alors d'après la question précédente $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = 0$. Puisque le vecteur $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2]'$ est gaussien d'après la question précédente, ceci implique que $\hat{\beta}_1$ et $\hat{\beta}_2$ sont indépendants.

2. Dans ce qui suit, on ne suppose pas l'orthogonalité de W et Z. On suppose que $\|W\| = \|Z\| = 1$ et on note θ l'angle formé par les deux vecteurs, c'est-à-dire que $\langle W, Z\rangle = \cos\theta$.

(a) D'après le cours,

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_{2,2}}} \sim \mathcal{T}_{n-2},$$

or puisque $\|W\| = \|Z\| = 1$ et $\langle W, Z\rangle = \cos\theta$,

$$\sqrt{[(X'X)^{-1}]_{2,2}} = \sqrt{\frac{\|W\|^2}{\|W\|^2\|Z\|^2 - \langle W, Z\rangle^2}} = \sqrt{\frac{1}{1 - (\cos\theta)^2}} = \frac{1}{|\sin\theta|}.$$

Au total, on a donc

$$U = \frac{|\sin\theta|(\hat{\beta}_2 - \beta_2)}{\hat{\sigma}} \sim \mathcal{T}_{n-2},$$

loi de Student à $(n-2)$ degrés de liberté.

(b) En notant $t_{1-\alpha/2}$ le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-2)$ degrés de liberté, un intervalle de confiance exact et bilatère de β_2 de niveau $(1-\alpha)$ est donc

$$\left[\hat{\beta}_2 - \frac{t_{1-\alpha/2}\ \hat{\sigma}}{|\sin\theta|}; \hat{\beta}_2 + \frac{t_{1-\alpha/2}\ \hat{\sigma}}{|\sin\theta|}\right].$$

(c) Puisque $\hat{\sigma}^2$ est un estimateur sans biais de σ^2, il vient

$$L^2(\theta) = \frac{4t_{1-\alpha/2}^2\ \hat{\sigma}^2}{(\sin\theta)^2} \implies g(\theta) = \mathbb{E}\left[L^2(\theta)\right] = \frac{4t_{1-\alpha/2}^2\ \sigma^2}{(\sin\theta)^2}.$$

La fonction g est donc continue décroissante sur $]0, \pi/2[$, avec

$$\lim_{\theta\to 0} g(\theta) = +\infty \qquad \text{et} \qquad \lim_{\theta\to\pi/2} g(\theta) = 4t_{1-\alpha/2}^2\ \sigma^2 = \inf_{\theta\in]0,\pi/2[} g(\theta).$$

L'estimateur $\hat{\beta}_2$ est d'autant plus précis que les vecteurs Y et Z sont orthogonaux. Lorsque $\sin\theta = 0$, la matrice X est de rang 1 et le modèle n'est pas identifiable : il est impossible de donner la moindre information pertinente sur β_2 au vu des observations.

Exercice 4

On considère la densité suivante par rapport à la mesure de Lebesgue :

$$f_\theta(x) = \theta^2 x e^{-\theta x} \mathbb{1}_{[0,+\infty[}(x),$$

où $\theta > 0$ est un paramètre que l'on cherche à estimer.

1. C'est bien un modèle exponentiel puisque f_θ a la forme

$$f_t(x) = C(\theta)h(x)e^{Q(\theta)T(x)},$$

avec $Q(\theta) = -\theta$ et $T(x) = x$.

2. Puisque Q est C^1 sur l'intervalle ouvert $\Theta =]0, +\infty[$, le modèle est régulier. Le modèle étant exponentiel, l'information de Fisher pour une observation vaut

$$I_1(\theta) = Q'(\theta)^2 \mathrm{Var}(T(X)) = \mathrm{Var}(X) = \frac{2}{\theta^2},$$

car $X \sim \Gamma(2, \theta)$.

3. Puisque le modèle est exponentiel, une statistique exhaustive est donnée par

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

4. Toujours par l'exponentialité du modèle, l'estimateur du maximum de vraisemblance est solution de

$$\mathbb{E}_{\hat{\theta}_n}[T(X)] = \frac{1}{n}\sum_{i=1}^{n} T(X_i) \iff \frac{2}{\hat{\theta}_n} = \bar{X}_n \iff \hat{\theta}_n = \frac{2}{\bar{X}_n}.$$

Puisque Q est C^1 et ne s'annule pas sur l'intervalle ouvert $\Theta =]0, +\infty[$, le résultat du cours assure que l'EMV $\hat{\theta}_n$ est asymptotiquement efficace.

