Sorbonne Université
Master Mathématiques et Applications
Master 1, MU4MA015

Année 2024/2025
Premier Semestre

Statistique

Arnaud Guyader

Table des matières

Chapitre 1

Modélisation statistique

Introduction

Considérons un exemple jouet qui servira de fil rouge dans tout ce chapitre. 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 $x_i = 0$ si le i-ème lancer a donné Face et $x_i = 1$ pour Pile. 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 : i.i.d.) selon la loi de Bernoulli de paramètre θ, ce que l'on notera

$$(X_1, \dots, X_n) \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\theta),$$

où la probabilité $\theta \in]0, 1[$ de tomber sur Pile est inconnue. Au vu de la réalisation $(x_1, \dots, x_n)$ de cet échantillon $(X_1, \dots, X_n)$, on souhaite par exemple estimer le paramètre θ, ou encore tester si la pièce est équilibrée ou non, autrement dit si $\theta = 1/2$ ou si $\theta \neq 1/2$. 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 statistique. En probabilités, le paramètre θ est supposé connu, donc la loi de $X = X_1$ aussi, et on répond à des questions du type : 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 on cherche à en déduire des résultats impliquant cette loi de X. En statistique, 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 ici le paramètre θ.

Il n'en reste pas moins que les outils utilisés dans les deux domaines sont rigoureusement 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 réelle, 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. Dans ce cas, on dit parfois que la loi de X présente un atome en x_0.

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 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 à droite).

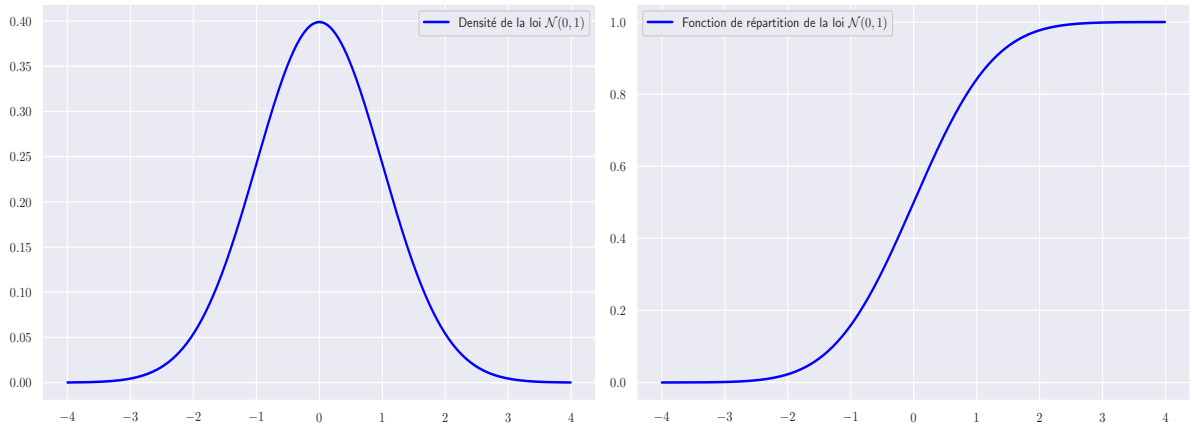

FIGURE 1.1 – Densité et fonction de répartition de la loi normale centrée réduite.

1.1.1 Modes de convergence

Nous nous focaliserons dans la suite sur les modes de convergence suivants : la convergence en probabilité, la convergence presque sûre, la convergence en loi et la convergence en moyenne quadratique (ou L_2).

Définition 1 (Convergence en probabilité, convergence presque sûre)
Les variables aléatoires (X_n) et X étant définies sur le même espace probabilisé, on dit que la suite (X_n) converge en probabilité vers la variable aléatoire X et on note

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

si

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

On dit que la suite (X_n) converge presque sûrement vers la variable aléatoire X et on note

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

si

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) := \mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

La convergence en probabilité dit que, si n est grand, X_n est proche de X avec grande probabilité. 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 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 déduit directement de la loi forte des grands nombres (cf. Section 1.1.3, Théorème 2) ou se démontre par l'intermédiaire du Lemme de Borel-Cantelli. Celui-ci assure en effet que 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. En effet, en notant $A_n^\varepsilon := \{|X_n - X| \geq \varepsilon\}$, on voit que $\{\lim_{n\to\infty} X_n \neq X\} = \cup_{\varepsilon>0} \limsup A_n^\varepsilon$, or la convergence de la série assure que $\mathbb{P}(\limsup A_n^\varepsilon) = 0$ et on conclut par sous-sigma-additivité.

Passons maintenant à la convergence en loi, d'usage constant en statistique en raison du Théorème Central Limite. Nous ne donnons ici qu'une des nombreuses caractérisations de ce mode de convergence (voir par exemple le Théorème porte-manteau).

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

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

si pour toute fonction continue et bornée φ, on a

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

Dans cette définition, on peut remplacer « pour toute fonction continue et bornée » par « pour toute fonction $\mathcal{C}^\infty$ à support compact ».

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 ! Du reste, la définition ne suppose même pas que les variables sont définies sur le même espace probabilisé.

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

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

1. Noter que $\mathbb{P}(|X_n - X| \geq \varepsilon) = \mathbb{E}[\mathbb{1}_{\{|X_n - X| \geq \varepsilon\}}]$ et appliquer la convergence dominée à $Y_n = \mathbb{1}_{\{|X_n - X| \geq \varepsilon\}}$.

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

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

Le critère de la définition ci-dessus n'est pas forcément 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

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

Exemple : Soit d'une part (X_n) suite de variables déterministes respectivement égales à $1/n$, et d'autre part $X = 0$. Toute fonction φ continue et bornée sur $\mathbb{R}$ est en particulier continue en 0 donc

$$\mathbb{E}[\varphi(X_n)] = \varphi(1/n) \xrightarrow[n\to\infty]{} \varphi(0) = \mathbb{E}[\varphi(X)],$$

ce qui prouve que (X_n) converge en loi vers X. En revanche, il est tout aussi clair que $F_{X_n}(0) = 0$ pour tout n, qui ne tend pas vers $F_X(0) = 1$.

Remarque : Supposons les variables X_n et X absolument continues de densités respectives f_n et f, alors pour que X_n converge en loi vers X, il suffit que f_n converge presque partout vers f. Cette condition n'est cependant pas nécessaire : il suffit pour s'en convaincre de considérer la suite de variables X_n de densités $f_n(x) = (1 - \cos(2\pi n x))\mathbb{1}_{[0,1]}(x)$ pour $n \geq 1$, qui tend en loi vers une uniforme sur $[0,1]$ (cf. fonctions de répartition) bien que $f_n(x)$ ne converge pour aucun x de $]0,1[$.

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). Noter que ceci ne correspond pas à la notation anglo-saxonne, pour laquelle une parenthèse est un crochet ouvert.

Exemple : Supposons que (X_n) converge en loi vers X, avec a et b des points de continuité de F_X, alors on peut montrer, par exemple grâce au Théorème porte-manteau, 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.

Si la fonction de répartition de la loi limite est continue sur $\mathbb{R}$, la convergence en loi équivaut donc à la convergence simple de la suite des fonctions de répartition. Le résultat suivant, donné à titre culturel, montre qu'on a en fait convergence uniforme.

Proposition 2 (Convergence uniforme d'une suite de fonctions de répartition)
Si la fonction de répartition F_X est continue sur tout $\mathbb{R}$ et si (X_n) converge en loi vers X, alors la suite de fonctions (F_{X_n}) converge uniformément vers F_X, c'est-à-dire

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

A nouveau, ce résultat s'applique en particulier lorsqu'il y a convergence vers une loi normale. Il correspond en fait au deuxième théorème de Dini appliqué à notre cadre.

Notons enfin 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 indépendantes. 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

$$\begin{array}{rccl} \Phi_X : & \mathbb{R} & \to \mathbb{C} & \\ & t & \mapsto \Phi_X(t) = \mathbb{E}\left[e^{itX}\right] = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)] \end{array}$$

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 loi 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[2] :

$$X \perp Y \implies \Phi_{X+Y} = \Phi_X \times \Phi_Y.$$

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

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

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.

Définition 3 (Convergence en moyenne quadratique)
On dit que la suite de variables aléatoires (X_n) de carrés intégrables tend vers X en moyenne quadratique, ou dans L_2, si

$$\mathbb{E}\left[(X_n - X)^2\right] \xrightarrow[n\to\infty]{} 0.$$

L'inégalité de Markov ci-dessous assure que la convergence en moyenne quadratique implique la convergence en probabilité.

2. Plus précisément, si on note $\Phi_{(X,Y)}(s,t) = \mathbb{E}[e^{i(sX+tY)}]$ la fonction caractéristique du couple (X, Y), alors on a l'équivalence : X et Y indépendantes si et seulement si $\forall (s,t) \in \mathbb{R}^2$, $\Phi_{(X,Y)}(s,t) = \Phi_X(s)\Phi_Y(t)$.

1.1.2 Majorations classiques

Si un résultat de convergence en loi se démontre souvent grâce à l'un des critères vus ci-dessus (fonctions tests, de répartition ou caractéristiques), une convergence en probabilité ou presque sûre découle typiquement de l'une des inégalités que nous rappelons maintenant. Elles quantifient 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, souvent plus faciles d'accès.

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

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

Ce résultat vient tout simplement de la décomposition

$$|X|^p = |X|^p \mathbb{1}_{\{|X|<c\}} + |X|^p \mathbb{1}_{\{|X|\geq c\}} \geq c^p \mathbb{1}_{\{|X|\geq c\}}.$$

Notons que si $\mathbb{E}\left[|X|^p\right] = +\infty$, cette inégalité reste valide, mais elle ne nous apprend rien. En prenant $p = 2$ 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 :

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

Noter qu'on a en fait démontré mieux, à savoir la convergence en moyenne quadratique de $(\hat{\theta}_n)$ vers θ puisque

$$\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] = \mathrm{Var}(\hat{\theta}_n) = \frac{\mathrm{Var}(S_n)}{n^2} = \frac{\theta(1-\theta)}{n} \xrightarrow[n\to\infty]{} 0.$$

La borne de Bienaymé-Tchebychev n'est cependant pas suffisamment précise pour montrer la convergence presque sûre via Borel-Cantelli puisque la série majorante $\sum 1/n$ en (1.1) 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 concentration : il en existe de multiples variantes, dont voici l'une des plus classiques.

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 $c \geq 0$

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

et

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

d'où il vient

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

Preuve : Si Y est une variable aléatoire à valeurs dans $[0,1]$, alors pour tout réel λ la convexité de la fonction $x \mapsto e^{\lambda x}$ implique

$$e^{\lambda Y} = e^{\lambda(Y\times 1 + (1-Y)\times 0)} \leq Y e^{\lambda} + (1 - Y),$$

d'où, en notant $p = \mathbb{E}[Y] \in [0,1]$,

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq p e^{\lambda} + 1 - p,$$

puis, en passant au logarithme,

$$\log \mathbb{E}\left[e^{\lambda(Y-p)}\right] \leq \log\Big(p e^{\lambda} + 1 - p\Big) - p\lambda =: \varphi(\lambda).$$

La formule de Taylor-Lagrange implique qu'il existe ℓ entre 0 et λ tel que

$$\varphi(\lambda) = \varphi(0) + \varphi'(0)\lambda + \frac{1}{2}\varphi''(\ell)\lambda^2.$$

On vérifie sans peine que $\varphi(0) = \varphi'(0) = 0$, tandis que

$$\varphi''(\ell) = \frac{p(1-p)e^{\lambda}}{\left(pe^{\lambda} + 1 - p\right)^2} = \frac{1}{4}\frac{\left(pe^{\lambda} + (1-p)\right)^2 - \left(pe^{\lambda} - (1-p)\right)^2}{\left(pe^{\lambda} + 1 - p\right)^2} \leq \frac{1}{4},$$

ce qui permet d'affirmer que

$$\log \mathbb{E}\left[e^{\lambda(Y - \mathbb{E}[Y])}\right] \leq \frac{\lambda^2}{8}.$$

Si maintenant $a \leq X \leq b$, alors $0 \leq (X-a)/(b-a) \leq 1$ et l'inégalité précédente en remplaçant λ par $\lambda(b-a)$ implique

$$\log \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq \frac{\lambda^2 (b-a)^2}{8}.$$

L'indépendance des X_i donne directement

$$\log \mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right] \leq \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2.$$

Notons $v = \sum_{i=1}^n (b_i - a_i)^2$, alors pour tous $\lambda > 0$ et $c \geq 0$, la stricte croissance de $x \mapsto e^{\lambda x}$ et l'inégalité de Markov assurent que

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq c\right) = \mathbb{P}\left(e^{\lambda(S_n - \mathbb{E}[S_n])} \geq e^{\lambda c}\right) \leq e^{-\lambda c}\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right] \leq e^{-\lambda c + \frac{v}{8}\lambda^2}.$$

Ceci étant vrai pour tout $\lambda > 0$, on a en particulier

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq c\right) \leq \inf_{\lambda>0} e^{-\lambda c + \frac{v}{8}\lambda^2} = e^{\inf_{\lambda>0}(\frac{v}{8}\lambda^2 - \lambda c)} = e^{-\frac{2c^2}{v}},$$

la dernière égalité étant obtenue par simple minimisation d'un trinôme sur $\mathbb{R}_+^*$. La première inégalité est donc établie. La deuxième s'obtient en changeant les X_i en $-X_i$ et la troisième découle naturellement des deux précédentes.

■

Remarque : 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 $m = \mathbb{E}[X_1]$, il appert que

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

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.

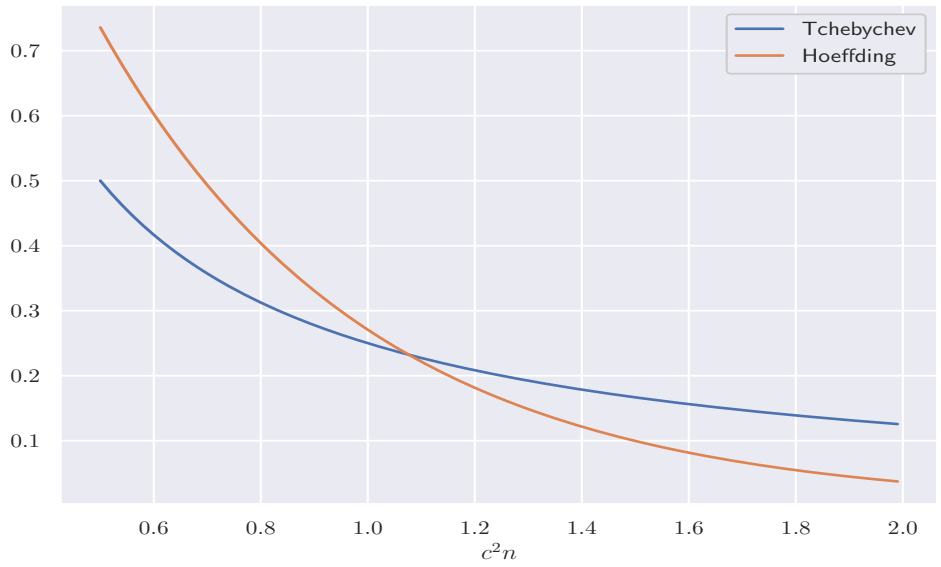

FIGURE 1.2 – Bornes de Bienaymé-Tchebychev et de Hoeffding pour $\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq c\right)$.

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 toutefois obtenir des bornes exponentielles explicites en supposant "seulement" que la variable X admet des moments exponentiels, c'est-à-dire que $\mathbb{E}[\exp(\lambda X)] < \infty$. L'idée est la même que dans la dernière étape de la démonstration précédente : pour tout réel c et tout $\lambda > 0$,

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

Si on sait minimiser φ sur $\mathbb{R}_+^*$ et si ce minimum est atteint en $\lambda_0 > 0$, ceci donne

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

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

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 de conclure directement : 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 i.i.d. 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.

Si l'on suppose que les X_i admettent un moment d'ordre 2, donc une variance $\sigma^2 < +\infty$, alors la loi faible des grands nombres est une simple conséquence de l'inégalité de Tchebychev puisque, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2 n} \xrightarrow[n\to\infty]{} 0.$$

Le résultat général du Théorème 2 montre que l'on n'a pas besoin de supposer l'existence d'un moment d'ordre 2 pour avoir la convergence, laquelle a même lieu presque sûrement.

Exemple : Dans notre exemple, les X_i étant effectivement i.i.d. 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 (à gauche) représente des trajectoires $(\hat{\theta}_n)$ pour une pièce déséquilibrée (2 fois plus de chances de tomber sur Face que sur Pile).

Remarque : Si les variables aléatoires X_n n'ont pas d'espérance, la suite S_n/n connaît des variations brusques et ne converge pas en général : ceci est illustré Figure 1.3 (à droite). On peut néanmoins montrer que si les X_n sont i.i.d. positives avec $\mathbb{E}[X_1] = +\infty$, alors S_n/n tend presque sûrement vers $+\infty$.

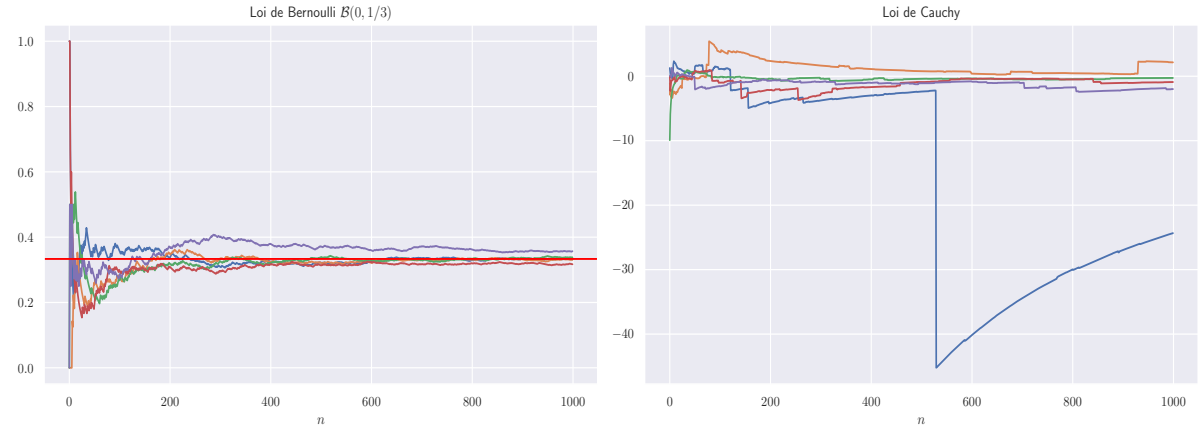

FIGURE 1.3 – Gauche : plusieurs réalisations de $(\hat{\theta}_n)_{1\leq n\leq 10^3}$ lorsque $\theta = 1/3$. Droite : plusieurs réalisations de $(S_n/n)_{1\leq n\leq 10^3}$ lorsque les X_i suivent une loi de Cauchy.

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 (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 i.i.d. admettant une variance $\sigma^2 = \mathrm{Var}(X_1) > 0$, alors

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

ce qui est équivalent à dire que

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

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, par exemple R et Python adoptent 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[3] 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 i.i.d., on a

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

Le TCL nous dit que, si l'on additionne un grand nombre de variables i.i.d., 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 en loi, c'est-à-dire

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

qui est exactement le Théorème Central Limite.

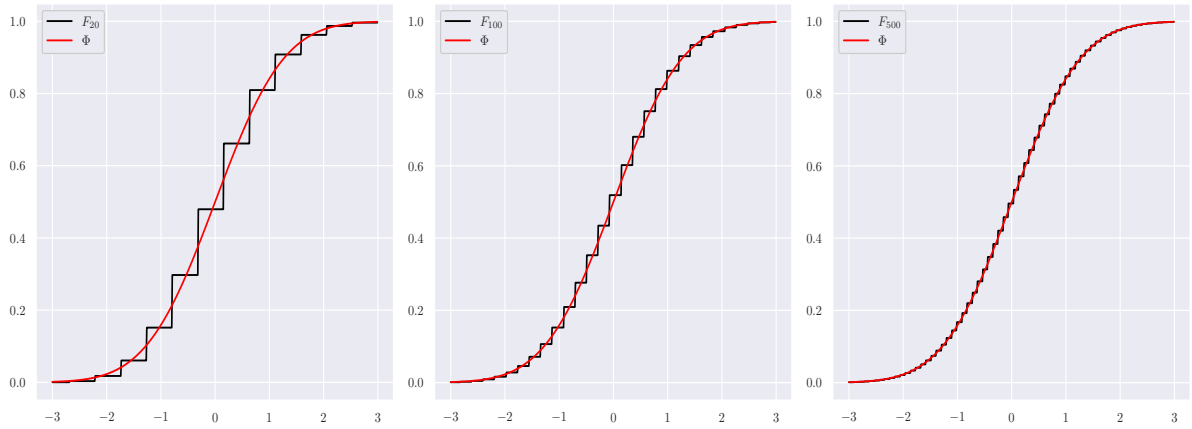

FIGURE 1.4 – Illustration du TCL via la convergence des fonctions de répartition F_n vers Φ pour le Pile ou Face avec $\theta = 1/3$ et respectivement $n = 20$, $n = 100$ et $n = 500$.

Exemple : Dans notre exemple, on a donc

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

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,

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

3. 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...