5. Il suffit de faire apparaître des densités de lois Gamma et d'appliquer les relations

$$\Gamma(r) = (r-1)\Gamma(r-1) = (r-1)(r-2)\Gamma(r-2),$$

valables pour tout $r > 2$. Le Théorème de Transfert donne ainsi

$$\mathbb{E}[1/Y] = \int_0^{+\infty} \frac{1}{y}\frac{(\lambda y)^{r-1}}{\Gamma(r)}\lambda e^{-\lambda y} dy = \frac{\lambda}{r-1}\int_0^{+\infty} \frac{(\lambda y)^{r-2}}{\Gamma(r-1)}\lambda e^{-\lambda y} dy = \frac{\lambda}{r-1},$$

puisque la quantité intégrée n'est rien d'autre que la densité d'une loi $\Gamma(r-1, \lambda)$. La seconde relation s'obtient de la même façon.

6. Puisque les X_i sont iid de loi $\Gamma(2, \theta)$, un résultat classique des lois Gamma implique que $\sum_{i=1}^{n} X_i \sim \Gamma(2n, \theta)$. Un autre résultat classique des lois Gamma implique alors que

$$\bar{X}_n \sim \Gamma(2n, n\theta).$$

Si $n \geq 2$ alors $2n > 2$ et on peut appliquer les résultats de la question précédente :

$$\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[2/\bar{X}_n] = 2\mathbb{E}[1/\bar{X}_n] = \frac{2n}{2n-1}\theta,$$

et

$$\mathrm{Var}(\hat{\theta}_n) = 4\mathrm{Var}(1/\bar{X}_n) = 4(\mathbb{E}[1/\bar{X}_n^2] - \mathbb{E}[1/\bar{X}_n]^2) = \frac{2n^2}{(n-1)(2n-1)^2}\theta^2.$$

7. Un estimateur sans biais pour θ est donc

$$\tilde{\theta}_n = \frac{2n-1}{2n}\hat{\theta}_n,$$

de variance

$$\mathrm{Var}(\tilde{\theta}_n) = \left(\frac{2n-1}{2n}\right)^2 \mathrm{Var}(\hat{\theta}_n) = \frac{\theta^2}{2(n-1)} > \frac{\theta^2}{2n} = \frac{1}{nI_1(\theta)}.$$

Cet estimateur n'est donc pas efficace. Par contre, il hérite de $\hat{\theta}_n$ la propriété d'efficacité asymptotique puisque

$$\sqrt{n}(\tilde{\theta}_n - \theta) = \sqrt{n}(\hat{\theta}_n - \theta) - \frac{\hat{\theta}_n}{2\sqrt{n}}.$$

Or $\hat{\theta}_n$ est asymptotiquement efficace donc

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/I_1(\theta)),$$

d'où en particulier

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

et par Slutsky

$$\frac{\hat{\theta}_n}{2\sqrt{n}} \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

et toujours par Slutsky

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/I_1(\theta)),$$

ce qui montre le résultat annoncé.

Exercice 5

Soit F une fonction de répartition sur $\mathbb{R}$ et soit $\theta \in \mathbb{R}_+$ un paramètre inconnu. On dispose d'un échantillon iid $(X_1, \ldots, X_n)$ de fonction de répartition $\mathbb{P}(X_i \leq x) = F_\theta(x) = F(x-\theta)$. On considère la variable aléatoire

$$W_n = \sum_{i=1}^{n} \mathbb{1}_{\{X_i>0\}}.$$

1. On suppose F connue.
   (a) Pour $n \in \mathbb{N}^\star$ fixé, la variable W_n est la somme des variables aléatoires iid $\mathbb{1}_{\{X_i>0\}}$, qui sont des variables de Bernoulli, donc $W_n \sim \mathcal{B}(n,p)$ loi binomiale de paramètres n et p avec
   $$p = \mathbb{P}(\mathbb{1}_{\{X_1>0\}} = 1) = \mathbb{P}(X_1 > 0) = 1 - \mathbb{P}(X_1 \leq 0) = 1 - F(-\theta).$$
   (b) Puisque W_n est une somme de variables iid qui ont pour moyenne p et pour variance $p(1-p)$, le TCL assure que
   $$\frac{W_n - np}{\sqrt{n}} = \sqrt{n}\left(\frac{W_n}{n} - p\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, p(1-p)).$$
2. On suppose désormais F inconnue. On sait cependant qu'elle est continue et symétrique, c'est-à-dire que $F(-x) = 1 - F(x)$ pour tout réel x.
   (a) Pour $x = 0$, la relation de symétrie donne $F(0) = 1-F(0)$ donc $F(0) = 1/2$. On souhaite tester l'hypothèse $H_0 : \theta = 0$ contre $H_1 : \theta > 0$. Avec les notations précédentes, sous H_0, on a $p = 1/2$ et
   $$2\,\frac{W_n - n/2}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$
   En notant $q_{1-\alpha} = \Phi^{-1}(1-\alpha)$ le quantile d'ordre $(1-\alpha)$ de la loi normale centrée réduite, on a sous H_0
   $$\mathbb{P}\left(W_n > \frac{n}{2} + q_{1-\alpha}\frac{\sqrt{n}}{2}\right) \xrightarrow[n\to\infty]{} \alpha.$$
   On rejette H_0 si $W_n > n/2 + q_{1-\alpha}\sqrt{n}/2$, ce qui fournit un test de niveau asymptotique α basé sur W_n.
   (b) La p-value α_0 vérifie
   $$W_n = \frac{n}{2} + q_{1-\alpha_0}\frac{\sqrt{n}}{2} \iff q_{1-\alpha_0} = -12 \iff \alpha_0 = \Phi(12) \implies \alpha_0 > 0.999.$$
   En particulier, il est clair qu'on accepte H_0.