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

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

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

Puisque $\theta(1-\theta)\leq 1/4$, il vient

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

Comme le montre la Figure 1.5, cette borne est toujours meilleure que celle donnée par l'inégalité de Hoeffding vue précédemment, à savoir

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

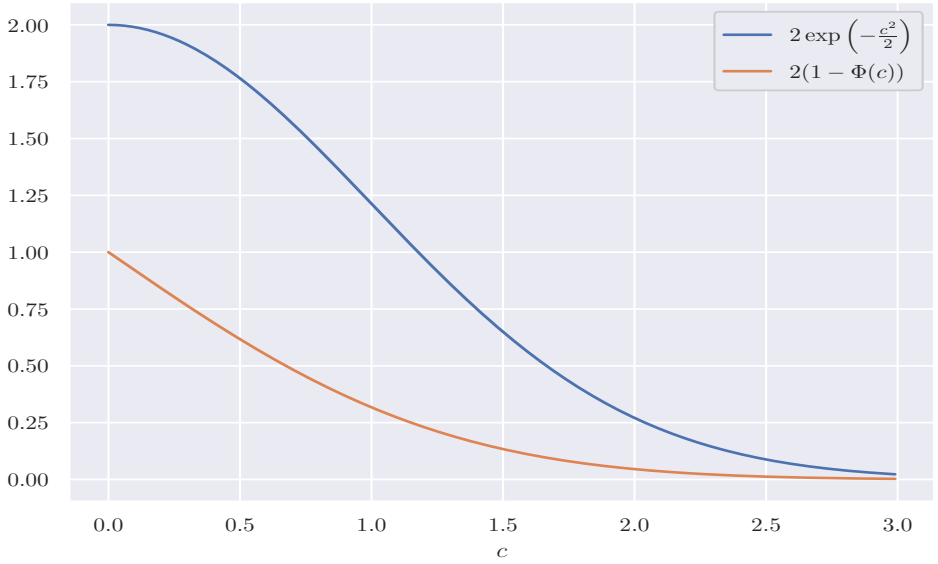

FIGURE 1.5 – Majorations de $\mathbb{P}\left(\left|\hat{\theta}_n-\theta\right|\geq\frac{c}{2\sqrt{n}}\right)$ par $2\left(1-\Phi(c)\right)$ et $2\exp\left(-\frac{c^2}{2}\right)$.

1.1.4 Opérations sur les limites

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[4] $(g(X_n))$. 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*.

4. Toutes les fonctions considérées dans ce cours sont boréliennes, donc $g(X_n)$ est bien une variable aléatoire.

Théorème 4 (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. C'est l'équivalent aléatoire du résultat bien connu sur les suites déterministes, à savoir que si (x_n) converge vers $L \in \mathbb{R}$ et si g est continue en L alors $(g(x_n))$ converge vers $g(L)$: g n'a nullement besoin d'être continue partout. Ici la limite n'est plus déterministe, mais aléatoire, donc il faut juste s'assurer du fait que g se comporte bien là où vit cette limite.

Exemple : Dans le jeu du Pile ou Face, puisque $\theta \in]0,1[$, la fonction $g : x \mapsto 1/\sqrt{x(1-x)}$ est continue en θ. Puisque $(\hat{\theta}_n)$ converge presque sûrement vers θ, on en déduit que $(1/\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)})$ converge presque sûrement vers $1/\sqrt{\theta(1-\theta)}$. La multiplication par une constante étant aussi une application continue, il s'ensuit que

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

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 $\Rightarrow$ 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 :

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

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

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

On retient : Les convergences *p.s.* et L_2 impliquent toutes deux la convergence en probabilité, laquelle implique la convergence en loi.

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.

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

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

Preuve : Nous allons montrer un résultat plus fort, en l'occurrence que le couple (X_n, Y_n) tend en loi vers (X, a). La définition de la convergence en loi pour un couple aléatoire est la même que celle de la Définition 2 pour une suite de variables aléatoires : on dit que (X_n, Y_n) converge en loi vers (X, Y) si pour toute fonction continue et bornée $\varphi : \mathbb{R}^2 \to \mathbb{R}$, on a

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

De plus, modulo l'introduction de la fonction caractéristique d'un couple aléatoire (X, Y) via

$$\forall (s,t) \in \mathbb{R}^2 \qquad \Phi_{(X,Y)}(s,t) = \mathbb{E}\left[e^{i(sX+tY)}\right],$$

le Théorème de Paul Lévy est toujours valide : (X_n, Y_n) converge en loi vers (X, Y) si et seulement si

$$\forall (s,t) \in \mathbb{R}^2 \qquad \Phi_{(X_n,Y_n)}(s,t) \xrightarrow[n\to\infty]{} \Phi_{(X,Y)}(s,t).$$

Ici, nous devons donc prouver que, pour tout couple de réels (s, t), on a bien

$$\mathbb{E}\left[e^{i(sX_n+tY_n)}\right] \xrightarrow[n\to\infty]{} \mathbb{E}\left[e^{i(sX+ta)}\right].$$

On part de la décomposition

$$e^{i(sX_n+tY_n)} - e^{i(sX+ta)} = e^{isX_n}\left(e^{itY_n} - e^{ita}\right) + e^{ita}\left(e^{isX_n} - e^{isX}\right).$$

d'où

$$\left|\Phi_{(X_n,Y_n)}(s,t) - \Phi_{(X,a)}(s,t)\right| = \left|\mathbb{E}\left[e^{isX_n}\left(e^{itY_n} - e^{ita}\right)\right] + e^{ita}\mathbb{E}\left[e^{isX_n} - e^{isX}\right]\right|$$

et on applique l'inégalité triangulaire pour en déduire

$$\left|\Phi_{(X_n,Y_n)}(s,t) - \Phi_{(X,a)}(s,t)\right| \leq \mathbb{E}\left[\left|e^{itY_n} - e^{ita}\right|\right] + \left|\mathbb{E}\left[e^{isX_n} - e^{isX}\right]\right|.$$

Par le Théorème de Paul Lévy, le second terme tend vers 0. Par ailleurs, (Y_n) tend en probabilité donc en loi vers a et la fonction $y \mapsto \left|e^{ity} - e^{ita}\right|$ est continue bornée donc, par définition de la convergence en loi, le premier terme tend également vers 0. On a donc prouvé que (X_n, Y_n) tend en loi vers (X, a). Il est alors facile de voir que $(X_n + Y_n)$ tend en loi vers $X + a$. Soit en effet $\psi : \mathbb{R} \to \mathbb{R}$ continue bornée, alors la fonction $\varphi : \mathbb{R}^2 \to \mathbb{R}$ définie par $\varphi(x, y) = \psi(x + y)$ est continue bornée donc

$$\mathbb{E}[\psi(X_n + Y_n)] = \mathbb{E}[\varphi(X_n, Y_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[\varphi(X, a)] = \mathbb{E}[\psi(X + a)],$$

ce qui est le résultat voulu. Le raisonnement est le même pour le produit $(X_n Y_n)$ et plus généralement pour toute suite de variables aléatoire de la forme $(\varphi(X_n, Y_n))$ où $\varphi : \mathbb{R}^2 \to \mathbb{R}$ est continue.

■

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

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

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 ? Nous avons vu ci-dessus que

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

ce qui implique bien sûr que

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

Il suffit alors d'appliquer le Théorème de Slutsky :

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

Ceci permet de construire des intervalles de confiance asymptotiques, comme nous le verrons plus loin.

Attention ! La convergence en loi n'est pas stable par addition. Soit $X \sim \mathcal{N}(0,1)$, $X_n = Y_n = X$ pour tout n, et $Y = -X$, alors (X_n) converge en loi vers X, (Y_n) converge en loi vers Y, mais $(X_n + Y_n)$ ne converge pas en loi vers $X + Y = 0$.

Le résultat suivant n'a rien d'étonnant et montre grosso modo qu'un TCL implique une convergence en probabilité. Pour le prouver, il suffit de prendre $Y_n = 1/\sqrt{n}$ dans le Théorème de Slutsky.

Corollaire 2
Soit (X_n) une suite de variables aléatoires, X une variable aléatoire et a un nombre réel tels que

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

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 méthode Delta ci-dessous.

Entre autres choses, cette méthode Delta 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 4 qui est un résultat de continuité, celui-ci peut se voir comme un résultat de dérivabilité.

L'idée est la suivante : supposons par exemple que $\sqrt{n}\,(X_n - 1)$ tende en loi vers une gaussienne centrée réduite et considérons par ailleurs une fonction g dérivable en 1, alors sans souci de rigueur on écrirait

$$X_n \approx 1 + \frac{1}{\sqrt{n}}\mathcal{N}(0,1) \qquad \text{et} \qquad g(1+h) \approx g(1) + g'(1)h$$

d'où

$$g(X_n) \approx g(1 + \frac{1}{\sqrt{n}}\mathcal{N}(0,1)) \approx g(1) + g'(1) \times \frac{1}{\sqrt{n}}\mathcal{N}(0,1),$$

c'est-à-dire

$$\sqrt{n}\,(g(X_n) - g(1)) \approx g'(1)\mathcal{N}(0,1) = \mathcal{N}(0,(g'(1))^2).$$

La méthode Delta traduit cette heuristique de façon rigoureuse.

Théorème 6 (méthode Delta)
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

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

Si g est une fonction dérivable au point a, alors

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

En particulier, si $v_n = \sqrt{n}$ et $X \sim \mathcal{N}(0,\sigma^2)$ alors

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

Preuve : D'après le Corollaire 2 et la remarque qui la suit, on sait que

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

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 :

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

■

Exemple : Nous avons vu que

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

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

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

Remarque : Si $g'(a) = 0$, alors $g'(a)X = 0$ et la loi limite est un Dirac en 0, ce qui nous apprend seulement que $g(X_n)$ tend vers $g(a)$ à vitesse plus rapide que $1/\sqrt{n}$. Pour connaître la vitesse

effective, il suffit souvent, comme en analyse, de pousser le développement limité jusqu'au premier terme non nul. Reprenons l'exemple où $\sqrt{n}\,(X_n - 1)$ tend en loi vers une gaussienne centrée réduite avec cette fois $g'(1) = 0$ mais $g''(1) \neq 0$. Alors, toujours sans souci de rigueur, on écrit

$$X_n \approx 1 + \frac{1}{\sqrt{n}}\mathcal{N}(0,1) \qquad \text{et} \qquad g(1+h) \approx g(1) + \frac{1}{2}g''(1)h^2$$

d'où

$$g(X_n) \approx g(1 + \frac{1}{\sqrt{n}}\mathcal{N}(0,1)) \approx g(1) + \frac{1}{2}g''(1)\left(\frac{1}{\sqrt{n}}\mathcal{N}(0,1)\right)^2,$$

c'est-à-dire, puisque le carré d'une loi $\mathcal{N}(0,1)$ est une loi du khi-deux à un degré de liberté, notée χ^2_1,

$$n\,(g(X_n) - g(1)) \approx \frac{g''(1)}{2}\chi^2_1.$$

En adaptant la preuve de la méthode Delta, on peut montrer rigoureusement que

$$\frac{2}{g''(1)}n\,(g(X_n) - g(1)) \xrightarrow[n\to\infty]{d} \chi^2_1.$$

Il y a donc convergence à vitesse $1/n$ et la loi limite n'est plus gaussienne.

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 espace mesuré $(E, \mathcal{E}, \mu)$, i.e. 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 [5] δ_k :

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

   Dans ce cadre, en munissant comme d'habitude $\mathbb{R}$ de la tribu borélienne $\mathcal{B}(\mathbb{R})$, toute fonction $\varphi : \mathbb{N} \to \mathbb{R}$ est $(\mathcal{P}(\mathbb{N}), \mathcal{B}(\mathbb{R}))$-mesurable et correspond à une suite $(\varphi(n))_{n\geq 0}$. Si la série $\sum_{n\geq 0} \varphi(n)$ est absolument convergente, la suite $(\varphi(n))_{n\geq 0}$ est dite intégrable par rapport à μ, d'intégrale la somme de la série :

$$\int_E \varphi(x)\mu(dx) = \sum_{n=0}^{\infty} \varphi(n)\mu(\{n\}) = \sum_{n=0}^{\infty} \varphi(n).$$

5. On rappelle que $\delta_k(A) = 1$ si $k \in A$ et $\delta_k(A) = 0$ sinon.

2. Mesure de Lebesgue : $(E, \mathcal{E}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, où λ est la mesure de Lebesgue qui à un intervalle associe sa longueur, éventuellement infinie. Avec la notation (a, b) définie précédemment, ceci s'écrit :

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

avec la convention classique : $+\infty - a = +\infty - (-\infty) = b - (-\infty) = +\infty$. Une fonction $\varphi : \mathbb{R} \to \mathbb{R}$ supposée borélienne, i.e. $(\mathcal{B}(\mathbb{R}), \mathcal{B}(\mathbb{R}))$-mesurable, est dite intégrable si sa valeur absolue est intégrable au sens de Lebesgue, auquel cas on note

$$\int_E \varphi(x)\mu(dx) = \int_{\mathbb{R}} \varphi(x)\lambda(dx) = \int_{\mathbb{R}} \varphi(x)dx.$$

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

Définition 4 (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.
3. L'ensemble des réels n'étant pas dénombrable, la mesure de comptage sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ n'est pas σ-finie.

L'absolue continuité correspond à une relation de préordre (réflexivité et transitivité) entre mesures.

Définition 5 (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 7 (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 $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, laquelle est complètement spécifiée par les probabilités des singletons $P_X(\{n\}) = \mathbb{P}(X = n)$. Puisque le seul ensemble négligeable pour la mesure de comptage est l'ensemble vide, il est clair que $P_X \ll \mu$. Le théorème de Radon-Nikodym affirme donc qu'il existe une fonction, ici une suite $f(n)$, telle que pour toute suite $\varphi(n)$ intégrable par rapport à P_X on puisse écrire
$$\sum_{n=0}^{\infty} \varphi(n)P_X(\{n\}) = \sum_{n=0}^{\infty} \varphi(n)f(n)\mu(\{n\}),$$
autrement dit
$$\sum_{n=0}^{\infty} \varphi(n)\mathbb{P}(X = n) = \sum_{n=0}^{\infty} \varphi(n)f(n).$$
En prenant comme fonctions tests $\varphi_k(n) = \mathbb{1}_k(n)$, on en déduit que la densité $f(n)$ au point n n'est rien d'autre que $\mathbb{P}(X = n)$. Sous réserve d'intégrabilité, on retrouve ainsi que l'espérance de la variable aléatoire $\varphi(X)$ s'écrit
$$\mathbb{E}[\varphi(X)] := \sum_{n=0}^{\infty} \varphi(n)P_X(\{n\}) = \sum_{n=0}^{\infty} \varphi(n)\mathbb{P}(X = n)\mu(\{n\}) = \sum_{n=0}^{\infty} \varphi(n)\mathbb{P}(X = n).$$
Par exemple, la loi de Bernoulli de paramètre θ est $P_\theta = (1-\theta)\delta_0 + \theta\delta_1$, qui est absolument continue par rapport à la mesure de comptage sur $\mathbb{N}$, et même par rapport à la mesure de comptage sur $\{0, 1\}$.
2. Mesure de Lebesgue : une variable aléatoire réelle 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 borélienne 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
$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x)P_X(dx) = \int_{\mathbb{R}} \varphi(x)f(x)dx.$$
Cette densité f correspond exactement à la dérivée de Radon-Nikodym de P_X par rapport à la mesure de Lebesgue, i.e. $f(x) = \frac{dP_X}{d\lambda}(x)$.

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 (physique, chimie, biologie, 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 6 (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.

Attention ! Dans toute la suite, afin d'alléger les écritures, nous adopterons généralement l'abus de notation consistant à utiliser la même lettre θ pour la vraie valeur du paramètre (i.e. θ^*) et pour une valeur générique de celui-ci (comme dans la notation $(P_\theta)_{\theta\in\Theta}$). Le contexte doit cependant permettre d'éviter toute confusion.

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 $\mathbf{X} = (X_1,\dots,X_n)$. 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 i.i.d., 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 $\mathbf{x} = (x_1,\dots,x_n)$ de $\mathbf{X}$ a, sous P_θ, la probabilité

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

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

2. Dans une population donnée, la taille des femmes adultes est modélisée par une loi normale de moyenne et variance inconnues. On veut estimer ces dernières à partir d'un échantillon de n femmes choisies 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 $\mathbf{X} = (X_1,\dots,X_n)$ avec les X_i

i.i.d. 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 i.i.d. appelées des observations[6]. Lorsque, comme dans ces exemples, ces variables sont i.i.d. 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 plus loin sur le modèle de régression linéaire. Par ailleurs, ces deux exemples ont un autre point commun : la taille de l'espace des paramètres.

Définition 7 (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 "en pratique" tout se passe comme si les variables gaussiennes étaient 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 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 (borne de l'union). 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.

Définition 8 (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.

6. Tandis que $\mathbf{x} = \mathbf{X}(\omega) = (x_1, \ldots, x_n)$ correspond à des réalisations de ces variables aléatoires.

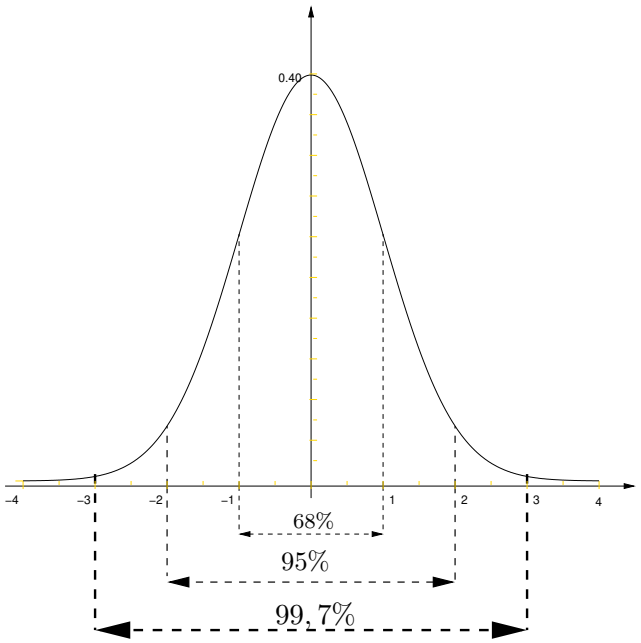

FIGURE 1.6 – Concentration de la loi normale standard autour de sa moyenne.

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 intervenir la notion d'absolue continuité rappelée en Section 1.1.5.

Définition 9 (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 spécifiée par la Définition 6 et on inventorie quelques questions classiques en statistique inférentielle. Comme précédemment, l'exemple du jeu de Pile ou Face servira de fil conducteur 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 10 (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.

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 11 (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 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 $\mathbb{P}$. Ainsi on pourra aussi écrire
$$R(\hat{\theta},\theta) = \mathbb{E}_\theta\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathbb{E}_\theta\left[\left(\hat{\theta}(\mathbf{X})-\theta\right)^2\right] = \int_E \left(\hat{\theta}(\mathbf{x})-\theta\right)^2 P_\theta(\mathbf{dx}).$$
Afin d'alléger les écritures, nous n'adopterons ces notations qu'en cas d'ambiguïté possible. Quoi qu'il en soit, il importe 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, pour tout $c>0$,
$$\mathbb{P}\left(\left|\hat{\theta}-\theta\right| \geq c\right) \leq \frac{\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right]}{c^2} = \frac{R(\hat{\theta},\theta)}{c^2}.$$
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 d'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
$$R(\hat{\theta},\theta) = \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]\right)^2\right] + \left(\mathbb{E}[\hat{\theta}]-\theta\right)^2 = \mathrm{Var}(\hat{\theta}) + B(\hat{\theta})^2.$$
■

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. Cette décomposition biais-variance se généralise en dimension supérieure lorsque $\Theta \subseteq \mathbb{R}^k$ est muni 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

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

Définition 12 (Convergence et normalité asymptotique)
Soit θ un paramètre réel inconnu. On dit que la suite d'estimateurs $(\hat{\theta}_n)_{n\geq 1}$ est :

— *consistante, ou convergente, si*

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

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

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

Remarques :

1. Par abus de langage, on dira simplement que $\hat{\theta}_n$ est un estimateur consistant et asymptotiquement normal de θ. D'autre part, on dira que $\hat{\theta}_n$ est un estimateur fortement consistant si la convergence vers θ a lieu presque sûrement.
2. De façon plus générale, s'il existe une suite (v_n) tendant vers l'infini et une variable X non dégénérée (i.e. non p.s. égale à 0) telles que $v_n(\hat{\theta}_n - \theta)$ tend en loi vers X, alors on dit que l'estimateur $\hat{\theta}_n$ converge à vitesse $1/v_n$.

Rappelons que, d'après le Corollaire 2, la normalité asymptotique de $(\hat{\theta}_n)_{n\geq 1}$ implique sa consistance (mais pas sa consistance forte). Par ailleurs, si l'on dispose d'une suite $(\hat{\sigma}_n^2)_{n\geq 1}$ d'estimateurs consistante pour σ^2, alors le Théorème de Slutsky entraîne que

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

ce qui permet de construire des intervalles de confiance asymptotiques pour θ (cf. Section 1.3.2).

En estimation paramétrique, le cadre d'application typique de la méthode Delta 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 méthode Delta 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 partout dérivable sans que sa réciproque soit dérivable partout.

Proposition 6 (méthode Delta et fonction inversible)
Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires i.i.d. de loi P_θ, avec θ point intérieur à Θ intervalle de $\mathbb{R}$, et φ un C^1-difféomorphisme de Θ dans $\varphi(\Theta)$. Si $\hat{\varphi}_n = \hat{\varphi}_n(X_1, \ldots, X_n)$ est un estimateur consistant de $\varphi(\theta)$, 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

$$v_n(\hat{\varphi}_n - \varphi(\theta)) \xrightarrow[n\to\infty]{d} Z_\theta,$$

alors

$$v_n(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \frac{1}{\varphi'(\theta)} Z_\theta.$$

Dans le cas particulier où $v_n = \sqrt{n}$ et $Z_\theta \sim \mathcal{N}(0, \sigma_\theta^2)$, on a donc

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, (\sigma_\theta/\varphi'(\theta))^2).$$

La preuve consiste simplement à adapter le théorème de continuité et la méthode Delta dans un contexte un peu spécial.

Preuve : Le point θ étant intérieur à Θ et φ continue bijective, le point $\varphi(\theta)$ est intérieur à $\varphi(\Theta)$. En particulier, pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que la boule centrée en $\varphi(\theta)$ et de rayon δ soit contenue dans $\varphi(\Theta)$ et, par continuité de φ^{-1} en $\varphi(\theta)$,

$$|u - \varphi(\theta)| < \delta \implies |\varphi^{-1}(u) - \theta| < \varepsilon.$$

Il convient maintenant de définir $\hat{\theta}_n$ de façon générale. De deux choses l'une : ou bien $\hat{\varphi}_n \in \varphi(\Theta)$, auquel cas $\hat{\theta}_n = \varphi^{-1}(\hat{\varphi}_n)$; ou bien $\hat{\varphi}_n \notin \varphi(\Theta)$, auquel cas on peut considérer un point arbitraire θ_0 de Θ et poser $\hat{\theta}_n = \theta_0$. On a donc, avec la convention $\varphi^{-1}(\hat{\varphi}_n)\mathbb{1}_{\hat{\varphi}_n \notin \varphi(\Theta)} = 0$,

$$\hat{\theta}_n = \varphi^{-1}(\hat{\varphi}_n)\mathbb{1}_{\hat{\varphi}_n \in \varphi(\Theta)} + \theta_0 \mathbb{1}_{\hat{\varphi}_n \notin \varphi(\Theta)}.$$

Ainsi l'estimateur $\hat{\theta}_n$ est-il bien défini au sens de l'énoncé dès que $\hat{\varphi}_n \in \varphi(\Theta)$, or

$$\mathbb{P}(|\hat{\varphi}_n - \varphi(\theta)| < \delta) \leq \mathbb{P}(\hat{\varphi}_n \in \varphi(\Theta))$$

et le membre de gauche tend vers 1 lorsque n tend vers l'infini car $\hat{\varphi}_n$ tend en probabilité vers $\varphi(\theta)$, donc $\hat{\theta}_n$ est bien défini (au sens de l'énoncé) avec une probabilité qui tend vers 1. De plus, puisque

$$|\hat{\varphi}_n - \varphi(\theta)| < \delta \implies |\varphi^{-1}(\hat{\varphi}_n) - \theta| = |\hat{\theta}_n - \theta| < \varepsilon,$$

il en résulte que

$$\mathbb{P}(|\hat{\varphi}_n - \varphi(\theta)| < \delta) \leq \mathbb{P}(|\hat{\theta}_n - \theta| < \varepsilon).$$

Il reste à nouveau à faire tendre n vers l'infini pour en déduire que, pour tout $\varepsilon > 0$,

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

c'est-à-dire

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

Pour la convergence en loi, on peut partir de la décomposition

$$v_n(\hat{\theta}_n - \theta) = v_n(\varphi^{-1}(\hat{\varphi}_n) - \varphi^{-1}(\varphi(\theta)))\mathbb{1}_{\hat{\varphi}_n \in \varphi(\Theta)} + v_n(\theta_0 - \theta)\mathbb{1}_{\hat{\varphi}_n \notin \varphi(\Theta)}. \tag{1.3}$$

D'après ci-dessus, pour tout $\varepsilon > 0$,

$$\mathbb{P}(|v_n(\theta_0 - \theta)\mathbb{1}_{\hat{\varphi}_n \notin \varphi(\Theta)}| \geq \varepsilon) \leq \mathbb{P}(\hat{\varphi}_n \notin \varphi(\Theta)) \xrightarrow[n\to\infty]{} 0,$$

donc le dernier terme du membre de droite de (1.3) tend en probabilité vers 0. Pour le premier, le même raisonnement assure que

$$\mathbb{1}_{\hat{\varphi}_n \in \varphi(\Theta)} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

Par ailleurs, la dérivabilité de φ^{-1} en $\varphi(\theta)$ et la relation [7] $(\varphi^{-1})'(\varphi(\theta)) = 1/\varphi'(\theta)$ donne pour tout $u \in \varphi(\Theta)$

$$\varphi^{-1}(u) = \theta + (u - \varphi(\theta))(1/\varphi'(\theta) + r(u)),$$

où r est définie sur $\varphi(\Theta)$ et continue en $\varphi(\theta)$ avec $r(\varphi(\theta)) = 0$. Par conséquent

$$v_n(\varphi^{-1}(\hat{\varphi}_n) - \varphi^{-1}(\varphi(\theta)))\mathbb{1}_{\hat{\varphi}_n \in \varphi(\Theta)} = v_n(\hat{\varphi}_n - \varphi(\theta))(1/\varphi'(\theta) + r(\hat{\varphi}_n))\mathbb{1}_{\hat{\varphi}_n \in \varphi(\Theta)},$$

et le Lemme de Slutsky donne

$$v_n(\varphi^{-1}(\hat{\varphi}_n) - \varphi^{-1}(\varphi(\theta)))\mathbb{1}_{\hat{\varphi}_n \in \varphi(\Theta)} \xrightarrow[n\to\infty]{d} \frac{1}{\varphi'(\theta)} Z_\theta.$$

En revenant à (1.3), une nouvelle application du Lemme de Slutsky donne finalement bien le résultat annoncé, à savoir

$$v_n(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \frac{1}{\varphi'(\theta)} Z_\theta.$$

■

Remarque : Soit $\theta \in \Theta :=]0, +\infty[$ un paramètre inconnu que l'on cherche à estimer et $(X_i)_{i\geq 1}$ des variables i.i.d. selon une loi de Poisson de paramètre $1/\theta$. Pour estimer θ, il suffit de considérer le C^1-difféomorphisme $\varphi : \Theta \to \Theta$ défini par $\varphi(\theta) = 1/\theta$. Par la Loi des Grands Nombres et le TCL, la moyenne empirique $\hat{\varphi}_n := \bar{X}_n$ est un estimateur consistant et asymptotiquement normal de $\varphi(\theta) = 1/\theta$. Le résultat précédent assure alors que $\hat{\theta}_n = \varphi^{-1}(\hat{\varphi}_n) = 1/\bar{X}_n$ est un estimateur consistant et asymptotiquement normal de θ, avec

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \theta^3).$$

On peut noter qu'il n'est pas nécessairement défini pour tout n, mais seulement à partir du premier indice n_0 tel que $X_{n_0} > 0$. C'est en ce sens qu'il faut comprendre le passage "est défini avec une probabilité qui tend vers 1 lorsque $n \to \infty$" dans l'énoncé ci-dessus : la probabilité qu'il ne soit toujours pas défini lorsque l'on dispose de n données est égale à $e^{-n/\theta}$, quantité qui pour tout $\theta > 0$ tend bien vers zéro lorsque la taille de l'échantillon tend vers l'infini.

Nota Bene. La normalité asymptotique ne permet pas de contrôler le risque quadratique. Dans le modèle précédent des lois de Poisson $\mathcal{P}(1/\theta)$, $\theta > 0$, l'estimateur $\hat{\theta}_n = 1/\bar{X}_n$ est asymptotiquement normal, mais de risque quadratique infini puisque $\mathbb{P}(\bar{X}_n = 0) > 0$.

Outre l'estimation du paramètre inconnu θ, on peut chercher un intervalle dans lequel celui-ci a de grandes chances de se trouver : c'est ici qu'intervient la notion d'intervalles de confiance.

7. Noter que $\varphi'(\theta) \neq 0$ car φ est un C^1-difféomorphisme de Θ dans $\varphi(\Theta)$.

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 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 [8],

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

Cette fois, le doute n'est plus permis : il est à peu près certain que la pièce est déséquilibrée.

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 13 (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 $\theta \in \Theta$,

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

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. Pour l'exemple du Pile ou Face, $[0, 1]$ est un intervalle de confiance à 95% (et même à 100%), mais il est clair qu'il n'a aucun intérêt...
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
$$\forall \theta \in \Theta \qquad \mathbb{P}_\theta(\theta \in D(\mathbf{X})) \geq 1 - \alpha.$$

La méthode standard pour obtenir des intervalles de confiance est de passer par des inégalités classiques comme celles vues en Section 1.1.2 ou, pour des intervalles de confiance asymptotiques, par un résultat de convergence en loi tel que le Théorème Central Limite.

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 $c > 0$,

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

8. Qui est en fait excellente car $\theta = 1/2$.

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

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

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 rayon $2.24/\sqrt{n}$.

Par l'inégalité de Hoeffding, nous avons obtenu

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

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

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

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 $n \to \infty$.

Définition 14 (Intervalle de confiance asymptotique)
Supposons $\Theta \subseteq \mathbb{R}$, $\mathbf{X} = (X_1, \ldots, X_n)$ et $\alpha \in]0,1[$. On appelle intervalles de confiance pour θ de niveau asymptotique $(1-\alpha)$ toute suite d'intervalles aléatoires $(\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))$ dont les bornes sont des statistiques et telle que, pour tout $\theta \in \Theta$,

$$\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in (\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))) \geq 1-\alpha.$$

Dans tous nos exemples, la limite inférieure sera en fait une limite classique. Illustrons l'idée sur l'exemple du Pile ou Face.

Exemple : Le Théorème Central Limite a permis d'établir, pour tout $0 < \theta < 1$, la convergence en loi

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

Rappelons que $q_{1-\alpha/2}$ désigne 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

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

Le paramètre inconnu θ apparaissant dans les bornes de l'intervalle, ce n'est pas un intervalle de confiance ! 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 tout $0 < \theta < 1$ donc

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

et en particulier

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

Plus précisément, on peut noter que la limite inférieure est en fait une limite usuelle puisque

$$\mathbb{P}_\theta\left(\left|\hat{\theta}_n - \theta\right| \leq \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right) \xrightarrow[n\to\infty]{} \mathbb{P}\left(|\mathcal{N}(0,1)| \leq \frac{q_{1-\alpha/2}}{2\sqrt{\theta(1-\theta)}}\right).$$

Mais en général, on fait plutôt 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)

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

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[9]. 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.

Remarques :

1. 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...
2. En Définition 14, si on requiert plutôt

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in (\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))) \geq 1-\alpha,$$

on parle d'intervalles de confiance asymptotiques **forts**. Il est facile de voir que cette condition implique celle donnée dans la définition. Cependant, un exemple permet de voir qu'elle

9. 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.

est bien plus exigeante (de même qu'en analyse la convergence uniforme d'une suite de fonctions implique strictement sa convergence simple). Considérons en effet des variables X_i i.i.d. selon une loi de Poisson de paramètre $\theta > 0$. Par le TCL et le Lemme de Slutsky, un intervalle de confiance asymptotique au sens de la Définition 14 est

$$IC(\theta, n) = \left[\bar{X}_n - \frac{\Phi^{-1}(1-\alpha/2)\sqrt{\bar{X}_n}}{\sqrt{n}}; \bar{X}_n + \frac{\Phi^{-1}(1-\alpha/2)\sqrt{\bar{X}_n}}{\sqrt{n}} \right].$$

Mais clairement le paramètre inconnu $\theta > 0$ n'appartient pas à cet intervalle si la borne de droite est nulle, i.e. si $\bar{X}_n = 0$, or pour tout $n \geq 1$ fixé

$$\inf_{\theta>0} \mathbb{P}_\theta(\theta \in (\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))) = 1 - \sup_{\theta>0} \mathbb{P}_\theta(\theta \notin (\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))),$$

avec

$$\mathbb{P}_\theta(\theta \notin (\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))) \geq \mathbb{P}_\theta(\bar{X}_n = 0) = e^{-n\theta},$$

et $\sup_{\theta>0} e^{-n\theta} = 1$, donc

$$\inf_{\theta>0} \mathbb{P}_\theta(\theta \in (\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))) = 0,$$

et a fortiori

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in (\underline{\theta_n}(\mathbf{X}), \overline{\theta_n}(\mathbf{X}))) = 0 < 1 - \alpha.$$

Bref, les $IC(\theta, n)$ ne sont pas des intervalles de confiance asymptotiques forts.

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 élections, ce sera plutôt : Alice va-t-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). On parle de test 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 $\Theta_1 = [0, 1/2[$. On parle cette fois de test unilatère.

Définition 15 (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 un 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 16 (Risques, niveau et puissance d'un test)
On appelle :

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

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

— *taille du test le réel*

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

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*

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

— *fonction puissance du test l'application*

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

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 commune et la fonction π est continue.

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

$$H_0 \ : \ \theta \leq 0 \quad \text{contre} \quad H_1 \ : \ \theta > 0$$

ce qui revient, en notant $\Theta_0 =]-\infty, 0]$ et $\Theta_1 =]0, +\infty[$, 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 $\mathcal{R}' =]0, +\infty[$, laquelle en Définition 15 correspond donc à la région de rejet $\mathcal{R} = \{\mathbf{x} \in \mathbb{R}^n,\ x_1 + \cdots + x_n > 0\}$. Calculons la fonction 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,

$$\pi(\theta) = \mathbb{P}_\theta(\hat{\theta}_n > 0) = 1 - \Phi(-\theta\sqrt{n}) = \Phi(\theta\sqrt{n}),$$

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 :

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

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

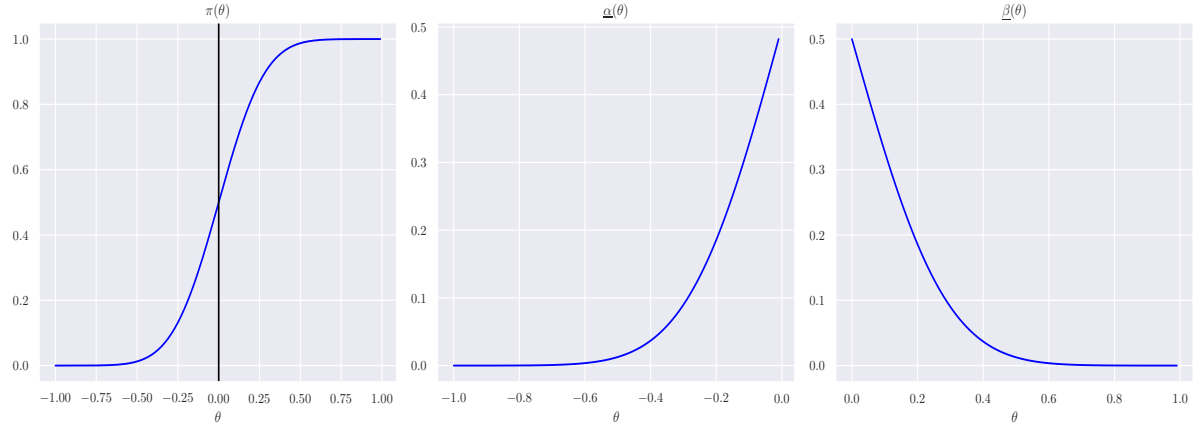

FIGURE 1.7 – Fonction puissance, risque de première espèce, risque de deuxième espèce ($n = 20$).

Dissymétrisation (Neyman & Pearson) : Pour s'en sortir, 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 ;
- 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 $\mathcal{R}'_\alpha =]c_\alpha, +\infty[$, avec $c_\alpha > 0$. Dit autrement, pour décider que le vrai paramètre θ est positif, la positivité de l'estimateur $\hat{\theta}_n$ ne suffit pas à nous convaincre : il faut que ce dernier soit supérieur à c_α, constante elle-même strictement positive. Reste à déterminer

c_α. Pour ce faire, il suffit d'écrire la condition sur le niveau du test en tenant compte du fait que $\hat{\theta}_n \sim \mathcal{N}(\theta, 1/n)$:

$$\sup_{\theta \leq 0} \mathbb{P}_\theta(\hat{\theta}_n > c_\alpha) \leq \alpha \iff \sup_{\theta \leq 0} \mathbb{P}(\mathcal{N}(\theta, 1/n) > c_\alpha) \leq \alpha \iff \sup_{\theta \geq 0} \mathbb{P}(\mathcal{N}(0,1) > (c_\alpha - \theta)\sqrt{n}) \leq \alpha$$

c'est-à-dire, puisque Φ est croissante :

$$\sup_{\theta \leq 0} \Phi((\theta - c_\alpha)\sqrt{n}) \leq \alpha \iff \Phi(-c_\alpha\sqrt{n}) \leq \alpha.$$

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

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

$$\pi(\theta) = \mathbb{P}_\theta(\hat{\theta}_n > q_{1-\alpha}/\sqrt{n}) = \Phi(\theta\sqrt{n} - q_{1-\alpha}).$$

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

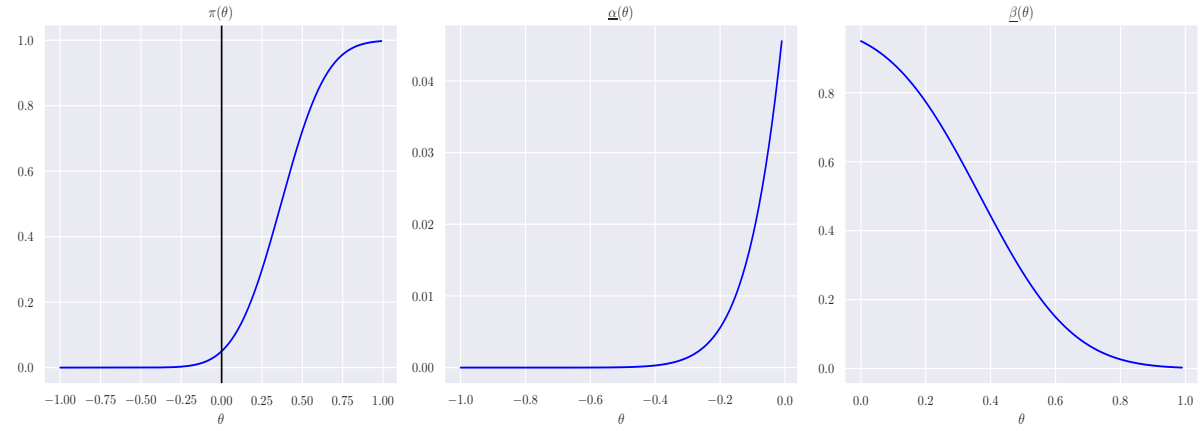

FIGURE 1.8 – Puissance, risque de première espèce, risque de deuxième espèce ($n = 20$, $\alpha = 5\%$).

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

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

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

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

Par conséquent, pour tout $\theta \in \Theta_0$,

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

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 :

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

et le test T est bien de niveau α.

■

Exemple : Dans l'exemple de l'échantillon gaussien, puisque $\hat{\theta}_n \sim \mathcal{N}(\theta, 1/n)$, on voit que pour tout θ, l'intervalle $[\hat{\theta}_n - q_{1-\alpha}/\sqrt{n}, +\infty[$ est un intervalle de confiance unilatère de niveau $(1-\alpha)$ pour θ. C'est en particulier vrai si $\theta \in \Theta_0 =]-\infty, 0]$. D'après ce qui vient d'être dit, on rejette H_0 lorsque

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

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

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

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 bien de niveau α pour tout $\alpha \in [0, 1]$. Il n'en reste pas moins qu'il n'a aucun intérêt. Notons enfin que si nous considérons l'intervalle de confiance a priori le plus pertinent pour θ, c'est-à-dire le plus court, il s'écrit

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

Par conséquent, le test résultant

$$T(\mathbf{X}) = \mathbf{1}_{I(\mathbf{X}) \cap \Theta_0 = \emptyset}$$

rejette H_0 si et seulement si $\hat{\theta}_n > q_{1-\alpha/2}/\sqrt{n}$ et est en fait de taille $\alpha/2$. Il est donc bien de niveau α, mais c'est en quelque sorte un excès de zèle par rapport à ce qui était requis, lequel se paie sur l'erreur de deuxième espèce, dont la borne supérieure sur Θ_1 vaut donc $1 - \alpha/2$ et non plus $1 - \alpha$.

Tout comme on parle de suite d'intervalles de confiance de niveau asymptotique $(1-\alpha)$, on peut introduire la notion de suite de tests de niveau asymptotique α. La définition générale fait intervenir la limite supérieure, mais dans tous les exemples que nous rencontrerons celle-ci sera en fait une limite classique.

Définition 17 (Niveau asymptotique d'une suite de tests)
On dit que la suite de tests $(T_n(\mathbf{X}))_{n\geq 1}$ est de niveau asymptotique α si

$$\forall \theta \in \Theta_0, \qquad \limsup_{n\to\infty} \mathbb{P}_\theta(T_n(\mathbf{X}) = 1) \leq \alpha.$$

Remarques :

1. Le raisonnement du Lemme 2 s'applique à nouveau et permet de faire le lien entre intervalle de confiance de niveau asymptotique $(1-\alpha)$ et test de niveau asymptotique α : si pour tout $\theta \in \Theta_0$, $(I_n(\mathbf{X}))_{n\geq 1}$ est une suite d'intervalles de confiance de niveau asymptotique $(1-\alpha)$, alors la suite de tests $(T_n(\mathbf{X}))_{n\geq 1}$ définie pour tout n par

$$T_n(\mathbf{X}) = \mathbf{1}_{I_n(\mathbf{X}) \cap \Theta_0 = \emptyset}$$

est de niveau asymptotique α.

2. A l'instar de la remarque faite pour les intervalles de confiance asymptotiques, noter que **nous n'exigeons pas** la condition plus forte

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

Exemples :

1. 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

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1), \tag{1.6}$$

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

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

Là encore, c'est en particulier vrai si $\theta \in \Theta_0 = [1/2, 1]$. Le Lemme 2 assure donc que le test $T_n(\mathbf{X}) = \mathbf{1}_{I_n(\mathbf{X})\cap\Theta_0=\emptyset}$ est de niveau asymptotique α. Ainsi, on rejette H_0 lorsque

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

Noter que si l'on définit la fonction puissance asymptotique sur tout Θ par

$$\pi_\infty(\theta) = \lim_{n\to\infty} \mathbb{P}_\theta(T_n(\mathbf{X}) = 1) = \lim_{n\to\infty} \mathbb{P}_\theta\left(\sqrt{n}\frac{\frac{1}{2} - \hat{\theta}_n}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} > q_{1-\alpha}\right),$$

on voit que $\pi_\infty(\theta) = \mathbf{1}_{\theta<1/2} + \alpha\mathbf{1}_{\theta=1/2}$.

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

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

Si $\theta = 1/2$, on déduit du TCL que

$$2\sqrt{n}\left(\hat{\theta}_n - \frac{1}{2}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

donc un intervalle bilatère et asymptotique de niveau $(1-\alpha)$ pour $\theta = 1/2$ est

$$\left[\hat{\theta}_n - \frac{q_{1-\alpha/2}}{2\sqrt{n}}, \hat{\theta}_n + \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right]. \tag{1.7}$$

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

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

Prenons par exemple $n = 1000$ et $\alpha = 5\%$, donc $q_{1-\alpha/2} = 1.96... \approx 2$. On rejette H_0 si $|\hat{\theta}_n - 1/2| > 0.03$.

Remarques :

1. L'exemple précédent a ceci de notable que l'intervalle de confiance (1.7) n'a aucun intérêt puisqu'un intervalle de confiance de niveau 100% pour la valeur $\theta = 1/2$ est tout simplement $I = \{1/2\}$. Néanmoins, le test construit à partir de cet intervalle stupide est, lui, pertinent. A contrario, si l'on applique le Lemme 2 à partir de l'intervalle de confiance optimal $I = \{1/2\}$, on aboutit à un test sans intérêt puisqu'on ne rejette jamais H_0.
2. Toujours sur l'exemple précédent du Pile ou Face, on peut recycler le raisonnement fait pour les élections : la normalité asymptotique (1.6) assure que, pour tout $\theta \in]0,1[$, un intervalle de confiance bilatère de niveau asymptotique $(1-\alpha)$ pour θ est

$$\left[\hat{\theta}_n - q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}, \hat{\theta}_n + q_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\right]. \tag{1.8}$$

Ceci est en particulier vrai pour $\theta = 1/2$ même si, à nouveau, cet intervalle de confiance n'a aucun intérêt a priori lorsqu'on connaît la valeur de θ. Quoi qu'il en soit, le test consistant à rejeter H_0 lorsque $1/2$ n'appartient pas à cet intervalle est de niveau asymptotique α. Comme $\hat{\theta}_n$ tend p.s. vers $1/2$ avec $\hat{\theta}_n(1-\hat{\theta}_n) \leq 1/4$, on constate que les tests sont asymptotiquement équivalents et que, à n fini, on rejette plus souvent H_0 en se basant sur (1.8) plutôt que sur (1.7).