3. On suppose toujours F inconnue, continue et symétrique. On suppose de plus qu'elle est dérivable, de dérivée f strictement positive sur $\mathbb{R}$.

   (a) La médiane de X_1 est $x_{1/2} = \theta$. On peut donc prendre comme estimateur de θ la médiane empirique $\hat{\theta}_n = x_{1/2}(n)$. Puisque $F_\theta(x) = F(x-\theta)$ avec F dérivable, les variables $X_1, \ldots, X_n$ admettent pour densité commune

   $$f_\theta(x) = (F_\theta(x))' = (F(x-\theta))' = f(x-\theta)$$

   et en particulier $f_\theta(\theta) = f(0)$. Par le résultat du cours, on a alors

   $$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/(4f_\theta(\theta)^2)) = \mathcal{N}(0, 1/(4f(0)^2)).$$

   (b) Puisqu'on ne connaît pas $f(0)$, on ne peut pas déduire de ce résultat un intervalle de confiance asymptotique pour θ.

   (c) Notons $F_\theta^n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$. On sait que, pour tout réel x,

   $$\sqrt{n}\left(F_\theta^n(x) - F_\theta(x)\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, F_\theta(x)(1-F_\theta(x))).$$

   En particulier, pour $x = \theta$, ceci donne

   $$\sqrt{n}\left(F_\theta^n(\theta) - \frac{1}{2}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/4).$$

   En considérant $\Phi^{-1}(0.975) \approx 2$, ceci implique en particulier que

   $$\mathbb{P}\left(\frac{1}{2} - \frac{1}{\sqrt{n}} \leq F_\theta^n(\theta) \leq \frac{1}{2} + \frac{1}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

   La croissance de $(F_\theta^n)^{-1}$ donne alors un intervalle de confiance bilatère de niveau asymptotique $(1-\alpha)$ pour θ :

   $$\left[(F_\theta^n)^{-1}\left(\frac{1}{2} - \frac{1}{\sqrt{n}}\right), (F_\theta^n)^{-1}\left(\frac{1}{2} + \frac{1}{\sqrt{n}}\right)\right].$$

   En notant $X_{(k)}$ la statistique d'ordre k de l'échantillon, un intervalle de confiance asymptotique bilatère de niveau 95% pour θ est donc $[X_{(i)}, X_{(j)}]$, avec $i = \lceil n/2 - \sqrt{n} \rceil$ et $j = \lceil n/2 + \sqrt{n} \rceil$.

Université Pierre et Marie Curie
Master Mathématiques et Applications
Arnaud Guyader

Mardi 3 mai 2016
Aucun document autorisé
Durée : 3 heures

Statistique mathématique

Exercice 1 (10 points)
On considère la densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$ définie par

$$f(y) = \tfrac{3}{2}(1-|y|)^2 \mathbb{1}_{[-1,1]}(y).$$

1. Représenter f. Si la variable Y a pour densité f, que valent son espérance, sa variance et la médiane de sa loi ? Pour $0 \leq x \leq 1$, déterminer $\mathbb{P}(Y \geq 1-x)$.

   Dans toute la suite, on suppose disposer de n observations iid $X_1, \ldots, X_n$ de densité $f_\theta(x) = f(x-\theta)$, où θ est un paramètre réel inconnu.

2. Représenter f_θ. Que vaut $\mathbb{E}[X_1]$? En déduire un estimateur $\hat{\theta}_{M,n}$ de θ par la méthode des moments. Déterminer la loi limite (après une renormalisation convenable) de $\hat{\theta}_{M,n} - \theta$.

3. Quelle est la médiane de la loi de X_1 ? En déduire un estimateur $\hat{\theta}_{Q,n}$ de θ et préciser la loi limite (après une renormalisation convenable) de $\hat{\theta}_{Q,n} - \theta$.

4. Soit $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Pour $0 \leq x \leq 1$, que vaut la probabilité $\mathbb{P}(X_{(n)} \leq \theta + 1 - x)$ (on pourra se servir de la question 1) ?

5. En déduire que $\hat{\theta}_{S,n} = X_{(n)} - 1$ est un estimateur convergent de θ. Est-il sans biais ?

6. Pour $u \geq 0$, rappeler ce que vaut $\lim_{n\to\infty}(1-u/n)^n$. En déduire que, pour un réel $\beta > 0$ que l'on précisera, $\mathbb{P}(\hat{\theta}_{S,n} \leq \theta - tn^{-\beta})$ converge, pour tout $t \geq 0$, vers une limite non nulle que l'on calculera.

7. En déduire la loi limite (fonction de répartition et densité) de $n^\beta(\theta - \hat{\theta}_{S,n})$.

8. Le modèle statistique défini par la famille de densités $(f_\theta)_{\theta\in\mathbb{R}}$ est-il régulier ? Si oui, calculer l'information de Fisher $I_1(\theta)$ du modèle à une observation.

9. Tout comme $X_{(n)} - 1$ (cf. question 5), on admettra que $X_{(1)} + 1$ est un estimateur convergent de θ. Calculer la vraisemblance des observations et montrer qu'elle s'annule en dehors d'un intervalle que l'on précisera. En déduire l'existence et la convergence d'un estimateur du maximum de vraisemblance $\hat{\theta}_{V,n}$ (on ne cherchera pas à le calculer). Préciser la loi limite (après une renormalisation convenable) de $\hat{\theta}_{V,n} - \theta$.

10. Si vous avez un grand nombre d'observations, lequel des estimateurs précédents choisir ?

11. **Bonus** : On dispose au contraire d'un petit nombre n d'observations et on veut trouver un intervalle de confiance bilatère pour θ au niveau $(1-\alpha)$. En utilisant un des estimateurs précédents mais sans appliquer d'approximation asymptotique, comment trouver un intervalle de confiance aussi petit que possible ?

Exercice 2 (4 points)
Pour $n \geq 4$, on considère le modèle linéaire

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \quad i = 1, \ldots, n$$

où les vecteurs X_j, $1 \leq j \leq 3$, sont déterministes, connus et appartiennent à $\mathbb{R}^n$, et ε est un vecteur gaussien inconnu centré de matrice de covariance $\sigma^2 I_n$. De plus, on suppose

$$\|X_j\| = 1, \ \ 1 \leq j \leq 3, \ \ \langle X_j, X_{j'} \rangle = 1/2, \ \ 1 \leq j, j' \leq 3 \text{ avec } j \neq j'.$$

Les paramètres réels β_j, $1 \leq j \leq 3$, et σ^2 sont supposés inconnus.

1. Pour un $q \in \mathbb{N} \backslash \{0\}$, calculer le produit AB des matrices $q \times q$ suivantes

$$A = \begin{bmatrix} 1 & 1/2 & \ldots & 1/2 \\ 1/2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1/2 \\ 1/2 & \ldots & 1/2 & 1 \end{bmatrix} ; \quad B = \begin{bmatrix} q & -1 & \ldots & -1 \\ -1 & q & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \ldots & -1 & q \end{bmatrix}.$$

2. Calculer les estimateurs des moindres carrés $\hat{\beta}$ et $\hat{\sigma}^2$. Déterminer la loi jointe de $(\hat{\beta}, \hat{\sigma}^2)$.
3. Donner un intervalle de confiance pour β_1 de niveau $1 - \alpha$. Quelle est l'espérance du carré de sa longueur ?
4. Donner un test de niveau α pour l'hypothèse nulle H_0 : "$\beta_2 = \beta_3 = 0$".

Exercice 3 (2 points)
On dit que la variable Λ suit une loi Bêta de paramètres a et b strictement positifs si elle a pour densité

$$f(\lambda) = \tfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \ \lambda^{a-1}(1-\lambda)^{b-1} \mathbb{1}_{[0,1]}(\lambda).$$

On rappelle que l'espérance d'une telle loi Bêta est $a/(a+b)$. On observe n variables aléatoires $X_1, \ldots, X_n$ iid de loi de Bernoulli de paramètre $\lambda \in [0,1]$ inconnu, c'est-à-dire que pour $x_i \in \{0,1\}$,

$$\mathbb{P}(X_i = x_i) = \lambda^{x_i}(1-\lambda)^{1-x_i}.$$

On suppose que la loi a priori de λ est une loi Bêta de paramètres a et b.

1. Déterminer la loi a posteriori de λ.
2. En déduire l'estimateur de Bayes $\tilde{\lambda}_n$ de λ pour la perte quadratique.

Exercice 4 (4 points)
On définit le carré de la distance de Hellinger h entre deux densités de probabilité f et g sur $\mathbb{R}$ par

$$h^2(f,g) = \frac{1}{2} \int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx.$$

1. Montrer que

$$h^2(f,g) = 1 - \int_{\mathbb{R}} \sqrt{f(x)g(x)} \, dx.$$

   Vérifier que h est une distance sur l'ensemble des densités sur $\mathbb{R}$, avec de plus $0 \leq h(f,g) \leq 1$.
2. Soit $\theta > 0$ et f_θ la densité de la loi uniforme sur $[0, \theta]$. Montrer que $h^2(f_\theta, f_{\theta'}) = 1 - \sqrt{\frac{\theta}{\theta'}}$ si $\theta \leq \theta'$.
3. Soit $X_1, \ldots, X_n$ iid de loi uniforme sur $[0, \theta]$ pour un certain $\theta > 0$. Déterminer $\hat{\theta}$, estimateur du maximum de vraisemblance de θ. Préciser la densité de $\hat{\theta}$ et en déduire $\mathbb{E}_\theta \left[\sqrt{\hat{\theta}} \right]$.
4. Montrer que $\mathbb{E}_\theta \left[h^2(f_{\hat{\theta}}, f_\theta) \right] = \frac{1}{2n+1}$. Comparer au risque quadratique $\mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right]$.