Règle : Il ressort de ces exemples que si l'on veut construire un test unilatère, on part d'intervalles de confiance unilatères de sens opposé à Θ_0. Pour un test bilatère avec $\Theta_0 = \{\theta_0\}$, on part idéalement de l'intervalle de confiance non trivial le plus court possible pour θ_0.

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 $\hat{\theta}_n(\mathbf{x}) = 0.3$. Pour cette valeur, 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

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

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

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

Exemple : Sur l'exemple de l'échantillon gaussien, $\mathcal{R}'_\alpha =]\Phi^{-1}(1-\alpha)/\sqrt{n}, +\infty[$ et la décroissance de la fonction $\alpha \mapsto \Phi^{-1}(1-\alpha)/\sqrt{n}$ montre que les régions sont en effet emboîtées.

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

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

Exemple : Pour l'exemple de l'échantillon gaussien, on a donc de façon générale

$$\alpha_0(\mathbf{x}) = 1 - \Phi(\sqrt{n}\hat{\theta}_n(\mathbf{x})),$$

et sur le cas particulier où $\hat{\theta}_n(\mathbf{x}) = 0.3$, ceci donne une p-value d'environ 10^{-3}.

Remarque : Pour une famille de suites de tests de niveaux asymptotiques α, on définit logiquement la p-value (sous-entendu : asymptotique) comme l'infimum des α tel que H_0 est rejetée au niveau asymptotique α.

Take-home message : C'est cette valeur $\alpha_0(\mathbf{x})$ qui est usuellement donnée par les logiciels de statistique 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 [10]. Si au contraire $\alpha_0(\mathbf{x})$ est grand (disons supérieur à $1/10$), il semble raisonnable de conserver H_0. Pour des valeurs intermédiaires de $\alpha_0(\mathbf{x})$, rien n'est clair...

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 $\theta \leq 0$, le scénario le plus vraisemblable pour observer une valeur positive 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 supérieure ou égale à 0.3 est, avec $n = 100$,

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

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 positive de $\hat{\theta}_n$ si H_0 était vraie. Le "au moins aussi positive" vient du test fait ici et de H_0, qui suppose $\theta \leq 0$. 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

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

10. Noter toutefois qu'en pratique ceci dépend complètement du domaine d'application !

On observe $\mathbf{x} = (x_1, \ldots, x_n)$: quelle est la p-value associée ? On a vu 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 asymptotique 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 $c \geq 0$

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

on peut aussi écrire

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

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'on peut donc écrire

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

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 la p-value se reformule comme suit :

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

où, pour chaque valeur de $\theta \in \Theta_0$, $\mathbf{X}$ (aléatoire !) a pour loi P_θ. Nous nous contentons d'établir ce résultat dans le cas confortable d'une fonction de répartition bijective (sous-entendu : de son support, supposé être un intervalle, vers l'image de celui-ci, donc continue et strictement croissante sur son support [11]).

11. Le support $\mathcal{S}$ de la loi d'une variable X est le plus petit fermé de mesure pleine, i.e. tel que $\mathbb{P}(X \in \mathcal{S}) = 1$.

Lemme 3 (Interprétation de la p-value)
Supposons qu'il existe $\theta_0 \in \Theta_0$ tel que le test rejette H_0 au niveau α si et seulement si $S(\mathbf{X}) > c_\alpha = F_{\theta_0}^{-1}(1-\alpha)$, où $F_{\theta_0}(s) = \mathbb{P}_{\theta_0}(S(\mathbf{X}) \leq s)$ est la fonction de répartition de $S(\mathbf{X})$ lorsque le paramètre est θ_0. F_{θ_0} est supposée bijective et telle que $F_{\theta_0}(s) = \inf_{\theta\in\Theta_0} F_\theta(s)$ pour tout s. Alors, pour une réalisation $\mathbf{x}$, la p-value $\alpha_0(\mathbf{x})$ s'écrit encore

$$\alpha_0(\mathbf{x}) = \mathbb{P}_{\theta_0}(S(\mathbf{X}) > S(\mathbf{x})) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(S(\mathbf{X}) > S(\mathbf{x})).$$

Preuve : Par définition du test, pour une réalisation $\mathbf{x}$ et puisque F_{θ_0} est strictement croissante, la p-value est

$$\alpha_0(\mathbf{x}) = \inf\left\{\alpha \in [0,1], S(\mathbf{x}) > F_{\theta_0}^{-1}(1-\alpha)\right\} = \inf\left\{\alpha \in [0,1], F_{\theta_0}(S(\mathbf{x})) > 1-\alpha\right\}.$$

On en déduit la première formule :

$$\alpha_0(\mathbf{x}) = 1 - F_{\theta_0}(S(\mathbf{x})) = \mathbb{P}_{\theta_0}(S(\mathbf{X}) > S(\mathbf{x})).$$

De plus, de par la minimalité de F_{θ_0} parmi les F_θ, il vient

$$\alpha_0(\mathbf{x}) = 1 - \inf_{\theta\in\Theta_0} F_\theta(S(\mathbf{x})) = \sup_{\theta\in\Theta_0} \left(1 - F_\theta(S(\mathbf{x}))\right),$$

c'est-à-dire

$$\alpha_0(\mathbf{x}) = \sup_{\theta\in\Theta_0} \left(1 - \mathbb{P}_\theta(S(\mathbf{X}) \leq S(\mathbf{x}))\right) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(S(\mathbf{X}) > S(\mathbf{x})).$$

■

Exemples :

— Echantillon gaussien : $\Theta_0 = \mathbb{R}_-$ et on a vu que $S(\mathbf{X}) = \sqrt{n}\hat{\theta}_n(\mathbf{X})$. Pour tout $\theta \leq 0$, $F_\theta(s) = \Phi(s - \theta\sqrt{n})$ donc $\inf_{\theta\leq 0} F_\theta(s) = \Phi(s) = F_0(s)$ qui est bijective et on retrouve bien le fait que

$$\alpha_0(\mathbf{x}) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(S(\mathbf{X}) > S(\mathbf{x})) = \mathbb{P}_0(S(\mathbf{X}) > S(\mathbf{x})) = \mathbb{P}(\mathcal{N}(0,1) > \sqrt{n}\hat{\theta}_n(\mathbf{x})).$$

— Alice et Bob : soit $\theta \in \Theta_0 = [1/2, 1]$ et

$$S(\mathbf{x}) = -\sqrt{n}\frac{\hat{\theta}_n(\mathbf{x}) - \frac{1}{2}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}.$$

Si $\theta = 1/2$, alors on sait que

$$S(\mathbf{X}) = -\sqrt{n}\frac{\hat{\theta}_n(\mathbf{X}) - \frac{1}{2}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Si $\theta > 1/2$, alors de la loi des grands nombres on déduit en raisonnant "ω par ω" que

$$S(\mathbf{X}) = -\sqrt{n}\frac{\hat{\theta}_n(\mathbf{X}) - \frac{1}{2}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to\infty]{p.s.} -\infty.$$

Asymptotiquement, on a donc $\inf_{\theta\geq 0} F_\theta(s) = \Phi(s) = F_0(s)$ et on conclut comme dans l'exemple précédent.

— Pile ou Face : $S(\mathbf{x}) = 2\sqrt{n}|\hat{\theta}_n(\mathbf{x}) - 1/2|$. Soit $Y \sim \mathcal{N}(0,1)$ et $Z = |Y|$, alors si $\theta = 1/2$ on a la convergence en loi

$$S(\mathbf{X}) = 2\sqrt{n}|\hat{\theta}_n(\mathbf{X}) - 1/2| \xrightarrow[n\to\infty]{d} Z,$$

et $c_\alpha = q_{1-\alpha/2} = F_Z^{-1}(1-\alpha)$. Le résultat s'applique à nouveau puisque $\Theta_0 = \{1/2\}$.

A retenir : On résume souvent le résultat du Lemme 3 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."

Remarque : D'un point de vue historique, il semblerait que la notion de p-value trouve ses origines dans une controverse à base de thé et de lait, donc typiquement britannique, entre Muriel Bristol et Ronald Fisher, anecdote connue sous le nom de The lady tasting tea.

Chapitre 2

Estimation unidimensionnelle

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 i.i.d. de loi inconnue P_X. La Section 2.1 présente les quantités dites empiriques liées à cet échantillon et quelques résultats afférents. On se restreint par la suite à des variables suivant 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. La Section 2.2 présente deux techniques classiques d'estimation : la méthode des moments et celle du maximum de vraisemblance. Finalement, dans le cadre des modèles réguliers, la Section 2.3 explique en quoi la notion d'information de Fisher permet de spécifier l'optimalité d'un estimateur.

2.1 Quantités empiriques

2.1.1 Moyenne et variance empiriques

Partant d'un échantillon $(X_n)_{n\geq 1}$ i.i.d. de variables intégrables, 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 forte des grands nombres et du théorème central limite.

Proposition 7 (Convergence et normalité asymptotique de la moyenne empirique) *Si les variables $(X_n)_{n\geq 1}$ sont i.i.d., ont un moment d'ordre 2 avec $\mathbb{E}[X_1] = \mu$ et $\mathrm{Var}(X_1) = \sigma^2 > 0$, alors la moyenne empirique $\bar{X}_n$ est un estimateur non biaisé, fortement consistant et asymptotiquement normal :*

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

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 7, 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 $\hat{\sigma}_n^2$ qui sera défini au Chapitre 3. Par ailleurs, c'est l'estimateur non biaisé $\hat{s}_n^2$ qui est considéré par de nombreux logiciels (cf. la commande `sd` de R, qui fournit l'écart-type associé).

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

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2. \tag{2.1}$$

La clé de la preuve est la relation $\mathbb{E}[Y^2] = \mathrm{Var}(Y) + \mathbb{E}[Y]^2$. Ainsi, 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

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

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

Proposition 8 (Convergence et normalité asymptotique de la variance empirique)
Si les variables $(X_n)_{n\geq 1}$ *sont i.i.d. 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 fortement consistants :*

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

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 :*

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

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 consistance, on part de la formule (2.1) à laquelle on applique deux fois la loi des grands nombres et le théorème de continuité :

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

Par la remarque ci-dessus, le même résultat s'applique à $\hat{s}_n^2$. Pour la normalité asymptotique, la ruse est de considérer les variables i.i.d. centrées $Y_i = (X_i - \mu)$ et de noter 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

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

d'où par Slutsky

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

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

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

Il reste à appliquer Slutsky pour recoller les morceaux :

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

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.

■

Remarque : Par le résultat précédent et le Lemme de Slutsky, un intervalle de confiance de niveau asymptotique $(1-\alpha)$ pour μ est donc

$$\left[\bar{X}_n - \frac{\Phi^{-1}(1-\alpha/2)\hat{\sigma}_n}{\sqrt{n}}; \bar{X}_n + \frac{\Phi^{-1}(1-\alpha/2)\hat{\sigma}_n}{\sqrt{n}}\right].$$

Ce résultat reste bien sûr valable avec $\hat{s}_n$ en lieu et place de $\hat{\sigma}_n$.

2.1.2 Fonction de répartition et quantiles empiriques

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

Définition 19 (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 i.i.d., 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 pour une loi sans atome, c'est-à-dire si la fonction de répartition des X_i est continue [1].

Définition 20 (Fonction de répartition empirique)
La fonction de répartition empirique F_n d'un échantillon $X_1, \dots, X_n$ est définie pour tout réel x par

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

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.

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 2.1). 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 9 (Loi, convergence et normalité asymptotique)
*Soit $(X_n)_{n \geq 1}$ des variables i.i.d. de fonction de répartition F, alors pour tout réel x **fixé**, on a :*

1. En effet, $\mathbb{P}(X_1 = X_2) = \lim_{N \to \infty} \mathbb{P}(|X_2 - X_1| \leq 1/N) = \lim_{N \to \infty} \mathbb{E}[\mathbf{1}_{|X_2 - X_1| \leq 1/N}]$ or par continuité de F et conditionnement $\mathbb{E}[\mathbf{1}_{|X_2 - X_1| \leq 1/N}] = \mathbb{E}[F(X_1 + 1/N) - F(X_1 - 1/N)]$ et on conclut par convergence dominée.

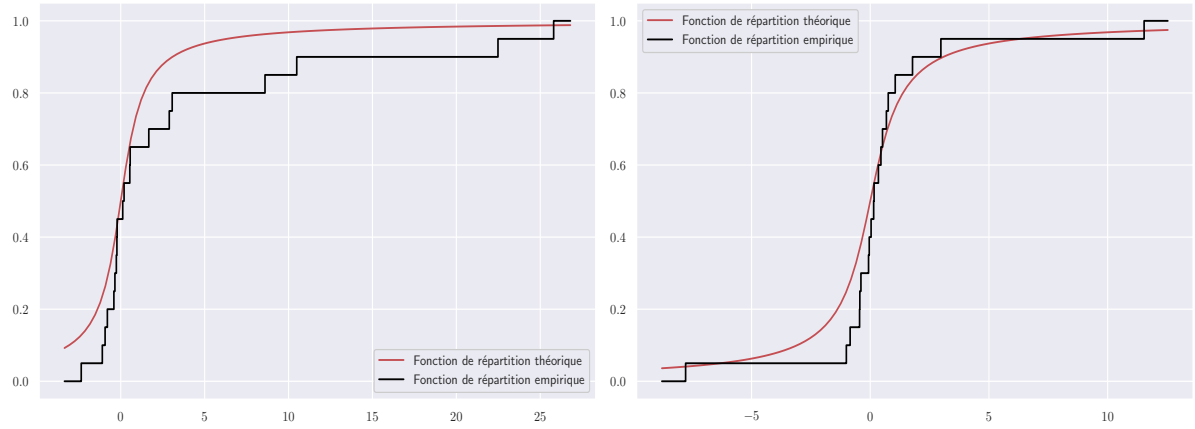

FIGURE 2.1 – Deux réalisations de F_{20} avec $X_1, \dots, X_{20}$ i.i.d. selon une loi de Cauchy.

— *Loi : la variable aléatoire $nF_n(x)$ suit une loi binomiale $\mathcal{B}(n, F(x))$. En particulier, on en déduit que $\mathbb{E}[F_n(x)] = F(x)$ et $\mathrm{Var}(F_n(x)) = F(x)(1 - F(x))$.*

— *Consistance forte :*

$$F_n(x) \xrightarrow[n\to\infty]{p.s.} F(x).$$

— *Normalité asymptotique :*

$$\sqrt{n}\,(F_n(x) - F(x)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, F(x)(1 - F(x))).$$

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 i.i.d. 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 forte 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

$$\sqrt{n}\,(F_n(x) - F(x)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \mathrm{Var}(Y_1)) = \mathcal{N}(0, F(x)(1 - F(x))).$$

■

Ainsi, pour tout réel x, il existe un ensemble $\Omega_0(x)$ de probabilité 1 tel que, pour tout $\omega \in \Omega_0(x)$, la réalisation $x_1 = X_1(\omega), x_2 = X_2(\omega), \dots$ vérifie

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(x_i) = \sum_{i=1}^{n} \frac{i}{n} \mathbb{1}_{[x_{(i)},x_{(i+1)}[}(x) \xrightarrow[n\to\infty]{} F(x).$$

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 on peut montrer que, de façon presque sûre, il y a bien convergence simple et même mieux, convergence uniforme : le Théorème de Glivenko-Cantelli, que nous ne démontrons pas ici [2], assure en effet que

$$\|F_n - F\|_\infty := \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{p.s.} 0.$$

Le fait que F_n s'approche de F lorsque n tend vers l'infini est illustré Figure 2.2.

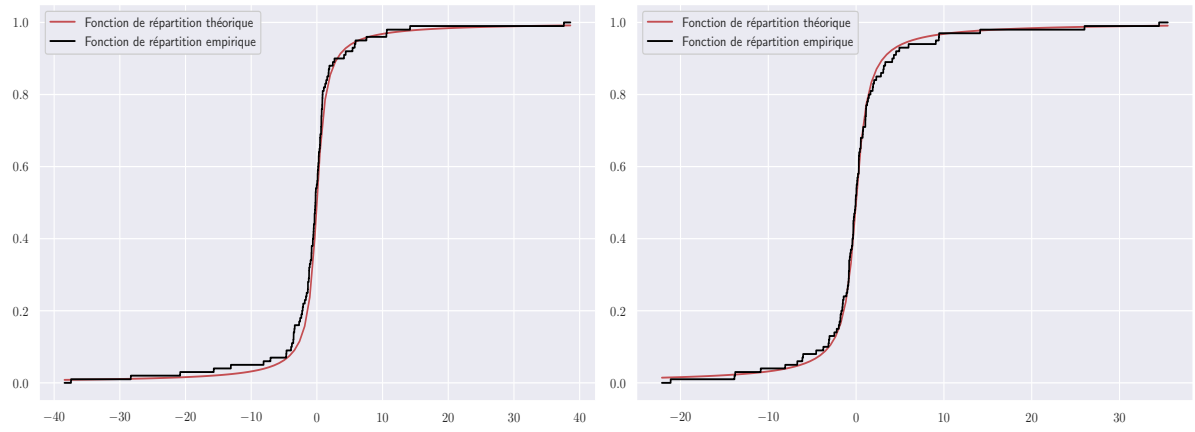

FIGURE 2.2 – Deux réalisations de F_{100} avec $X_1, \ldots, X_{100}$ i.i.d. selon une loi de Cauchy.

Un quantile est défini à partir de la fonction de répartition. Il n'y aucun problème lorsque celle-ci est bijective. Si tel n'est pas le cas, il faut faire un peu attention. Ceci arrivera en particulier pour les fonctions de répartition empiriques que nous aborderons ultérieurement.

Définition 21 (Inverse généralisée)
Soit F une fonction de répartition. On appelle inverse généralisée de F, ou fonction quantile, 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* $\inf \emptyset = +\infty$.

Remarque : ainsi, on peut noter que $F^{-1}(0) = -\infty$, tandis que $F^{-1}(1)$ est la borne supérieure du support de la loi de X lorsque cette variable a pour fonction de répartition F.

Si F est bijective, il est clair que cette fonction quantile coïncide avec l'inverse classique (au sens de fonction réciproque) de F, avec les conventions évidentes aux limites. C'est en particulier le cas pour Φ, c'est-à-dire la fonction de répartition de la loi normale standard (voir Figure 2.3).

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 $u \in]0,1[$,

$$F^{-1}(u) = \begin{cases} x_1 & \text{si } 0 < u \leq p_1 \\ x_2 & \text{si } p_1 < u \leq p_1 + p_2 \\ \vdots & \\ x_m & \text{si } p_1 + \cdots + p_{m-1} < u \leq 1 \end{cases}$$

2. Voir par exemple [8], Théorème 2.5.

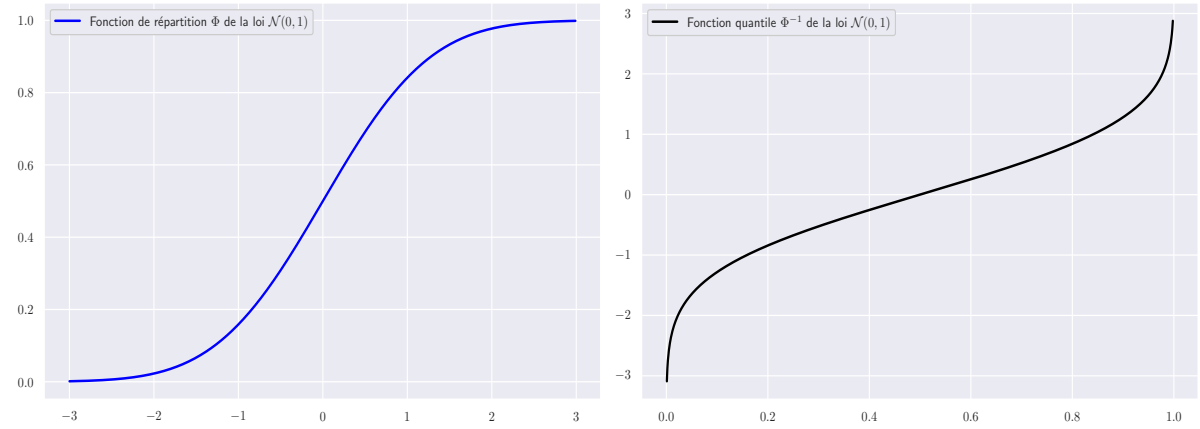

FIGURE 2.3 – Fonction de répartition Φ et fonction quantile Φ^{-1} de la loi normale centrée réduite.

c'est-à-dire

$$F^{-1}(u) = \sum_{k=1}^{m} x_k \mathbb{1}_{p_1+\cdots+p_{k-1}<u\leq p_1+\cdots+p_k}. \tag{2.2}$$

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 : $\forall u \in [0, 1]$,*

$$F(x) \geq u \iff x \geq F^{-1}(u). \tag{2.3}$$

5. *Inversibilité : $\forall u \in [0, 1]$, on a $(F \circ F^{-1})(u) \geq u$. De plus :*
   — *si F est continue alors $F \circ F^{-1} = Id$, mais si elle n'est pas injective il existe x_0 tel que $(F^{-1} \circ F)(x_0) < x_0$;*
   — *si F est injective alors $F^{-1} \circ F = Id$, mais si elle n'est pas continue 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.*

Preuve : Les deux premiers points découlent de la définition de F^{-1}. Etablissons l'équivalence (2.3) : avec la convention $F^{-1}(0) = -\infty$, il n'y a rien à montrer pour $u = 0$, donc on peut considérer $u \in]0, 1]$. Par définition de $F^{-1}(u)$, si $F(x) \geq u$, alors $x \geq F^{-1}(u)$. Inversement, si $F^{-1}(u) \leq x$, alors pour tout $\varepsilon > 0$ on a $F^{-1}(u) < x + \varepsilon$, donc par définition de $F^{-1}(u)$, il vient $u \leq F(x + \varepsilon)$. Puisque F est continue à droite, on en déduit que $u \leq F(x)$ et l'équivalence (2.3) est établie.