Université Pierre et Marie Curie
Master Mathématiques et Applications
Arnaud Guyader

Mardi 3 mai 2016
Aucun document autorisé
Durée : 3 heures

Statistique mathématique

Corrigé

Exercice 1
On considère la densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$ définie par

$$f(y) = \tfrac{3}{2}(1-|y|)^2 \mathbb{1}_{[-1,1]}(y).$$

1. Si la variable Y a pour densité f, alors par symétrie par rapport à 0, sa moyenne et sa médiane valent 0, et un calcul élémentaire montre que sa variance vaut $1/10$. Pour $0 \leq x \leq 1$, un autre calcul élémentaire donne

$$\mathbb{P}(Y \geq 1-x) = \frac{x^3}{2}.$$

2. C'est un modèle de translation donc $\mathbb{E}[X_1] = \theta$ et $\mathrm{Var}(X_1) = 1/10$. Un estimateur par la méthode des moments de θ est donc $\hat{\theta}_{M,n} = \bar{X}_n$. Par le TCL, il s'ensuit que

$$\sqrt{n}\left(\hat{\theta}_{M,n} - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/10).$$

3. La médiane de la loi de X_1 est θ. Un estimateur $\hat{\theta}_{Q,n}$ de θ est donc la médiane empirique, encore notée $\hat{\theta}_{Q,n} = x_{1/2}(n)$, et on a

$$\sqrt{n}\left(\hat{\theta}_{Q,n} - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/9).$$

4. Soit $X_{(n)} = \max_{1\leq i\leq n} X_i$. Pour $0 \leq x \leq 1$, on a

$$\mathbb{P}(X_{(n)} \leq \theta + 1 - x) = \left(1 - \tfrac{x^3}{2}\right)^n.$$

5. Puisque $\hat{\theta}_{S,n} = X_{(n)} - 1 < \theta$ presque sûrement, on a pour tout $\varepsilon \in]0,1]$

$$\mathbb{P}\left(|\hat{\theta}_{S,n} - \theta| \geq \varepsilon\right) = \mathbb{P}\left(\hat{\theta}_{S,n} \leq \theta - \varepsilon\right) = \left(1 - \tfrac{\varepsilon^3}{2}\right)^n \xrightarrow[n\to\infty]{} 0,$$

ce qui assure que $\hat{\theta}_{S,n}$ converge en probabilité vers θ. Néanmoins, puisque $\hat{\theta}_{S,n} < \theta$ presque sûrement, on a $\mathbb{E}[\hat{\theta}_{S,n}] < \theta$ et cet estimateur est biaisé.

6. Pour $u \geq 0$, on a $\lim_{n\to\infty}(1-u/n)^n = e^{-u}$. Pour tout $t \geq 0$, posons $x = t \times n^{-1/3}$, alors pour n assez grand x est entre 0 et 1, et on peut appliquer l'inégalité ci-dessus, ce qui donne

$$\mathbb{P}(\hat{\theta}_{S,n} \leq \theta - tn^{-1/3}) = \left(1 - \tfrac{t^3}{2n}\right)^n \xrightarrow[n\to\infty]{} e^{-\frac{t^3}{2}}.$$

7. La suite de variables aléatoires $n^{1/3}(\theta - \hat{\theta}_{S,n})$ est positive donc si elle converge en loi, c'est encore vrai pour la loi limite. Pour tout $t \geq 0$, on vient de voir que

$$\mathbb{P}\left(n^{1/3}(\theta - \hat{\theta}_{S,n}) \geq t\right) \xrightarrow[n\to\infty]{} e^{-\frac{t^3}{2}}.$$

En notant W une variable aléatoire positive de fonction de répartition

$$\forall t \geq 0 \qquad F_W(t) = \mathbb{P}(W \leq t) = 1 - e^{-\frac{t^3}{2}},$$

on vient donc de prouver que

$$n^{1/3}(\theta - \hat{\theta}_{S,n}) \xrightarrow[n\to\infty]{\mathcal{L}} W.$$

La densité de W est

$$f_W(t) = \tfrac{3}{2}t^2 e^{-\frac{t^3}{2}}\ \mathbb{1}_{t\geq 0}.$$

Pour la culture, c'est une loi de Weibull, apparaissant typiquement dans la théorie des valeurs extrêmes (avec les lois de Gumbel et de Fréchet).