La continuité à gauche en découle : puisqu'il n'y a rien à prouver pour $u = 0$, il suffit en effet de montrer, grâce à la croissance de F^{-1}, que pour tout $u \in]0, 1]$ et tout $\varepsilon > 0$, on peut trouver $\delta > 0$ tel que $F^{-1}(u - \delta) > F^{-1}(u) - \varepsilon =: x'$. Puisque $x' < F^{-1}(u)$, (2.3) assure que $F(x') < u$ donc $F(x') < u - \delta$ pour δ assez petit. Ceci implique $x' < F^{-1}(u - \delta)$, c'est-à-dire précisément ce qu'il fallait établir.

Pour le dernier point, il n'y a rien à prouver si $u = 0$. Si $u \in]0, 1]$, d'après (2.3), on a

$$F^{-1}(u) \leq F^{-1}(u) \implies u \leq (F \circ F^{-1})(u).$$

Supposons maintenant F continue. Alors, pour tout $u \in]0, 1]$ et pour tout $\varepsilon > 0$, on a, toujours par (2.3),

$$F^{-1}(u) - \varepsilon < F^{-1}(u) \implies F(F^{-1}(u) - \varepsilon) < u.$$

Etant donné que $u \in]0, 1]$ et que F est supposée continue, le passage à la limite lorsque $\varepsilon \to 0$ donne $(F \circ F^{-1})(u) \leq u$. Au total, on a donc prouvé que, pour tout $u \in]0, 1]$, $(F \circ F^{-1})(u) = u$. Avec les conventions prises pour F et F^{-1}, ceci est encore vrai pour $u = 0$. Supposons F non injective, ce qui signifie qu'il existe $x_0' < x_0$ tels que $F(x_0') = F(x_0) = u_0$, donc

$$(F^{-1} \circ F)(x_0) = F^{-1}(u_0) \leq x_0' < x_0.$$

Dans le même ordre d'idée, si F est injective, alors quel que soit le réel x, il n'existe pas de réel $x' < x$ tel que $F(x') = F(x)$, donc

$$F^{-1}(F(x)) = \inf\{x' \in \mathbb{R},\ F(x') \geq F(x)\} = x.$$

Si F n'est pas continue en un point x_0, il existe u_0 tel que $F(x_0^-) < u_0 < F(x_0)$, auquel cas

$$(F \circ F^{-1})(u_0) = F(F^{-1}(u_0)) = F(x_0) > u_0.$$

Quant au dernier point, il correspond exactement à la définition de la réciproque d'une fonction bijective, de sorte qu'il n'y a rien à démontrer.

■

Remarque : La preuve ci-dessus montre que si F est continue en $F^{-1}(u_0)$ alors $(F \circ F^{-1})(u_0) = u_0$.

Exemples : Illustrons le dernier point des Propriétés 1.

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 $X = BY$, 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 2.4). Elle est injective mais pas continue, et on voit que

$$(F \circ F^{-1})(1/2) = F(0) = \frac{3}{4} > \frac{1}{2}.$$

Le résultat suivant est utile tant d'un point de vue pratique, par exemple pour les méthodes Monte-Carlo, que théorique, typiquement pour l'étude du processus empirique.

Lemme 2 (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 :

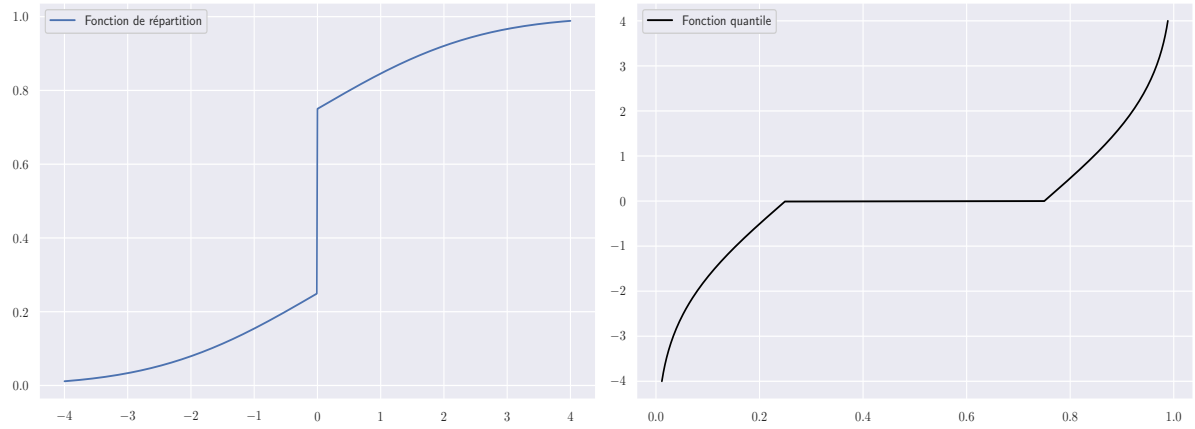

FIGURE 2.4 – Fonction de répartition de la variable $X = 2BY$ et son inverse généralisée.

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. Par exemple, si $X \sim \mathcal{B}(1/3)$, alors $F(X)$ est une variable discrète prenant les valeurs $F(0) = 2/3$ et $F(1) = 1$ avec les probabilités respectives $2/3$ et $1/3$.

Application : méthode d'inversion en Monte-Carlo. Supposons que l'on dispose d'un générateur aléatoire de variables uniformes, comme c'est le cas pour tous les logiciels[3]. Alors, si la fonction de répartition F est facilement inversible, on déduit du résultat précédent une méthode simple pour générer une variable de fonction de répartition F à partir de la simulation d'une variable uniforme.

Exemples :

1. Simulation d'une variable exponentielle. On veut générer une variable X selon la loi exponentielle de paramètre $\lambda > 0$ fixé connu. Pour tout $x > 0$, $F(x) = 1 - e^{-\lambda x}$, bijective de $]0, \infty[$ vers $]0, 1[$. Il s'ensuit que pour tout $u \in]0, 1[$, $F^{-1}(u) = -(\log(1 - u))/\lambda$. Dès lors, si U suit une loi uniforme sur $[0, 1]$, la variable $X = -(\log(1 - U))/\lambda$ suit une loi exponentielle de paramètre 1. Puisque U a la même loi que $1 - U$, on peut même aller plus vite en considérant $X = -(\log U)/\lambda$.

3. C'est en fait un générateur pseudo-aléatoire, mais passons.

2. Simulation d'une variable de Cauchy. On veut générer une variable X selon la loi de Cauchy standard, c'est-à-dire de densité $f(x) = 1/(\pi(1+x^2))$, donc de fonction de répartition $F(x) = (\pi/2 + \arctan x)/\pi$, bijective de $\mathbb{R}$ vers $]0,1[$. Par la méthode d'inversion, si U suit une loi uniforme sur $]0,1[$, $X = \tan(\pi(U-1/2))$ suit une loi de Cauchy.

Maintenant qu'on a défini l'inverse d'une fonction de répartition en toute généralité, on peut passer aux quantiles.

Définition 22 (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 $+\infty$). De plus, la Proposition 1 assure que

$$\forall p \in [0,1] \qquad F(x_p) = F(F^{-1}(p)) \geq p. \tag{2.4}$$

On peut aussi définir les quantiles empiriques : ils coïncident avec les points de l'échantillon puisque c'est uniquement en ceux-ci que la fonction de répartition empirique varie.

Notation : pour tout réel x, $\lceil x \rceil$ désigne la partie entière supérieure de x, c'est-à-dire le plus petit entier supérieur ou égal à x. En particulier, elle vérifie : $x \leq \lceil x \rceil < x+1$.

Lemme 3 (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, avec la convention $X_{(0)} = -\infty$,

$$x_p(n) = F_n^{-1}(p) = \inf\{x \in \mathbb{R} :\ F_n(x) \geq p\} = X_{(\lceil np \rceil)}.$$

Preuve : Le but est de prouver la dernière égalité. Celle-ci est évidente si $p = 0$ avec la convention adoptée. Si $0 < p \leq 1$, alors $1 \leq \lceil np \rceil \leq n$ et, puisque $X_{(1)} \leq \cdots \leq X_{(\lceil np \rceil)} \leq \cdots \leq X_{(n)}$, il est clair que

$$F_n(X_{(\lceil np \rceil)}) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{X_{(j)} \leq X_{(\lceil np \rceil)}} \geq \frac{\lceil np \rceil}{n} \geq p,$$

donc $x_p(n) = F_n^{-1}(p) \leq X_{(\lceil np \rceil)}$. Supposons maintenant que $F_n^{-1}(p) < X_{(\lceil np \rceil)}$. Rappelons que $F_n^{-1}(p)$ est l'un des points de l'échantillon. Dès lors, si $F_n^{-1}(p) < X_{(\lceil np \rceil)}$, alors il y a au plus $\lceil np \rceil - 1$ indices j tels que $X_j \leq F_n^{-1}(p)$, donc

$$F_n(F_n^{-1}(p)) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{X_j \leq F_n^{-1}(p)} \leq \frac{\lceil np \rceil - 1}{n} < p,$$

ce qui est en contradiction avec (2.4).

■

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 $(X_1, \ldots, X_n)$ i.i.d. selon F, que dire du p-quantile empirique $x_p(n)$? Sans prendre de précautions, ça peut mal se passer...

Théorème 8 (Convergence et normalité asymptotique du quantile empirique)
Soit $(X_1, \dots, X_n)$ i.i.d. 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*

$$\sqrt{n}(x_p(n) - x_p) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{p(1-p)}{f(x_p)^2}\right).$$

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

$$\mathbb{P}(|x_p(n) - x_p| > \varepsilon) \leq \alpha \exp(-\beta_{p,\varepsilon} n),$$

et Borel-Cantelli permettra de conclure. Vu la dissymétrie induite par l'inverse généralisée, on commence par scinder le terme à majorer :

$$\mathbb{P}(|x_p(n) - x_p| > \varepsilon) = \mathbb{P}(x_p(n) < x_p - \varepsilon) + \mathbb{P}(x_p(n) > x_p + \varepsilon). \tag{2.5}$$

Pour le premier, il découle de l'équivalence (2.3) que

$$\mathbb{P}(x_p(n) < x_p - \varepsilon) \leq \mathbb{P}(F_n^{-1}(p) \leq x_p - \varepsilon) = \mathbb{P}(nF_n(x_p - \varepsilon) \geq np) = \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{X_i \leq x_p - \varepsilon} \geq np\right),$$

où l'on reconnaît une somme de variables de Bernoulli i.i.d. :

$$S_n = \sum_{i=1}^n B_i = \sum_{i=1}^n \mathbb{1}_{]-\infty, x_p - \varepsilon]}(X_i) \sim \mathcal{B}(n, F(x_p - \varepsilon)) \implies \mathbb{E}[S_n] = nF(x_p - \varepsilon).$$

Ainsi

$$\mathbb{P}(x_p(n) < x_p - \varepsilon) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \geq n(p - F(x_p - \varepsilon))).$$

Or, par définition de $x_p = \inf\{x,\ F(x) \geq p\}$, on a, pour tout $\varepsilon > 0$, $F(x_p - \varepsilon) < p$ donc

$$n(p - F(x_p - \varepsilon)) =: n\delta > 0.$$

A ce stade, Hoeffding s'impose (cf. Chapitre 1 Proposition 4) :

$$\mathbb{P}(x_p(n) < x_p - \varepsilon) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \geq n\delta) \leq \exp(-2\delta^2 n),$$

terme général d'une série convergente. Le second terme de l'équation (2.5) se traite de façon comparable :

$$\mathbb{P}(x_p(n) > x_p + \varepsilon) = \mathbb{P}(F_n^{-1}(p) > x_p + \varepsilon) = \mathbb{P}(nF_n(x_p + \varepsilon) < np) \leq \mathbb{P}(nF_n(x_p + \varepsilon) \leq np),$$

c'est-à-dire

$$\mathbb{P}(x_p(n) > x_p + \varepsilon) \leq \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{X_i \leq x_p + \varepsilon} \leq np\right),$$

où l'on a cette fois

$$S_n = \sum_{i=1}^n \mathbb{1}_{]-\infty, x_p + \varepsilon]}(X_i) \sim \mathcal{B}(n, F(x_p + \varepsilon)) \implies \mathbb{E}[S_n] = nF(x_p + \varepsilon),$$

d'où

$$\mathbb{P}(x_p(n) > x_p + \varepsilon) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \leq n(p - F(x_p + \varepsilon))).$$

Or F étant globalement croissante et, par hypothèse, strictement croissante en x_p, l'inégalité (2.4) implique que pour tout $\varepsilon > 0$

$$F(x_p + \varepsilon) > F(x_p) \geq p \implies n(p - F(x_p + \varepsilon)) =: -n\gamma < 0.$$

On peut donc à nouveau appliquer Hoeffding :

$$\mathbb{P}(x_p(n) > x_p + \varepsilon) \leq \mathbb{P}(S_n - \mathbb{E}[S_n] \leq -n\gamma) \leq \exp(-2\gamma^2 n),$$

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 9 implique que

$$Z_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0, p(1-p)).$$

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, si x est positif,

$$n(F_n(x_p + x/\sqrt{n}) - F_n(x_p)) = \sum_{i=1}^{n} \mathbb{1}_{x_p < X_i \leq x_p + x/\sqrt{n}} \sim \mathcal{B}(n, F(x_p + x/\sqrt{n}) - F(x_p)) =: \mathcal{B}(n, \delta_n).$$

Si x est négatif, le même raisonnement montre que

$$-n(F_n(x_p + x/\sqrt{n}) - F_n(x_p)) \sim \mathcal{B}(n, F(x_p) - F(x_p + x/\sqrt{n})) = \mathcal{B}(n, -\delta_n).$$

Dans tous les cas, $\sqrt{n}|Y_n - Z_n|$ correspond en loi à une binomiale $\mathcal{B}(n, |\delta_n|)$ recentrée. L'inégalité de Tchebychev et la continuité de F en x_p assurent donc que, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(|Y_n - Z_n| \geq \varepsilon\right) \leq \frac{\delta_n(1 - \delta_n)}{\varepsilon^2} \xrightarrow[n\to\infty]{} 0,$$

c'est-à-dire que $(Y_n - Z_n)$ tend en probabilité vers 0. Au total, par le Lemme de Slutsky,

$$Y_n = Z_n + (Y_n - Z_n) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, p(1-p)).$$

Par ailleurs, (y_n) converge de façon déterministe, donc a fortiori en probabilité, vers $-f(x_p)x$ donc une nouvelle application du Lemme de Slutsky donne

$$Y_n - y_n - f(x_p)x \xrightarrow[n\to\infty]{d} \mathcal{N}(0, p(1-p)),$$

ce qui implique, pour tout réel t,

$$\mathbb{P}(Y_n - y_n - f(x_p)x > t) \xrightarrow[n\to\infty]{} 1 - \Phi\left(\frac{t}{\sqrt{p(1-p)}}\right) = \Phi\left(-\frac{t}{\sqrt{p(1-p)}}\right).$$

La valeur $t = -f(x_p)x$ donne

$$\mathbb{P}(\sqrt{n}(x_p(n) - x_p) \leq x) = 1 - G_n(y_n) = \mathbb{P}(Y_n > y_n) \xrightarrow[n\to\infty]{} \Phi\left(\frac{f(x_p)}{\sqrt{p(1-p)}}\, x\right),$$

ce qui est le résultat voulu.

■

Exemples :

1. On considère $(X_1, \ldots, X_n)$ i.i.d. 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

$$x_{1/2}(n) \xrightarrow[n\to\infty]{p.s.} x_{1/2} = \theta,$$

avec plus précisément

$$\sqrt{n}(x_{1/2}(n) - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \pi^2/4\right).$$

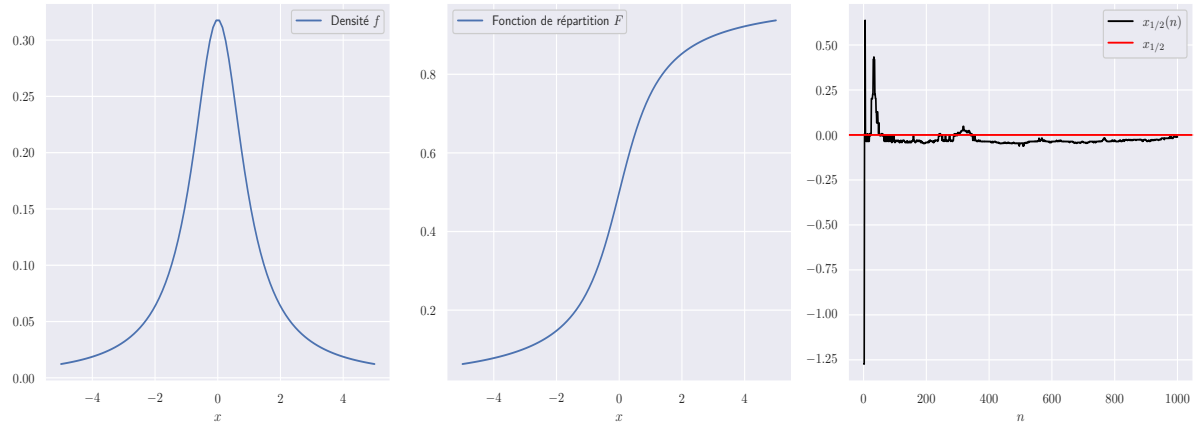

FIGURE 2.5 – Densité de Cauchy, fonction de répartition et convergence de la médiane empirique.

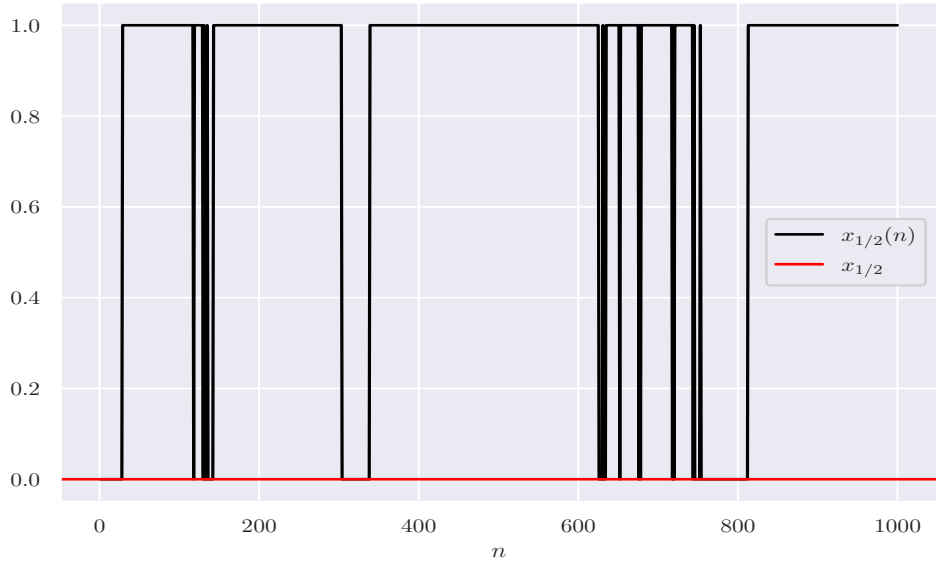

FIGURE 2.6 – Oscillation de la médiane empirique pour des variables de Bernoulli $\mathcal{B}(1/2)$.

Via l'approximation usuelle $\Phi^{-1}(0.975) \approx 2$, on en déduit par exemple qu'un intervalle de confiance de niveau asymptotique 95% pour θ est donné par

$$\left[x_{1/2}(n) - \frac{\pi}{\sqrt{n}} \; ; \; x_{1/2}(n) + \frac{\pi}{\sqrt{n}}\right].$$

Lorsque $\theta = 0$, la densité de la loi de Cauchy symétrique, sa fonction de répartition et la convergence de la médiane empirique sont illustrées Figure 2.5.

2. 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) de la valeur 0 à la valeur 1 (voir Figure 2.6).

3. 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 2.7 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 2.7 à droite).

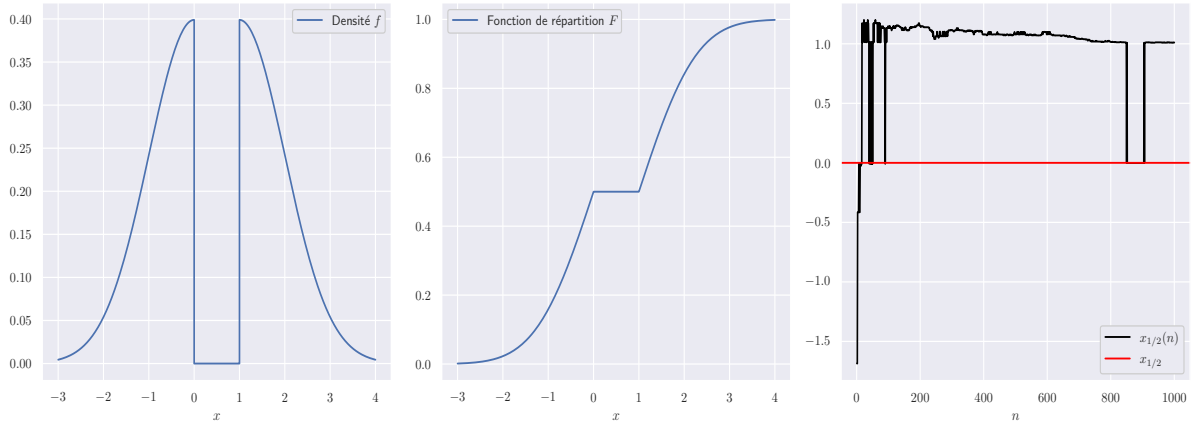

FIGURE 2.7 – Densité de $X = Y\mathbb{1}_{Y<0} + (1+Y)\mathbb{1}_{Y\geq 0}$, fonction de répartition et oscillation de la médiane empirique.

4. 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 -3 et +3. Formellement, en notant Y et Z les variables gaussiennes en question et B une variable de Bernoulli de paramètre 1/2, indépendante des 2 précédentes, ceci s'écrit [4] :

$$X = B \times Y + (1-B) \times Z \implies f(x) = \frac{1}{2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-3)^2}{2}} + \frac{1}{2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+3)^2}{2}}.$$

4. Pour trouver la densité, on peut commencer par calculer la fonction de répartition.

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 (voir Figure 2.8). 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.

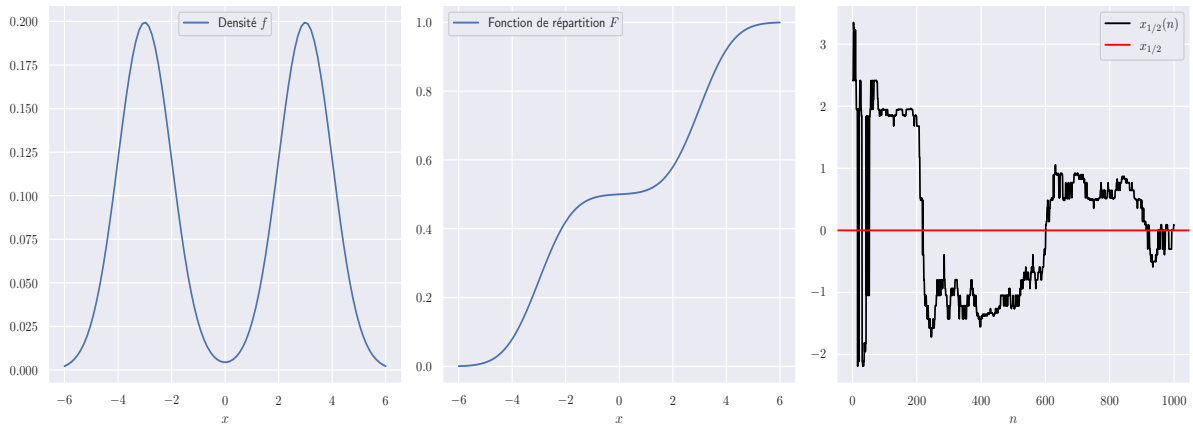

FIGURE 2.8 – Densité d'un mélange de gaussiennes, fonction de répartition et médiane empirique.

Remarque : Le résultat de normalité asymptotique du Théorème 8 ne permet pas de construire des intervalles de confiance si on ne connaît pas $f(x_p)$. 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 9, si $F(x_p) = p$, c'est-à-dire si F est continue en x_p, on a

$$\sqrt{n}\,(F_n(x_p) - F(x_p)) = \sqrt{n}\,(F_n(x_p) - p) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, p(1-p)),$$

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 l'équivalence (2.3) des Propriétés 1 avec F_n :

$$F_n(x) \geq u \iff x \geq F_n^{-1}(u) \qquad \text{et} \qquad F_n(x) < v \iff x < F_n^{-1}(v)$$

pour en déduire un intervalle de confiance de niveau asymptotique $(1-\alpha)$ pour x_p, à savoir :

$$\left[F_n^{-1}\left(p - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right), F_n^{-1}\left(p + \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\right[,$$

ou encore (qui peut le plus peut le moins) :

$$\left[F_n^{-1}\left(p - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right), F_n^{-1}\left(p + \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\right].$$

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 $[X_{(4900)}, X_{(5100)}]$.

Remarque : Le raisonnement précédent a ceci de remarquable qu'il ne suppose ni la connaissance de $f(x_p)$ ni sa stricte positivité ! La seule chose requise est la continuité de F en x_p. Un exemple d'application est donné en fin de section 2.2.2.

2.2 Estimation paramétrique unidimensionnelle

On se limite désormais au modèle paramétrique unidimensionnel, c'est-à-dire qu'on dispose d'un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires réelles i.i.d. de loi P_θ paramétrée par $\theta \in \Theta$, où θ est inconnu et Θ est un intervalle de $\mathbb{R}$. Cette section présente deux techniques classiques d'estimation de θ : méthodes des moments et du maximum de vraisemblance.

2.2.1 La méthode des moments

Nous avons vu en Proposition 6 que si

$$\sqrt{n}(\hat{\varphi}_n - \varphi(\theta)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2),$$

alors, sous des hypothèses idoines et en notant $\hat{\theta}_n := \varphi^{-1}(\hat{\varphi}_n)$, on a

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, (\sigma/\varphi'(\theta))^2).$$

Sous le nom de méthode des moments 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 $\varphi(\theta) = \mathbb{E}_\theta[X_1^k]$ pour un certain entier k, ou plus généralement $\varphi(\theta) = \mathbb{E}_\theta[h(X_1)]$. L'exemple le plus connu est celui où l'on estime $\varphi(\theta) = \mathbb{E}_\theta[X_1]$ par la moyenne empirique $\bar{X}_n$. Nous allons décliner cette idée 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

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1/3).$$

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 : les variables étant bornées, on peut écrire

$$\mathbb{P}(|\bar{X}_n - \theta| \geq c) \leq 2e^{-\frac{c^2 n}{2}},$$

d'où

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

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

$$\bar{X}_n \xrightarrow[n\to\infty]{p.s.} \frac{1}{\lambda} \qquad \text{et} \qquad \sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1/\lambda^2).$$

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 méthode Delta donne

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \lambda^2).$$

Lois Gamma

En guise de mise en bouche, on rappelle que la fonction Gamma, définie pour tout réel $r > 0$ par

$$\Gamma(r) = \int_0^{+\infty} x^{r-1}e^{-x}dx, \tag{2.6}$$

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 2 (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 khi-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 i.i.d. de loi* $\mathcal{E}(\lambda)$, *alors*

$$\sum_{i=1}^{n} X_i \sim \Gamma(n, \lambda) \qquad \text{et} \qquad \bar{X}_n \sim \Gamma(n, n\lambda).$$

— *Si* $(X_1, \dots, X_n)$ *sont i.i.d. 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 (voir Figure 2.9). 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

$$\frac{\lambda}{\sqrt{r}}\left(X_r - \frac{r}{\lambda}\right) \xrightarrow[r\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \iff \forall x \in \mathbb{R},\ \left|\mathbb{P}\left(\frac{\lambda}{\sqrt{r}}\left(X_r - \frac{r}{\lambda}\right) \le x\right) - \Phi(x)\right| \xrightarrow[r\to\infty]{} 0.$$

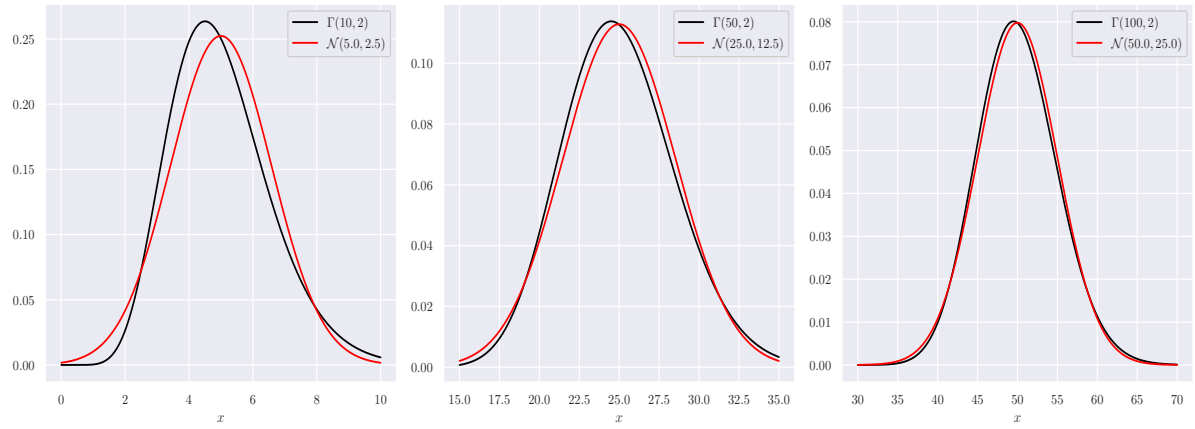

FIGURE 2.9 – Densités de lois $\Gamma(r,\lambda)$ et $\mathcal{N}(r/\lambda, r/\lambda^2)$ avec $\lambda = 2$ et $r \in \{10, 50, 100\}$.

Pour l'estimation de paramètres, partant d'un échantillon $(X_1, \dots, X_n)$ i.i.d. 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

$$\sqrt{n}\left(\bar{X}_n - \frac{r}{\lambda}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, r/\lambda^2) \iff \sqrt{\frac{n}{r}}\left(\lambda\bar{X}_n - r\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

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

$$\sqrt{n}\,\frac{\lambda\bar{X}_n - r}{\sqrt{\lambda\bar{X}_n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

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 $(Y_1, \ldots, Y_n)$ i.i.d. 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

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

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

$$\sqrt{n}(\hat{\sigma}_n - \sigma) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, (\sigma s/m)^2) \iff \sqrt{n}\,\frac{m}{s}\left(\frac{\hat{\sigma}_n}{\sigma} - 1\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

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

$$\sqrt{n}(\hat{\sigma}_n^2-\sigma^2) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma^4\mathrm{Var}(X^2)/s^4) \iff \sqrt{n}\,\frac{s^2}{\sqrt{\mathrm{Var}(X^2)}}\left(\frac{\hat{\sigma}_n^2}{\sigma^2}-1\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

et on peut à nouveau obtenir des intervalles de confiance asymptotiques.

Comparaison avec les quantiles empiriques

Nous avons vu en Section 2.1.2 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 $(X_1,\dots,X_n)$ i.i.d. selon la loi normale $\mathcal{N}(\theta,1)$, alors par le TCL [5]

$$\sqrt{n}(\bar{X}_n-\theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

tandis qu'en notant $x_{1/2}(n)$ la médiane empirique, on a

$$\sqrt{n}(x_{1/2}(n)-\theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\pi/2).$$

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**.

5. Noter que, dans ce cas particulier, il y a en fait égalité en loi pour tout $n \geq 1$ puisque $\bar{X}_n \sim \mathcal{N}(\theta, 1/n)$.

Rappel ! Revenons sur la médiane empirique dans un cadre général. Comme expliqué précédemment, le résultat de normalité asymptotique

$$\sqrt{n}(x_{1/2}(n) - x_{1/2}) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{1}{4f(x_{1/2})^2}\right)$$

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) :

$$[X_{(\lceil n/2-\sqrt{n}\rceil)}, X_{(\lceil n/2+\sqrt{n}\rceil)}].$$

2.2.2 Le maximum de vraisemblance

On considère 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 $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. selon P_{θ^*}, 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, et ce **pour toute valeur** $\theta \in \Theta$. C'est pour éviter toute confusion que nous notons ici exceptionnellement θ^* la vraie valeur du paramètre.

Définition 23 (Maximum de vraisemblance)
Avec les notations précédentes, un estimateur du maximum de vraisemblance (EMV) est, sous réserve d'existence, une statistique $\hat{\theta} = \hat{\theta}(\mathbf{X}) \in \Theta$ 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 d'un modèle d'échantillonnage où $\mathbf{X} = (X_1, \ldots, X_n)$ avec les X_i i.i.d., autrement dit $g_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \ldots f_\theta(x_n)$, 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 i.i.d., 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 par la suite.

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 24 (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 i.i.d. de loi $\mathcal{N}(\theta_*, 1)$. La log-vraisemblance s'écrit (voir aussi Figure 2.10)

$$\ell_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2.$$

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 $(\bar{X}_n)^2$.

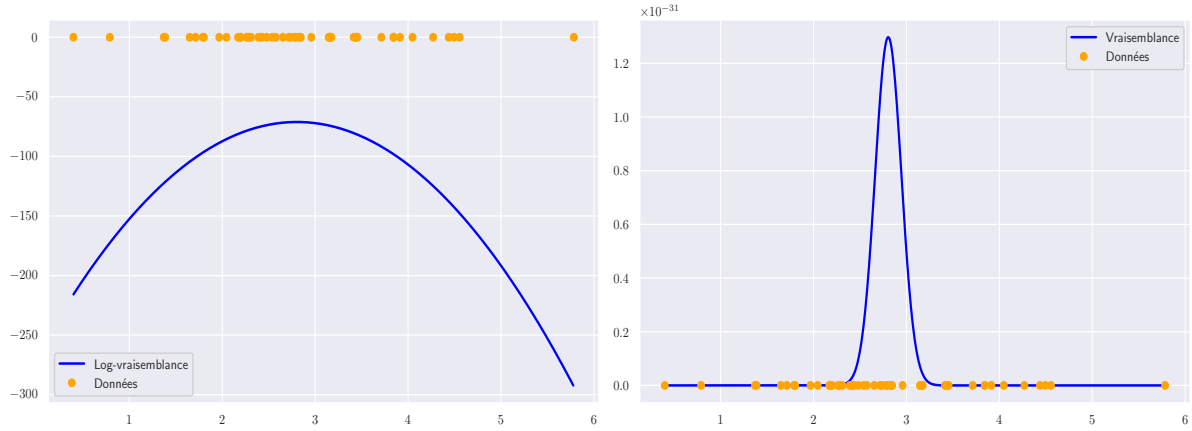

FIGURE 2.10 – Echantillon de 50 variables i.i.d. de loi $\mathcal{N}(3, 1)$, log-vraisemblance et vraisemblance.

Remarque : Pour un modèle $(P_\theta)_{\theta\in\theta}$ dominé, l'EMV dépend de la densité choisie ! Reprenons l'exemple précédent du modèle de translation gaussien, i.e. $X \sim \mathcal{N}(\theta_*, 1)$, mais plutôt que la densité classique $f_\theta(x) = f(x-\theta)$ avec $f(x) = (2\pi)^{-1/2}e^{-x^2/2}$, considérons $g_\theta(x) = g(x-\theta)$ où $g(x) = f(x)\mathbf{1}_{x\neq 1} + \mathbf{1}_{x=1}$. Puisque f et g sont égales presque partout, g est encore une densité par rapport à la mesure de Lebesgue, de loi associée la gaussienne standard, et le modèle de translation défini à partir de cette densité est le même que précédemment. Néanmoins, il est facile de voir que si $n = 1$, c'est-à-dire que l'on dispose d'une seule observation $X \sim \mathcal{N}(\theta_*, 1)$, l'EMV pour les densités g_θ est $\tilde{\theta} = X - 1$ et non plus $\hat{\theta} = X$. Dans la suite, afin d'éviter ce genre de tracas et même si on ne le précisera pas, on considérera toujours les versions "usuelles" des densités.

Nous présentons maintenant quelques exemples illustrant 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.

Loi de Poisson. On passe maintenant à un exemple discret, à savoir l'ensemble des lois de Poisson $(\mathcal{P}(\lambda))_{\lambda>0}$. 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 i.i.d. $(X_1, \ldots, X_n)$ selon $\mathcal{P}(\lambda^*)$ étant donné, sa log-vraisemblance vaut donc, après quelques bidouillages,

$$\ell_n(\lambda) = n(\bar{X}_n \log\lambda - \lambda) - \sum_{i=1}^n \log(X_i!),$$

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.

Remarque : Si l'on s'intéresse aux propriétés asymptotiques de l'EMV, c'est-à-dire consistance et vitesse de convergence, le fait qu'il ne soit pas toujours proprement défini, mais seulement avec une probabilité qui tend vers 1, n'a pas d'importance. En effet, il suffit de suivre le raisonnement de la preuve de la Proposition 6 (c'est-à-dire donner une valeur $\theta_0 \in \Theta$ arbitraire lorsque l'EMV n'est pas défini) pour voir que lesdites propriétés asymptotiques restent valables.

On retient : Souvent, même si la variable aléatoire $\hat{\theta}_n$ maximisant la vraisemblance n'appartient à Θ qu'avec une probabilité qui tend vers 1, nous dirons que $\hat{\theta}_n$ est un EMV de θ^*.

Loi uniforme sur $[0, \theta]$. Le modèle est l'ensemble des lois uniformes $(\mathcal{U}_{[0,\theta]})_{\theta>0}$ et la vraie valeur du paramètre est notée θ_*. La densité de la loi $\mathcal{U}_{[0,\theta]}$ 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 $F_{\hat{\theta}}(t) = \mathbb{P}_{\theta_*}(X_{(n)} \leq t) = (t/\theta_*)^n$ pour tout $t \in [0, \theta_*]$, d'où sa densité et son espérance :

$$f_{\hat{\theta}}(t) = \frac{n}{\theta_*^n}t^{n-1}\mathbb{1}_{[0,\theta_*]}(t) \implies \mathbb{E}_{\theta_*}[\hat{\theta}] = \frac{n}{n+1}\theta_*,$$

ce qui prouve qu'il est biaisé (biais en $\mathcal{O}(1/n)$). Le moment d'ordre 2 permet de calculer le risque quadratique :

$$\mathbb{E}_{\theta_*}[\hat{\theta}^2] = \frac{n}{n+2}\theta_*^2 \implies R(\hat{\theta}, \theta_*) = \mathbb{E}_{\theta_*}\left[(\hat{\theta} - \theta_*)^2\right] = \frac{2\theta_*^2}{(n+1)(n+2)}.$$

Grâce à la fonction de répartition, on note que, pour tout $\alpha \in]0,1[$,

$$\mathbb{P}_{\theta_*}(\hat{\theta} \leq \alpha^{1/n}\theta_*) = \alpha \implies \mathbb{P}_{\theta_*}(\hat{\theta} \leq \theta_* \leq \alpha^{-1/n}\hat{\theta}) = 1-\alpha,$$

ce qui fournit un intervalle de confiance (non asymptotique !) de niveau $(1-\alpha)$.

Puisque $\mathbb{E}_{\theta_*}[\bar{X}_n] = \theta_*/2$, 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, et ce bien que l'EMV soit biaisé, puisque

$$R(\tilde{\theta}, \theta_*) = \mathrm{Var}_{\theta_*}(2\bar{X}_n) = \frac{\theta_*^2}{3n}.$$

Par ailleurs, le calcul de la fonction de répartition montre que, pour tout $t \geq 0$,

$$\mathbb{P}_{\theta_*}\left(n(\theta_* - \hat{\theta}) \geq t\right) = F_{\hat{\theta}}\left(\theta_* - \frac{t}{n}\right) = \left(1 - \frac{t}{\theta_* n}\right)^n \mathbb{1}_{[0,n\theta_*]}(t) \xrightarrow[n\to\infty]{} e^{-\frac{t}{\theta_*}}\mathbb{1}_{[0,\infty[}(t),$$

ce qui prouve que

$$n(\theta_* - \hat{\theta}) \xrightarrow[n\to\infty]{d} \mathcal{E}(1/\theta_*).$$

Ainsi, l'EMV $\hat{\theta}$ converge à vitesse $1/n$ vers θ_* et la loi limite est une loi exponentielle.

Loi uniforme sur $[\theta-1, \theta+1]$. Le modèle est l'ensemble des lois uniformes $(\mathcal{U}_{[\theta-1,\theta+1]})_{\theta\in\mathbb{R}}$ et la vraie valeur du paramètre est à nouveau notée θ_*. 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[6]. 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 $X_{(n)}$ comme en exemple précédent, on montre facilement que $X_{(1)}$ tend en probabilité vers (θ_*-1) et $X_{(n)}$ vers (θ_*+1). Par conséquent, quel que soit le choix de $\hat{\theta}_n$ dans l'intervalle $[X_{(n)}-1, X_{(1)}+1]$, on aura convergence en probabilité 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. On considère le modèle des lois de Cauchy translatées, donc de densités, pour tout paramètre réel θ,

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

et θ_* désigne la vraie valeur de celui-ci. 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 2.11). Bref, on préférera de loin l'estimateur $x_{1/2}(n)$ de la médiane empirique vu en Section 2.1.2, 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 : $x_{1/2}(n) = X_{(\lceil n/2\rceil)}$.

6. Noter que $[X_{(n)}-1, X_{(1)}+1]$ est toujours non vide car $0 < X_{(n)} - X_{(1)} < 2$.

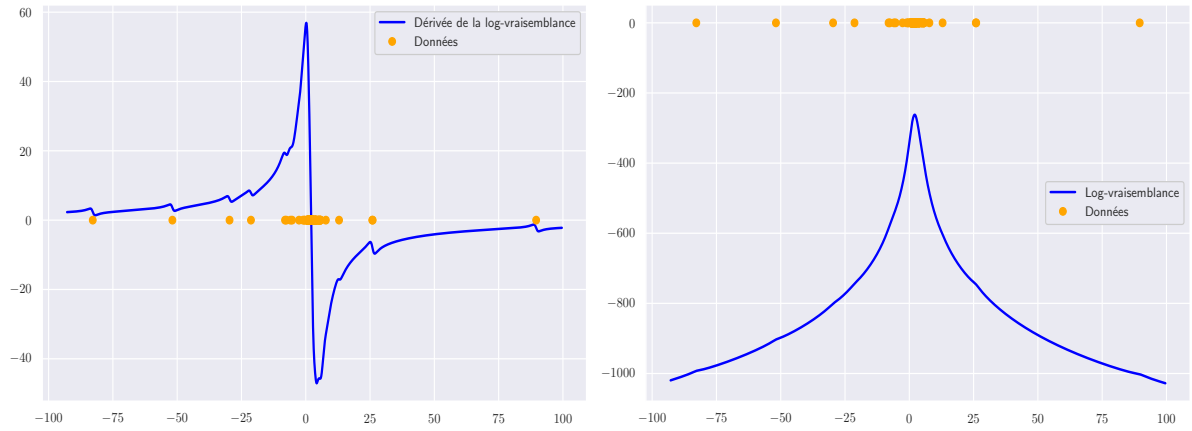

FIGURE 2.11 – 10 variables de Cauchy avec $\theta = 2$, dérivée de la log-vraisemblance et log-vraisemblance.

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 d'être discontinue en 0, où elle explose (mais $f(0) = 0$). On considère alors la famille de densités $(f_\theta)_{\theta\in\mathbb{R}}$ obtenues par translation de f, c'est-à-dire pour tous réels θ et x,

$$f_\theta(x) = f(x-\theta) = \frac{1}{6}\left(\frac{1}{\sqrt{|x-\theta|}}\mathbb{1}_{]0,1]}(|x-\theta|) + \frac{1}{(x-\theta)^2}\mathbb{1}_{]1,+\infty[}(|x-\theta|)\right). \tag{2.7}$$

Pour un n-échantillon $(X_1,\dots,X_n)$ tiré selon la densité f_{θ^*}, 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, mais vaut 0 en chacun des X_i par définition de f. Il n'y a donc pas d'estimateur du maximum de vraisemblance (voir Figure 2.12). On peut également 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 :

$$x_{1/2}(n) \xrightarrow[n\to\infty]{p.s.} \theta^*.$$

En revanche, la normalité asymptotique telle qu'énoncée en Théorème 8 est hors-sujet puisque $f_{\theta^*}(\theta^*) = 0$. 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%.