8. Nous avons affaire à un modèle de translation : la fonction f est continue sur $\mathbb{R}$ et C^1 sauf en un point (l'origine). De plus, un calcul élémentaire donne

$$\int_{\mathbb{R}} \frac{f'(y)^2}{f(y)} \mathbb{1}_{f(y)>0}\ dy = 12 < \infty$$

donc le modèle est régulier d'information de Fisher $I_1(\theta) = 12$ pour le modèle à une observation.

9. La vraisemblance des observations s'écrit

$$L_n(\theta) = \left(\tfrac{3}{2}\right)^n \left(\prod_{i=1}^{n} (1 - |X_i - \theta|)^2\right) \mathbb{1}_{X_{(n)}-1\leq\theta\leq X_{(1)}+1},$$

laquelle s'annule en dehors de l'intervalle $[X_{(n)} - 1, X_{(1)} + 1]$. Celui-ci est d'intérieur non vide puisque $X_{(n)} - 1 < \theta < X_{(1)} + 1$ presque sûrement. La fonction L_n étant continue sur ce compact, elle y atteint son maximum en au moins un point, qui définit donc un estimateur du maximum de vraisemblance $\hat{\theta}_{V,n}$. Puisque $X_{(n)} - 1$ et $X_{(1)} + 1$ convergent tous deux vers θ en probabilité, on en déduit que $\hat{\theta}_{V,n}$ est un estimateur consistant de θ. Puisque le modèle est régulier d'information de Fisher $I_1(\theta) = 12$, on a d'après le théorème du cours

$$\sqrt{n}\left(\hat{\theta}_{V,n} - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1/12).$$

10. En régime asymptotique, l'estimateur $\hat{\theta}_{S,n}$ converge seulement à vitesse $n^{-1/3}$, tandis que les trois autres convergent à vitesse $n^{-1/2}$. Parmi ceux-ci, on choisit celui de plus petite variance asymptotique, c'est-à-dire (sans surprise) celui du maximum de vraisemblance. Bien entendu, on fait ici abstraction des possibles erreurs de modèles et autres données aberrantes.

11. Si l'on dispose seulement d'un petit nombre d'observations, a priori seul le résultat sur $\hat{\theta}_{S,n}$ s'applique. On a vu que pour tout $x \in [0, 1]$,

$$\mathbb{P}\left(\hat{\theta}_{S,n} \leq \theta \leq \hat{\theta}_{S,n} + x\right) = 1 - \left(1 - \tfrac{x^3}{2}\right)^n.$$

Par conséquent, si $\alpha \in [2^{-n}, 1]$, un intervalle de confiance non asymptotique de niveau $(1-\alpha)$ pour θ est

$$\left[\hat{\theta}_{S,n} \; ; \; \hat{\theta}_{S,n} + \left(2\left(1-\alpha^{1/n}\right)\right)^{1/3}\right].$$

Pour $x \in [1,2]$, un calcul similaire à celui de la première question montre que

$$\mathbb{P}\left(\hat{\theta}_{S,n} \leq \theta \leq \hat{\theta}_{S,n} + x\right) = 1 - \left(\tfrac{(2-x)^3}{2}\right)^n.$$

De fait, si $\alpha \in [0, 2^{-n}]$, un intervalle de confiance non asymptotique de niveau $(1-\alpha)$ pour θ s'écrit

$$\left[\hat{\theta}_{S,n} \; ; \; \hat{\theta}_{S,n} + 2 - \left(2\alpha^{1/n}\right)^{1/3}\right].$$

Ceci étant, l'estimateur $\hat{\theta}_{M,n}$ conduit également à des intervalles de confiance non asymptotiques grâce aux inégalités classiques. Ainsi l'inégalité de Tchebychev conduit à

$$\forall c > 0 \qquad \mathbb{P}\left(\left|\hat{\theta}_{M,n} - \theta\right| \geq c\right) \leq \frac{\mathrm{Var}(X_1)}{c^2 n} = \frac{1}{10c^2 n},$$

d'où, pour tout $\alpha \in [0,1]$, l'intervalle de confiance de niveau $(1-\alpha)$:

$$\left[\hat{\theta}_{M,n} - \frac{1}{\sqrt{10n\alpha}} \; ; \; \hat{\theta}_{M,n} + \frac{1}{\sqrt{10n\alpha}}\right].$$

Puisque les variables X_i sont bornées, l'inégalité de Hoeffding s'applique elle aussi :

$$\forall c > 0 \qquad \mathbb{P}\left(\left|\hat{\theta}_{M,n} - \theta\right| \geq c\right) \leq 2\exp\left(-\tfrac{1}{2}c^2 n\right),$$

d'où, pour tout $\alpha \in [0,1]$, l'intervalle de confiance de niveau $(1-\alpha)$:

$$\left[\hat{\theta}_{M,n} - \sqrt{-\tfrac{2}{n}\log(\alpha/2)} \; ; \; \hat{\theta}_{M,n} + \sqrt{-\tfrac{2}{n}\log(\alpha/2)}\right].$$

Parmi ces trois intervalles de confiance non asymptotiques, on privilégiera celui de longueur minimale, ce qui dépendra à la fois de n et de α.