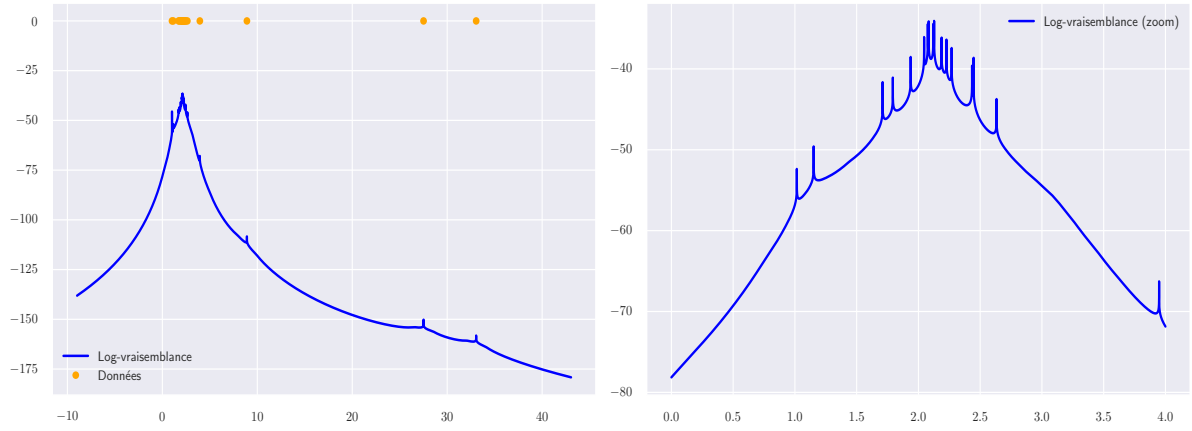

FIGURE 2.12 – Echantillon de 20 variables de loi (2.7) avec $\theta = 2$ et log-vraisemblance "explosive".

2.3 Comparaison d'estimateurs

On reste dans le cadre 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 [7], lequel doit-on choisir ? Plus généralement, existe-t-il un estimateur "optimal", et si oui en quel sens ? Cette section se propose de donner quelques éléments de réponses.

2.3.1 Principes généraux

Comparaison des risques

Comme on l'a vu, une façon de quantifier la qualité d'un estimateur $\hat{\theta} = \hat{\theta}(\mathbf{X})$ de θ est de passer par son risque quadratique, i.e.

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

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

$$\forall \theta \in \Theta \qquad R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta}).$$

Cependant, 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 guère plus avancé. C'est précisément ce qui arrive dans le modèle gaussien déjà croisé où les n variables X_i sont i.i.d. suivant une loi $\mathcal{N}(\theta, 1)$ avec θ paramètre réel inconnu. Considérons les deux estimateurs $\hat{\theta} = \bar{X}_n$ et $\tilde{\theta} = 0$, alors $R(\hat{\theta}, \theta) = 1/n$ et $R(\tilde{\theta}, \theta) = \theta^2$, donc $\hat{\theta}$ est meilleur que $\tilde{\theta}$ si $|\theta| \geq 1/\sqrt{n}$ mais moins bon sinon.

L'approche minimax consiste, pour un estimateur $\hat{\theta}$, à définir son risque maximal $R_{\max}(\hat{\theta}) = \sup_{\theta\in\Theta} R(\theta, \hat{\theta})$, quantité qui ne dépend donc plus de θ, puis à chercher un estimateur $\check{\theta}$ qui minimise

7. Nous revenons ici à la convention selon laquelle θ désigne la vraie valeur du paramètre.

ce risque maximal, c'est-à-dire tel que

$$R_{\max}(\check{\theta}) = \inf_{\hat{\theta}} R_{\max}(\hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} R(\theta, \hat{\theta}),$$

où l'infimum est pris sur tous les estimateurs possibles $\hat{\theta}$ de θ. S'il existe, un tel estimateur $\check{\theta}$ est dit minimax : c'est donc un estimateur optimal dans le pire des cas. Dans l'exemple gaussien ci-dessus, on constate que $R_{\max}(\hat{\theta}) = 1/n$ tandis que $R_{\max}(\tilde{\theta}) = +\infty$, donc au sens du critère minimax le premier estimateur est préférable au second. On peut en fait montrer que, dans ce modèle, $\hat{\theta} = \bar{X}_n$ est un estimateur minimax. De façon plus générale, on peut cependant reprocher à ce critère d'être trop pessimiste, notamment lorsque l'intervalle Θ n'est pas compact.

Le point de vue bayésien revient quant à lui à mettre une loi a priori Π sur le paramètre θ, dès lors vu comme une variable aléatoire $\boldsymbol{\theta}$, et à définir le risque de Bayes

$$R_B(\Pi, \hat{\theta}) = \mathbb{E}\left[(\hat{\theta}(\mathbf{X}) - \boldsymbol{\theta})^2\right] = \int_\Theta \mathbb{E}_\theta\left[(\hat{\theta}(\mathbf{X}) - \theta)^2\right] \Pi(d\theta),$$

où le premier symbole d'espérance signifie qu'on moyennise par rapport à $\mathbf{X}$ et par rapport à $\boldsymbol{\theta}$, tandis que le second considère $\boldsymbol{\theta}$ fixé à la valeur θ (ce n'est rien d'autre que Fubini). A nouveau, l'intérêt est que la quantité $R_B(\Pi, \hat{\theta})$ ne dépend plus de θ. Un estimateur est alors dit de Bayes pour la loi a priori Π et le risque quadratique s'il minimise le risque de Bayes[8]. Contrairement à un estimateur minimax, c'est un estimateur qui est optimal en moyenne, ce qui semble un critère plus raisonnable. Cette solution est attrayante, mais elle dépend tout de même de la loi a priori Π sur $\boldsymbol{\theta}$, laquelle est bien entendu 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

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

et on voit qu'un bon estimateur doit avoir un biais et une variance qui sont **tous deux** petits.

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

$$\lambda = \mathbb{E}_\lambda[\hat{\lambda}(\mathbf{X})] = \sum_{k=0}^{n} \binom{n}{k} \lambda^{-k} \left(1 - \frac{1}{\lambda}\right)^{n-k} \hat{\lambda}(k).$$

Dans cette écriture, les $\hat{\lambda}(k)$ ne sont rien de plus que des coefficients réels dépendant de k mais pas 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.

8. Pour le risque quadratique, on peut montrer que la moyenne a posteriori $\mathbb{E}[\boldsymbol{\theta}|\mathbf{X}]$ est un estimateur de Bayes.

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 préservée par transformation. Ceci est limpide lorsque φ est strictement convexe (ou concave), car l'inégalité de Jensen impose alors [9]

$$\mathbb{E}[\varphi(\hat{\theta})] > \varphi(\mathbb{E}[\hat{\theta}]) = \varphi(\theta),$$

donc l'estimateur $\varphi(\hat{\theta})$ est biaisé, alors que $\hat{\theta}$ ne l'était pas.

L'histoire du débiaisage. Supposons que l'on dispose d'un estimateur biaisé mais que ce biais soit 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 en Section 2.2.2, où θ désigne cette fois la vraie valeur du paramètre. L'estimateur du maximum de vraisemblance est $\hat{\theta} = X_{(n)}$, qui présente un biais puisque $\mathbb{E}[\hat{\theta}] = (n\theta)/(n+1)$. Par ailleurs nous avons vu que

$$\mathbb{E}[\hat{\theta}^2] = \frac{n}{n+2}\theta^2 \implies R(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \frac{2\theta^2}{(n+1)(n+2)}.$$

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 $\check{\theta} := (n+2)X_{(n)}/(n+1)$ est donc (un peu) meilleur que l'estimateur non biaisé $\tilde{\theta}$:

$$R(\check{\theta}, \theta) = \frac{\theta^2}{(n+1)^2} < \frac{\theta^2}{n(n+2)} = R(\tilde{\theta}, \theta).$$

Biais et parallélisation. Plaçons-nous du point de vue du risque quadratique. Très souvent, les estimateurs que l'on considère sont ou bien non biaisés ou bien biaisés en $\mathcal{O}(1/n)$. Leur variance étant typiquement [10] 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

$$b_n(\theta) = \mathbb{E}[\hat{\theta}_n] - \theta = \mathcal{O}(1/n),$$

9. Si $\hat{\theta}$ n'est pas constant, mais cette situation serait sans intérêt.
10. Mais pas toujours, par exemple l'EMV $X_{(n)}$ pour la loi $\mathcal{U}_{[0,\theta]}$ ne rentre pas dans ce cadre, bref passons.

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 i.i.d., 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 $\tilde{T}_n$ vérifiant

$$\mathbb{E}[\tilde{T}_n] = 0 \text{ et } \mathrm{Var}(\tilde{T}_n) = \frac{s_N^2(\theta)}{N} \implies R(\hat{T}_n, \theta) = \frac{s_N^2(\theta)}{N}.$$

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 !

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, pour tout $\theta \in \Theta$, on a $R(\hat{\theta}_n, \theta) = o(R(\tilde{\theta}_n, \theta))$ lorsque $n \to \infty$, on préférera $\hat{\theta}_n$ à $\tilde{\theta}_n$.

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, pour tout $\theta > 0$,

$$R(\hat{\theta}_n, \theta) = \frac{2\theta^2}{(n+1)(n+2)} = o\left(\frac{\theta^2}{3n}\right),$$

on choisira l'EMV, et ce malgré son biais.

Ce dernier exemple n'est cependant pas représentatif de la situation typique : en général, les risques quadratiques convergent à vitesse $1/n$ vers 0. Plus précisément, si l'on dispose pour les estimateurs $\hat{\theta}_n$ et $\tilde{\theta}_n$ de résultats de normalité asymptotique de la forme

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2(\theta)) \qquad \text{et} \qquad \sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, s^2(\theta)),$$

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 [11].

Rappel : 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 (méthode Delta), mais de risque quadratique infini puisque $\mathbb{P}(\bar{X}_n = 0) > 0$.

2.3.2 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.

Nous commençons par rappeler la notion d'absolue continuité d'une fonction. Celle-ci est bien entendu liée à l'absolue continuité d'une mesure par rapport à une autre, vue au Chapitre 1. Pour plus de détails sur ce thème, on pourra consulter [12], Chapitre VI, paragraphe 4, ou [2], Chapitre 6, Section 31.

La question initiale est la suivante : quand peut-on dire qu'une fonction dérivable presque partout est l'intégrale indéfinie de sa dérivée ? Clairement ce n'est pas toujours vrai, comme le montre la fonction $f(x) = \mathbf{1}_{[0,\infty[}(x)$. Il y a plusieurs caractérisations équivalentes de l'absolue continuité, nous adopterons la suivante.

Définition 25 (Absolue continuité)
On dit qu'une fonction f définie sur un intervalle ouvert I de $\mathbb{R}$ est absolument continue sur I de dérivée f' si pour tout segment $[a, b]$ de I on a

$$f(b) - f(a) = \int_a^b f'(x)dx.$$

Remarque : Ainsi, pour faire le lien avec la Section 1.1.5, il est équivalent de dire que la loi d'une variable aléatoire définit une mesure absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$, ou que la fonction de répartition associée à cette loi est absolument continue au sens de la définition ci-dessus.

En terme de régularité, la notion d'absolue continuité est plus forte que celle de continuité [12] + dérivabilité presque partout, mais plus faible que celle de lipschitziannité. En particulier, toute fonction absolument continue est continue. Une fonction peut même être continue sur I, dérivable presque partout sur I, sans être pour autant absolument continue : L'escalier du diable, ou fonction de Cantor, est un exemple typique de fonction continue et dérivable presque partout sans être pour autant absolument continue.

11. 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).
12. Par le théorème de convergence dominée appliqué à la Définition 25.

Par ailleurs, avec la définition précédente, la fonction f' n'est définie que presque partout. Le résultat suivant précise les choses. On rappelle qu'une fonction f est dérivable au sens usuel en x_0 s'il existe $\ell_0 \in \mathbb{R}$ tel que

$$\frac{f(x_0+h)-f(x_0)}{h} \xrightarrow[h\to 0]{} \ell_0.$$

Théorème 9 (Théorème de dérivation de Lebesgue)
Si f est absolument continue sur I, alors il existe un ensemble $I' \subseteq I$ avec $I \setminus I'$ de mesure de Lebesgue nulle tel que f est dérivable au sens usuel en tout point de I', de dérivée f'.

Dans la suite, pour définir sans ambiguïté la dérivée au sens de l'absolue continuité, nous considérons que f' est la dérivée au sens usuel quand celle-ci existe, et 0 sinon. Avec cette convention, lorsque f est absolument continue et positive sur I, alors $f(x) = 0$ implique $f'(x) = 0$. En effet, ou bien f est dérivable au sens usuel en x, mais alors puisque x est un minimum de $f \geq 0$, nécessairement $f'(x) = 0$. Ou bien f n'est pas dérivable au sens usuel en x, auquel cas par la convention précédente on a encore $f'(x) = 0$. Pour la suite, la conséquence de ceci est l'égalité

$$\forall x \in I \qquad f'(x) = f'(x)\mathbb{1}_{f(x)>0}. \tag{2.8}$$

Si f est absolument continue sur I, alors elle est continue sur I et à variation bornée sur tout segment de I. De plus, si f et g sont absolument continues sur I, alors fg l'est aussi, de dérivée égale à $f'g + fg'$ presque partout.

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** de $\mathbb{R}$ et μ une mesure de référence. Par ailleurs, les symboles de dérivation au sens de l'absolue continuité 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 $\theta \in \Theta$:

$$g'_\theta(\mathbf{x}) = \frac{\partial}{\partial\theta} g_\theta(\mathbf{x}) \qquad \text{et} \qquad g''_\theta(\mathbf{x}) = \frac{\partial^2}{\partial\theta^2} g_\theta(\mathbf{x}).$$

Si l'on note $\ell_\theta(\mathbf{X}) = \log g_\theta(\mathbf{X})$ le logarithme de la densité calculé en $\mathbf{X}$ avec $\mathbf{X} \sim P_\theta$, c'est-à-dire la log-vraisemblance évaluée en la vraie valeur θ du paramètre, on appelle **score** la **variable aléatoire**

$$\ell'_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta} \log g_\theta(\mathbf{X}) = \frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}.$$

Attention ! Il y a ici une subtilité : lorsque $\mathbf{X} = (X_1, \ldots, X_n) \sim P_{\theta^*}$, nous avons précédemment noté $\ell_n(\theta) = \log g_\theta(\mathbf{X})$ la log-vraisemblance de l'échantillon (cf. Section 2.2.2), fonction définie **pour toute valeur** $\theta \in \Theta$ et avons défini l'estimateur du maximum de vraisemblance comme une valeur de θ maximisant cette fonction. A contrario, dans toute la présente section, $\mathbf{X}$ est supposé suivre la loi P_θ. En particulier, lorsque nous parlerons des moments du score $\ell'_\theta(\mathbf{X})$, il faut bien avoir en tête que $\mathbf{X} \sim P_\theta$.

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 26 (Modèle régulier, score 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 $\mathbf{x} \in E$, l'application $\theta \mapsto g_\theta(\mathbf{x})$ est absolument continue sur Θ ;*
- *pour tout $\theta_0 \in \Theta$, il existe $E_0 \subseteq E$ avec $\mu(E \setminus E_0) = 0$ tel que pour tout $\mathbf{x} \in E_0$, l'application $\theta \mapsto g'_\theta(\mathbf{x})$ est continue en θ_0 ;*

— *pour tout $\theta \in \Theta$, le score doit admettre un moment d'ordre 2 et l'application*

$$\theta \mapsto I(\theta) = \mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] = \int_E \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\ \mathbb{1}_{g_\theta(\mathbf{x})>0}\ \mu(d\mathbf{x}) \tag{2.9}$$

doit être 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.

A retenir : si pour μ presque tout $\mathbf{x} \in E$, la fonction $\theta \mapsto g_\theta(\mathbf{x})$ est C^1, alors les deux premiers points sont clairement vérifiés [13]. A contrario, si pour μ presque tout $\mathbf{x} \in E$, la fonction $\theta \mapsto g_\theta(\mathbf{x})$ possède (au moins) une discontinuité, le modèle n'est pas régulier puisque le premier point n'est pas vérifié.

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 donc les deux premiers points sont clairs ;
— pour tout $\theta > 0$,

$$\ell_\theta(\mathbf{X}) = \log\theta - \theta\mathbf{X} \implies \ell'_\theta(\mathbf{X}) = \frac{1}{\theta} - \mathbf{X}.$$

Puisque $\mathbb{E}_\theta[\mathbf{X}] = 1/\theta$ et $\mathrm{Var}_\theta(\mathbf{X}) = 1/\theta^2$, on en déduit que

$$\mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] = \mathbb{E}_\theta\left[\left(\frac{1}{\theta} - \mathbf{X}\right)^2\right] = \mathbb{E}_\theta\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2\right] = \mathrm{Var}_\theta(\mathbf{X}) = 1/\theta^2$$

continue sur $\Theta =]0, +\infty[$. Ainsi le modèle défini par ces lois exponentielles est bien régulier, d'information de Fisher égale à $I(\theta) = 1/\theta^2$.

2. Loi de Bernoulli : soit $\mathbf{X} \sim \mathcal{B}(\theta)$ avec $\theta \in \Theta =]0, 1[$, alors $\mu = \delta_0 + \delta_1$ 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 donc les deux premiers points sont satisfaits ;
— Pour le dernier, on peut écrire pour tout $\mathbf{x} \in \{0, 1\}$

$$g_\theta(\mathbf{x}) = \theta^{\mathbf{x}}(1-\theta)^{1-\mathbf{x}} \implies \ell_\theta(\mathbf{X}) = \mathbf{X}\log\theta + (1-\mathbf{X})\log(1-\theta) \implies \ell'_\theta(\mathbf{X}) = \frac{\mathbf{X} - \theta}{\theta(1-\theta)}$$

et puisqu'une variable de Bernoulli de paramètre θ a pour moyenne θ et pour variance $\theta(1-\theta)$, on en déduit

$$\mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] = \mathbb{E}_\theta\left[\left(\frac{\mathbf{X} - \mathbb{E}_\theta[\mathbf{X}]}{\theta(1-\theta)}\right)^2\right] = \frac{\mathrm{Var}_\theta(\mathbf{X})}{(\theta(1-\theta))^2} = \frac{1}{\theta(1-\theta)}$$

13. Ceux-ci comprennent néanmoins des modèles plus généraux : par exemple, comme nous le verrons plus loin, le modèle de translation pour la loi de Laplace défini par $g_\theta(\mathbf{x}) = \frac{1}{2}\exp(-|\mathbf{x} - \theta|)$ est régulier. Pour le second point, il suffit en effet de prendre $E_0 = \mathbb{R} \setminus \{\theta_0\}$, lequel est bien de mesure de Lebesgue pleine.

continue sur $\Theta =]0,1[$. Par conséquent ce modèle est régulier, d'information de Fisher égale à $I(\theta) = 1/(\theta(1-\theta))$.

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.

4. On pourrait penser que si pour tout $\mathbf{x}$ de E la fonction $\theta \mapsto g_\theta(\mathbf{x})$ est continue et C^1 par morceaux sur Θ, alors les deux premiers points de la Définition 26 sont automatiquement vérifiés. Ce n'est pas le cas, comme le montre l'exemple suivant : soit $0 < \theta < 1$, U une variable de loi uniforme sur $[0,1]$, et $\mathbf{X}$ définie par : $\mathbf{X} = \mathbb{1}_{U \leq \theta/2}$ si $0 < \theta \leq 1/2$ et $\mathbf{X} = \mathbb{1}_{U \leq \theta - 1/4}$ si $1/2 < \theta < 1$. Que $\mathbf{x}$ soit égal à 0 ou 1, la fonction $\theta \mapsto g'_\theta(\mathbf{x})$ n'est pas continue en $\theta = 1/2$ et il est donc impossible de définir ce que serait l'information de Fisher en ce point. Ce modèle n'est donc pas régulier.

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 sur Θ 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 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.

Proposition 10 (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.

Preuve : Fixons $\theta_0 \in \Theta$ et $h > 0$ tel que $[\theta_0, \theta_0 + h] \subset \Theta$. Alors, par absolue continuité de $\theta \mapsto g_\theta(\mathbf{x})$ pour μ presque tout $\mathbf{x}$ de E, on a

$$\mathbb{E}_{\theta_0+h}[T(\mathbf{X})] - \mathbb{E}_{\theta_0}[T(\mathbf{X})] = \int_E T(\mathbf{x})(g_{\theta_0+h}(\mathbf{x}) - g_{\theta_0}(\mathbf{x}))\mu(d\mathbf{x}) = \int_E T(\mathbf{x})\left(\int_{\theta_0}^{\theta_0+h} g'_\theta(\mathbf{x})d\theta\right)\mu(d\mathbf{x}).$$

Pour pouvoir inverser l'ordre d'intégration, il faut commencer par vérifier l'absolue intégrabilité.