Exercice 2

On considère le modèle linéaire

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \quad i = 1, \ldots, n$$

où les vecteurs X_j, $1 \leq j \leq 3$, sont déterministes, connus et appartiennent à $\mathbb{R}^n$, et ε est un vecteur gaussien inconnu de matrice de covariance $\sigma^2 I_n$. De plus, on suppose

$$\|X_j\| = 1, \;\; 1 \leq j \leq 3, \;\; \langle X_j, X_{j'}\rangle = 1/2, \;\; 1 \leq j, j' \leq 3 \text{ avec } j \neq j'.$$

Les paramètres réels β_j, $1 \leq j \leq 3$, et σ^2 sont supposés inconnus.

1. La calcul donne : $AB = \frac{q+1}{2} I_q$.
2. Il suffit de remarquer que $X'X = A$ avec $q = 3$ pour en déduire que $(X'X)^{-1} = B/2$. Dès lors,
$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (X'X)^{-1}X'Y = \frac{1}{2} \times \begin{bmatrix} 3X_1'Y - X_2'Y - X_3'Y \\ -X_1'Y + 3X_2'Y - X_3'Y \\ -X_1'Y - X_2'Y + 3X_3'Y \end{bmatrix}.$$

D'où l'on déduit

$$\hat{\sigma}^2 = \frac{\left\|Y - X\hat{\beta}\right\|^2}{n-3}.$$

Les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants, avec

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(X'X)^{-1}\right) = \mathcal{N}\left(\beta, \frac{\sigma^2}{2} \times \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}\right),$$

et

$$(n-3)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-3}.$$

3. Puisque

$$\sqrt{\frac{2}{3}}\ \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sim \mathcal{T}_{n-3},$$

on en déduit, en notant t_α le quantile d'ordre $(1-\alpha/2)$ d'une loi de Student à $(n-3)$ ddl, que

$$\left[\hat{\beta}_1 - \sqrt{\frac{3}{2}}\ t_\alpha\hat{\sigma}\ ;\ \hat{\beta}_1 + \sqrt{\frac{3}{2}}\ t_\alpha\hat{\sigma}\right]$$

est un intervalle de confiance de niveau $(1-\alpha)$ pour β_1. L'espérance du carré de sa longueur vaut alors

$$\mathbb{E}\left[6t^2_\alpha\hat{\sigma}^2\right] = 6t^2_\alpha\sigma^2,$$

puisque $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.

4. On peut par exemple effectuer un test de Fisher entre modèles emboîtés. Notons $\hat{Y} = X\hat{\beta}$ et, suivant les notations du cours,

$$\hat{Y}_0 = P_0 Y = X_1(X_1'X_1)^{-1}X_1'Y = X_1X_1'Y,$$

alors sous H_0 : "$\beta_2 = \beta_3 = 0$", on sait que

$$F := \frac{\left\|\hat{Y} - \hat{Y}_0\right\|^2}{2\hat{\sigma}^2} \sim \mathcal{F}^2_{n-3}.$$

Soit f_α le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à 2 et $(n-3)$ ddl. Pour en déduire un test de niveau α, il suffit donc de comparer la réalisation $F(\omega)$ de la statistique de test à f_α : en l'occurrence, on accepte H_0 si $F(\omega) \leq f_\alpha$.

Exercice 3

On observe n variables aléatoires $X_1, \ldots, X_n$ iid de loi de Bernoulli de paramètre $\lambda \in [0,1]$ inconnu. On suppose que la loi a priori de λ est une loi Bêta de paramètres a et b.

1. La densité a posteriori de λ s'obtient comme suit :

$$f(\lambda|x_1, \ldots, x_n) \propto f(\lambda)\prod_{i=1}^{n} f(x_i|\lambda) \propto \lambda^{a+n\bar{x}_n-1}(1-\lambda)^{b+n-n\bar{x}_n-1}.$$

La loi a posteriori de λ est encore une loi Bêta, mais de paramètres $(a+n\bar{x}_n, b+n-n\bar{x}_n)$.

2. L'estimateur de Bayes $\tilde{\lambda}_n$ de λ pour la perte quadratique est la moyenne de cette loi a posteriori, donc

$$\tilde{\lambda}_n = \frac{a+n\bar{x}_n}{a+b+n}.$$

Exercice 4

On définit le carré de la distance de Hellinger h entre deux densités de probabilité f et g sur $\mathbb{R}$ par

$$h^2(f,g) = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx.$$

1. Puisque f et g sont deux densités, elles intègrent à 1, d'où

$$h^2(f,g) = \frac{1}{2}\int_{\mathbb{R}} \left(f(x) - 2\sqrt{f(x)g(x)} + g(x)\right) dx = 1 - \int_{\mathbb{R}} \sqrt{f(x)g(x)}\ dx.$$

   Noter que $\sqrt{2} \times h(f,g) = \|\sqrt{f} - \sqrt{g}\|_2$, distance L_2 entre les racines carrées de f et g. Ceci assure que h est positive, symétrique, vérifie l'inégalité triangulaire et que $h(f,g) = 0$ si et seulement si f et g sont presque partout égales. Enfin, la formule que l'on vient d'établir montre que $h(f,g) \leq 1$.

2. Partant de la formule de la question précédente et supposant $\theta \leq \theta'$, on obtient

$$h^2(f_\theta, f_{\theta'}) = 1 - \int_0^\theta \frac{1}{\sqrt{\theta\theta'}}\ dx = 1 - \sqrt{\frac{\theta}{\theta'}}.$$

3. Si $X_1, \ldots, X_n$ sont iid de loi uniforme sur $[0,\theta]$ pour un certain $\theta > 0$, alors la vraisemblance s'écrit

$$L_n(\theta) = \frac{1}{\theta^n}\mathbb{1}_{\theta \geq X_{(n)}},$$

   donc l'estimateur du maximum de vraisemblance est $\hat{\theta} = X_{(n)}$. Sa fonction de répartition F est nulle au-dessous de 0, égale à 1 au-dessus de θ, et vaut pour tout $t \in [0,\theta]$

$$F(t) = \mathbb{P}(X_{(n)} \leq t) = \mathbb{P}(X_1 \leq t)^n = \left(\frac{t}{\theta}\right)^n,$$

   d'où la densité

$$f(t) = \frac{nt^{n-1}}{\theta^n}\mathbb{1}_{[0,\theta]}(t).$$

   Par conséquent, le théorème de transfert donne

$$\mathbb{E}_\theta\left[\sqrt{\hat{\theta}}\right] = \int_0^\theta \sqrt{t}\ f(t)dt = \frac{n}{\theta^n}\int_0^\theta t^{n-\frac{1}{2}}\ dt = \frac{n\sqrt{\theta}}{n+\frac{1}{2}}.$$

4. On a

$$\mathbb{E}_\theta\left[h^2(f_{\hat{\theta}}, f_\theta)\right] = \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[h^2(f_{\hat{\theta}}, f_\theta)\middle|\,\hat{\theta}\right]\right]$$

   Or, d'après la question 2 et puisque $\hat{\theta} = X_{(n)} \leq \theta$ presque sûrement,

$$\mathbb{E}_\theta\left[h^2(f_{\hat{\theta}}, f_\theta)\middle|\,\hat{\theta}\right] = 1 - \sqrt{\frac{\hat{\theta}}{\theta}}.$$

   D'après la question précédente, il vient alors

$$\mathbb{E}_\theta\left[h^2(f_{\hat{\theta}}, f_\theta)\right] = \mathbb{E}_\theta\left[1 - \sqrt{\frac{\hat{\theta}}{\theta}}\right] = 1 - \frac{\mathbb{E}_\theta\left[\sqrt{\hat{\theta}}\right]}{\sqrt{\theta}} = \frac{1}{2n+1}.$$

   Un calcul facile (voir cours) montre que le risque quadratique vaut quant à lui

$$\mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right] = \frac{2\theta^2}{(n+1)(n+2)}.$$

   Ainsi le risque Hellinger est-il en $\mathcal{O}(1/n)$ tandis que le risque quadratique est en $\mathcal{O}(1/n^2)$.

Bibliographie

[1] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics*. Prentice Hall, 1976.

[2] Lucien Birgé. *Statistique mathématique*. Polycopié UPMC, 2014.

[3] Alexandr Alekseevich Borovkov. *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons Inc., 1991.

[5] Bernard Delyon. *Estimation paramétrique*. Format électronique, 2015.

[6] Benoît Cadre et Céline Vial. *Statistique mathématique - Master 1 et Agrégation*. Ellipses, 2012.

[7] Bernard Bercu et Djalil Chafaï. *Modélisation stochastique et simulation*. Dunod, 2007.

[8] Pierre-André Cornillon et Eric Matzner-Lober. *Régression avec R*. Springer, 2010.

[9] Vincent Rivoirard et Gilles Stoltz. *Statistique mathématique en action*. Vuibert, 2012.

[10] Jean Jacod et Philip Protter. *L'essentiel en théorie des probabilités*. Cassini, 2003.

[11] Dominique Fourdrinier. *Statistique inférentielle*. Dunod, 2002.

[12] Michel Lejeune. *Statistique - La théorie et ses applications*. Springer, 2005.

[13] Christian Robert. *Le choix bayésien*. Springer, 2010.

[14] Mark J. Schervish. *Theory of Statistics*. Springer-Verlag, 1995.

[15] Jun Shao. *Mathematical Statistics - Exercises and Solutions*. Springer, 2005.

[16] Larry Wasserman. *All of Statistics - A Concise Course in Statistical Inference*. Springer, 2004.

[17] Jan Wretman. A Simple Derivation of the Asymptotic Distribution of a Sample Quantile. *Scand. J. Statist.*, 5(2) :123–124, 1978.