La propriété (2.8) et l'inégalité de Cauchy-Schwarz donnent :

$$\begin{aligned}\int_{\theta_0}^{\theta_0+h}\left(\int_E |T(\mathbf{x})g'_\theta(\mathbf{x})|\mu(d\mathbf{x})\right)d\theta &= \int_{\theta_0}^{\theta_0+h}\left(\int_E |T(\mathbf{x})|\sqrt{g_\theta(\mathbf{x})}\frac{|g'_\theta(\mathbf{x})|}{\sqrt{g_\theta(\mathbf{x})}}\mathbb{1}_{g_\theta(\mathbf{x})>0}\mu(d\mathbf{x})\right)d\theta \\ &\leq \int_{\theta_0}^{\theta_0+h}\sqrt{\int_E T(\mathbf{x})^2 g_\theta(\mathbf{x})\mu(d\mathbf{x})\int_E \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\mathbb{1}_{g_\theta(\mathbf{x})>0}\mu(d\mathbf{x})}\;d\theta \\ &\leq \int_{\theta_0}^{\theta_0+h}\sqrt{\mathbb{E}_\theta[T(\mathbf{X})^2]I(\theta)}\;d\theta.\end{aligned}$$

Le modèle étant régulier et la fonction $\theta \mapsto \mathbb{E}_\theta[T(\mathbf{X})^2]$ localement bornée, le terme de droite est fini pour h assez petit et on peut donc appliquer le théorème de Fubini dans l'égalité initiale :

$$\mathbb{E}_{\theta_0+h}[T(\mathbf{X})] - \mathbb{E}_{\theta_0}[T(\mathbf{X})] = \int_{\theta_0}^{\theta_0+h}\left(\int_E T(\mathbf{x})g'_\theta(\mathbf{x})\mu(d\mathbf{x})\right)d\theta.$$

Pour montrer que l'application $\theta \mapsto \mathbb{E}_\theta[T(\mathbf{X})]$ est C^1 avec la dérivée de l'énoncé, il suffit ainsi de prouver que l'application $\theta \mapsto \int_E T(\mathbf{x})g'_\theta(\mathbf{x})\mu(d\mathbf{x})$ est continue en tout θ_0, c'est-à-dire que pour toute suite (θ_n) de limite θ_0, on a bien

$$\int_E T(\mathbf{x})g'_{\theta_n}(\mathbf{x})\mu(d\mathbf{x}) \xrightarrow[n\to\infty]{} \int_E T(\mathbf{x})g'_{\theta_0}(\mathbf{x})\mu(d\mathbf{x}),$$

ou, de façon équivalente, que

$$\int_E T(\mathbf{x})\sqrt{g_{\theta_n}(\mathbf{x})}\frac{g'_{\theta_n}(\mathbf{x})}{\sqrt{g_{\theta_n}(\mathbf{x})}}\mathbb{1}_{g_{\theta_n}(\mathbf{x})>0}\mu(d\mathbf{x}) \xrightarrow[n\to\infty]{} \int_E T(\mathbf{x})\sqrt{g_{\theta_0}(\mathbf{x})}\frac{g'_{\theta_0}(\mathbf{x})}{\sqrt{g_{\theta_0}(\mathbf{x})}}\mathbb{1}_{g_{\theta_0}(\mathbf{x})>0}\mu(d\mathbf{x}).$$

En notant $\varphi_n(\mathbf{x}) = T(\mathbf{x})\sqrt{g_{\theta_n}(\mathbf{x})}$ et $\psi_n(\mathbf{x}) = \frac{g'_{\theta_n}(\mathbf{x})}{\sqrt{g_{\theta_n}(\mathbf{x})}}\mathbb{1}_{g_{\theta_n}(\mathbf{x})>0}$, le but est donc de prouver que

$$\int_E (\varphi_n(\mathbf{x})\psi_n(\mathbf{x}) - \varphi_0(\mathbf{x})\psi_0(\mathbf{x}))\mu(d\mathbf{x}) \xrightarrow[n\to\infty]{} 0.$$

Si $\Delta\varphi_n(\mathbf{x}) = \varphi_n(\mathbf{x}) - \varphi_0(\mathbf{x})$ et $\Delta\psi_n(\mathbf{x}) = \psi_n(\mathbf{x}) - \psi_0(\mathbf{x})$, il vient

$$\begin{aligned}\left|\int_E (\varphi_n(\mathbf{x})\psi_n(\mathbf{x}) - \varphi_0(\mathbf{x})\psi_0(\mathbf{x}))\mu(d\mathbf{x})\right| &= \left|\int_E \varphi_n(\mathbf{x})\Delta\psi_n(\mathbf{x})\mu(d\mathbf{x}) + \int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mu(d\mathbf{x})\right| \\ &\leq \int_E |\varphi_n(\mathbf{x})\Delta\psi_n(\mathbf{x})|\mu(d\mathbf{x}) + \left|\int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mu(d\mathbf{x})\right|.\end{aligned} \tag{2.10}$$

Pour démontrer que le second terme de (2.10) tend vers 0, on adopte un raisonnement de type intégrabilité uniforme, en remarquant que pour tout $a > 0$ on peut écrire :

$$\begin{aligned}\int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mu(d\mathbf{x}) &= \int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mathbb{1}_{|\Delta\varphi_n(\mathbf{x})|\leq a|\psi_0(\mathbf{x})|}\mu(d\mathbf{x}) \\ &\quad + \int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mathbb{1}_{|\Delta\varphi_n(\mathbf{x})|>a|\psi_0(\mathbf{x})|}\mu(d\mathbf{x}).\end{aligned} \tag{2.11}$$

Concernant le premier terme de (2.11), puisque pour μ presque tout $\mathbf{x} \in E$, l'application $\theta \mapsto g_\theta(\mathbf{x})$ est absolument continue, elle est en particulier continue en θ_0, donc la fonction sous l'intégrale tend vers 0 pour μ presque tout $\mathbf{x}$. Elle est de plus majorée en valeur absolue par la fonction $\mathbf{x} \mapsto a\psi_0(\mathbf{x})^2$,

laquelle est intégrable par rapport à μ, d'intégrale $aI(\theta_0)$. Le théorème de convergence dominée assure donc que

$$\int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mathbb{1}_{|\Delta\varphi_n(\mathbf{x})|\leq a|\psi_0(\mathbf{x})|}\mu(d\mathbf{x}) \xrightarrow[n\to\infty]{} 0.$$

Pour le second terme de (2.11), via l'inégalité classique $(u-v)^2 \leq 2u^2+2v^2$, on a

$$\left|\int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mathbb{1}_{|\Delta\varphi_n(\mathbf{x})|>a|\psi_0(\mathbf{x})|}\mu(d\mathbf{x})\right| \leq \frac{1}{a}\int_E \Delta\varphi_n(\mathbf{x})^2\mu(d\mathbf{x}) \leq \frac{2}{a}\left(\mathbb{E}_{\theta_n}[T(\mathbf{X})^2]+\mathbb{E}_{\theta_0}[T(\mathbf{X})^2]\right).$$

Puisque la fonction $\theta \mapsto \mathbb{E}_\theta[T(\mathbf{X})^2]$ est localement bornée, elle est bornée au voisinage de θ_0 et il existe c indépendant de a tel que $\limsup_{n\to\infty}\mathbb{E}_{\theta_n}[T(\mathbf{X})^2] \leq c$. Cette borne étant également valide en remplaçant θ_n par θ_0, il vient

$$\limsup_{n\to\infty}\left|\int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mathbb{1}_{|\Delta\varphi_n(\mathbf{x})|>a|\psi_0(\mathbf{x})|}\mu(d\mathbf{x})\right| \leq \frac{2c}{a}.$$

Puisque a peut être choisi arbitrairement, on a bien établi que

$$\int_E \Delta\varphi_n(\mathbf{x})\psi_0(\mathbf{x})\mathbb{1}_{|\Delta\varphi_n(\mathbf{x})>a|\psi_0(\mathbf{x})|}\mu(d\mathbf{x}) \xrightarrow[n\to\infty]{} 0.$$

Au total, nous venons de prouver que le second terme de (2.10) tend vers 0. Pour le premier terme, l'inégalité de Cauchy-Schwarz donne

$$\int_E |\varphi_n(\mathbf{x})\Delta\psi_n(\mathbf{x})|\mu(d\mathbf{x}) \leq \sqrt{\mathbb{E}_{\theta_n}[T(\mathbf{X})^2]} \times \sqrt{\int_E \Delta\psi_n(\mathbf{x})^2\mu(d\mathbf{x})}.$$

Puisque $\limsup_{n\to\infty}\mathbb{E}_{\theta_n}[T(\mathbf{X})^2] \leq c$, la preuve sera complète une fois établi que le terme de droite tend vers 0. Pour ce faire, on écrit

$$\int_E \Delta\psi_n(\mathbf{x})^2\mu(d\mathbf{x}) = I(\theta_n) - I(\theta_0) - 2\int_E \psi_0(\mathbf{x})\Delta\psi_n(\mathbf{x})\mu(d\mathbf{x}).$$

Puisque l'information de Fisher est continue, $I(\theta_n)$ tend vers $I(\theta_0)$ et il suffit donc de prouver que le dernier terme tend vers 0. Une façon de procéder consiste à considérer la décomposition (2.11) en remplaçant $\Delta\varphi_n$ par $\Delta\psi_n$ et à voir que, mutatis mutandis, les arguments précédents passent encore. En particulier, le théorème de convergence dominée s'applique à nouveau en remarquant que

$$\int_E \psi_0(\mathbf{x})\Delta\psi_n(\mathbf{x})\mu(d\mathbf{x}) = \int_E \psi_0(\mathbf{x})\left(\psi_n(\mathbf{x})\mathbb{1}_{g_{\theta_0}(\mathbf{x})>0} - \psi_0(\mathbf{x})\right)\mu(d\mathbf{x}).$$

Le modèle étant régulier, il existe un ensemble E_0 de μ mesure pleine tel que pour tout $\mathbf{x} \in E_0$,

$$\psi_n(\mathbf{x})\mathbb{1}_{g_{\theta_0}(\mathbf{x})>0} = \frac{g'_{\theta_n}(\mathbf{x})}{\sqrt{g_{\theta_n}(\mathbf{x})}}\mathbb{1}_{g_{\theta_n}(\mathbf{x})>0}\mathbb{1}_{g_{\theta_0}(\mathbf{x})>0} \xrightarrow[n\to\infty]{} \frac{g'_{\theta_0}(\mathbf{x})}{\sqrt{g_{\theta_0}(\mathbf{x})}}\mathbb{1}_{g_{\theta_0}(\mathbf{x})>0} = \psi_0(\mathbf{x}).$$

■

Dans la Proposition 10, le cas particulier $T(\mathbf{X}) = 1$ assure que le score est centré, c'est-à-dire $\mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})\right] = 0$. Ceci donne une nouvelle formule pour l'information de Fisher, que nous avons en fait déjà rencontrée sur les modèles des lois exponentielles et de Bernoulli.

Corollaire 3 (Information de Fisher et variance du score)
Si le modèle est régulier, alors

$$I(\theta) = \mathrm{Var}_\theta(\ell'_\theta(\mathbf{X})),$$

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 10, alors $\theta \mapsto \mathbb{E}_\theta[T(\mathbf{X})^2] = 1$ est bien localement bornée, donc

$$0 = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[1] = \mathbb{E}_\theta\left[\frac{g'_\theta(\mathbf{X})}{g_\theta(\mathbf{X})}\right] = \mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})\right].$$

D'où l'on déduit, en partant de l'équation (2.9),

$$I(\theta) = \mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] = \mathbb{E}_\theta\left[(\ell'_\theta(\mathbf{X}))^2\right] - \left(\mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})\right]\right)^2 = \mathrm{Var}_\theta(\ell'_\theta(\mathbf{X})).$$

■

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

$$\forall\theta_0 \in \Theta,\ \exists\varepsilon = \varepsilon(\theta_0) > 0 \qquad \sup_{\theta_0-\varepsilon<\theta<\theta_0+\varepsilon} |\varphi_\theta(\mathbf{x})| \in L_1(\mu).$$

Si l'on considère pour μ la mesure de Lebesgue sur $\mathbb{R}$, un exemple de famille non localement dominée dans $L_1(\mu)$ est donné par $\varphi_\theta(\mathbf{x}) = \exp(-|\theta\mathbf{x}|)$ si $\theta \neq 0$ et $\varphi_0(\mathbf{x}) = 0$ lorsque $\theta = 0$. Toutes les fonctions $\mathbf{x} \mapsto \varphi_\theta(\mathbf{x})$ sont intégrables sur $\mathbb{R}$, mais si l'on prend $\theta_0 = 0$, il est clair que pour tout réel $\mathbf{x}$ et tout $\varepsilon > 0$, $\sup_{-\varepsilon<\theta<\varepsilon} |\varphi_\theta(\mathbf{x})| = 1$, qui n'est pas intégrable sur $\mathbb{R}$.

Quoi qu'il en soit, 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[14].

Lemme 4 (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 26.

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 26, 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.

■

14. On notera cependant qu'elle est plus restrictive et pas plus simple à vérifier que celle de la Définition 26.

En ajoutant une hypothèse du même tonneau, on aboutit à une nouvelle expression pour l'information de Fisher.

Proposition 11 (Information de Fisher et dérivée seconde)
Conservons les hypothèses du Lemme 4 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 (2.9) 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 $g''_{\theta_0}(\mathbf{x})$. 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

$$\Phi'(\theta_0) = \int_{E'} g''_{\theta_0}(\mathbf{x})\mu(d\mathbf{x}) = \int_{E'} \frac{g''_{\theta_0}(\mathbf{x})}{g_{\theta_0}(\mathbf{x})} g_{\theta_0}(\mathbf{x})\mu(d\mathbf{x}) = \mathbb{E}_{\theta_0} \left[\frac{g''_{\theta_0}(\mathbf{X})}{g_{\theta_0}(\mathbf{X})}\right].$$

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 3, Φ est identiquement nulle sur Θ, donc il en va de même pour sa dérivée.

■

Exemple : Illustrons ce résultat sur l'exemple des lois exponentielles. Quel que soit $\theta > 0$, le support est $[0, +\infty[$ donc indépendant de θ. Par ailleurs, on a vu que pour tout $x \geq 0$, l'application $\theta \mapsto g_\theta(\mathbf{x})$ est C^∞ sur $\Theta =]0, +\infty[$. Pour tout $\theta_0 > 0$ et $\varepsilon > 0$ tel que $\theta_0 - \varepsilon > 0$, on a pour tout $x \geq 0$:

$$\frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})} = \frac{(1-\theta\mathbf{x})^2}{\theta} e^{-\theta\mathbf{x}} \implies 0 \leq \sup_{\theta_0-\varepsilon<\theta<\theta_0+\varepsilon} \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})} \leq \psi(\mathbf{x}) = \frac{(1+(\theta_0+\varepsilon)\mathbf{x})^2}{\theta_0-\varepsilon} e^{-(\theta_0-\varepsilon)\mathbf{x}},$$

avec clairement

$$\int_0^{+\infty} \psi(\mathbf{x})\ dx < +\infty.$$

De la même façon,

$$g''_\theta(\mathbf{x}) = (\theta\mathbf{x} - 2)\mathbf{x}e^{-\theta\mathbf{x}} \implies \sup_{\theta_0-\varepsilon<\theta<\theta_0+\varepsilon} |g''_\theta(\mathbf{x})| \leq \phi(\mathbf{x}) = ((\theta_0+\varepsilon)\mathbf{x} + 2)\mathbf{x}e^{-(\theta_0-\varepsilon)\mathbf{x}},$$

avec clairement

$$\int_0^{+\infty} \phi(\mathbf{x})\ dx < +\infty.$$

Le modèle est donc régulier au sens de Fisher et on peut appliquer la formule de la Proposition 11 pour retrouver l'information de Fisher :

$$\ell'_\theta(\mathbf{x}) = \frac{1}{\theta} - \mathbf{x} \implies \ell''_\theta(\mathbf{x}) = -\frac{1}{\theta^2} \implies I(\theta) = -\mathbb{E}_\theta\left[\ell''_\theta(\mathbf{X})\right] = \frac{1}{\theta^2}.$$

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 5 (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 12 (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 on peut montrer que pour μ presque tout $\mathbf{x}$, la fonction $\eta \mapsto h_\eta(\mathbf{x})$ est absolument 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 ceci 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 i.i.d., l'information de Fisher croît linéairement avec la taille de l'échantillon. En d'autres termes, l'information apportée par n observations i.i.d. est n fois plus grande que l'information apportée par une seule.

Proposition 13 (Information de Fisher d'un échantillon)
Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon i.i.d., 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

$$\theta \mapsto k_\theta(\mathbf{x}, \mathbf{y}) = g_\theta(\mathbf{x}) h_\theta(\mathbf{y}),$$

c'est-à-dire la vérification des deux premiers points de la Définition 26, se déduit de celles de $\theta \mapsto g_\theta(\mathbf{x})$ et $\theta \mapsto h_\theta(\mathbf{y})$. Le produit de deux fonctions absolument continues étant lui-même absolument continu, on en conclut que 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

$$I(\theta) = \iint \frac{k'_\theta(\mathbf{x},\mathbf{y})^2}{k_\theta(\mathbf{x},\mathbf{y})} \mu(d\mathbf{x})\nu(d\mathbf{y})$$

est la somme de trois termes, le premier et le dernier étant comparables. Le premier s'écrit (l'intégration se faisant sur S_θ)

$$\iint \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})} h_\theta(\mathbf{y})\mu(d\mathbf{x})\nu(d\mathbf{y}) = \left(\int \frac{g'_\theta(\mathbf{x})^2}{g_\theta(\mathbf{x})}\mu(d\mathbf{x})\right)\left(\int h_\theta(\mathbf{y})\nu(d\mathbf{y})\right) = I_1(\theta),$$

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

$$\iint g'_\theta(\mathbf{x})h'_\theta(\mathbf{y})\mu(d\mathbf{x})\nu(d\mathbf{y}) = \left(\int g'_\theta(\mathbf{x})\mu(d\mathbf{x})\right)\left(\int h'_\theta(\mathbf{y})\nu(d\mathbf{y})\right) = 0,$$

ces deux intégrales étant nulles via la Proposition 10 : 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 13 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 :

$$g_\theta(\mathbf{X}) = \prod_{i=1}^n f_\theta(X_i) \Longrightarrow \ell_\theta(\mathbf{X}) = \sum_{i=1}^n \ell_\theta(X_i) \Longrightarrow I_n(\theta) = \mathrm{Var}_\theta(\ell'_\theta(\mathbf{X})) = \sum_{i=1}^n \mathrm{Var}_\theta(\ell'_\theta(X_i)) = nI_1(\theta).$$

Exemples

La Proposition 13 nous dit que l'information de Fisher d'un échantillon i.i.d. 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. Commençons par quelques lois classiques.

1. Loi binomiale : si $X \sim \mathcal{B}(n,\theta)$ avec $0 < \theta < 1$ inconnu et $n \in \mathbb{N}^\star$ connu, alors ce modèle est régulier pour les mêmes raisons que le modèle de Bernoulli. Cette fois, pour tout $x \in \{0,\dots,n\}$, on a

$$f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \implies \ell'_\theta(x) = \frac{x-n\theta}{\theta(1-\theta)} \implies I(\theta) = \mathrm{Var}(\ell'_\theta(X)) = \frac{n}{\theta(1-\theta)}.$$

   On note que cette information est égale à n fois celle du modèle de Bernoulli. Sans rentrer dans les détails : on sait qu'une variable binomiale $\mathcal{B}(n,\theta)$ correspond en loi à la somme de n variables i.i.d. $X_1,\dots,X_n$ de Bernoulli $\mathcal{B}(\theta)$, or on peut montrer que cette somme est une statistique exhaustive du vecteur $(X_1,\dots,X_n)$, c'est-à-dire grosso modo que la somme est un résumé sans perte de toute l'information sur le paramètre θ contenue dans le vecteur. Or la Proposition 13 nous assure justement que l'information de Fisher du modèle à n variables est égale à n fois l'information du modèle à 1 variable, laquelle vaut comme on l'a vu sur le modèle de Bernoulli $I_1(\theta) = 1/(\theta(1-\theta))$.

2. Loi de Poisson : si $X \sim \mathcal{P}(\lambda)$, avec $\lambda > 0$ paramètre inconnu, la vraisemblance vaut, pour tout $\lambda > 0$ et tout $x \in \mathbb{N}$,

$$\ell_\lambda(x) = \log \mathbb{P}_\lambda(X = x) = -\lambda + x \log \lambda - \log(x!).$$

Pour tout $x \in \mathbb{N}$, la fonction $\lambda \mapsto \ell_\lambda(x)$ est C^1 sur $]0, \infty[$ donc les deux premiers points de la Définition 26 sont satisfaits. Il reste à vérifier que le moment d'ordre 2 du score $\ell'_\lambda(X) = X/\lambda - 1$ est une fonction continue en λ. Rappelons qu'une variable de Poisson $\mathcal{P}(\lambda)$ a pour moyenne et pour variance λ, donc

$$\mathbb{E}_\lambda[\ell'_\lambda(X)^2] = \mathbb{E}_\lambda[(X/\lambda - 1)^2] = \frac{1}{\lambda^2}\mathbb{E}_\lambda[(X-\lambda)^2] = \frac{1}{\lambda^2}\mathrm{Var}_\lambda(X) = \frac{1}{\lambda}.$$

Puisque $\lambda \mapsto \frac{1}{\lambda}$ est continue sur $]0, \infty[$, ce modèle est régulier d'information $I(\lambda) = \frac{1}{\lambda}$.

3. 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 $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $Y = (X-\mu)/\sigma \sim \mathcal{N}(0,1)$, avec $\mathbb{E}[Y^2] = 1$ et $\mathrm{Var}[Y^2] = 2$.

— Si le paramètre est $\mu \in \mathbb{R}$ (i.e. σ^2 connu) : pour tout réel x, la fonction $\mu \mapsto \ell_\mu(x) = \log f(x)$ est C^1 sur $\mathbb{R}$. De plus

$$\mathbb{E}_\mu[\ell'_\mu(X)^2] = \frac{1}{\sigma^2}\mathbb{E}_\mu[((X-\mu)/\sigma)^2] = \frac{1}{\sigma^2}\mathbb{E}_\mu[Y^2] = \frac{1}{\sigma^2}.$$

Le modèle $(f_\mu)_{\mu\in\mathbb{R}}$ est donc régulier, d'information de Fisher constante $I(\mu) = 1/\sigma^2$.

— Si le paramètre est $\sigma^2 > 0$ (i.e. μ connu) : pour tout réel x, la fonction $\sigma^2 \mapsto \ell_{\sigma^2}(x) = \log f(x)$ est C^1 sur $]0, \infty[$. De plus

$$\mathbb{E}_{\sigma^2}[\ell'_{\sigma^2}(X)^2] = \frac{1}{4\sigma^4}\mathbb{E}_{\sigma^2}\left[\left\{((X-\mu)/\sigma)^2 - 1\right\}^2\right] = \frac{1}{4\sigma^4}\mathrm{Var}_{\sigma^2}(Y^2) = \frac{1}{2\sigma^4}.$$

Le modèle $(f_{\sigma^2})_{\sigma^2>0}$ est donc régulier, d'information de Fisher $I(\sigma^2) = 1/(2\sigma^4)$. Noter que si on considère $\sigma > 0$ comme paramètre, alors la Proposition 12 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. Ceci est en fait vrai pour tous les modèles de translation réguliers, comme nous allons le voir maintenant.

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. Rappelons qu'une fonction définie sur un segment $[a, b]$ est dite continue et C^1 par morceaux si elle est continue et s'il

existe une subdivision $a_0 = a < a_1 < \cdots < a_n = b$ telle que chaque restriction de f à $]a_i, a_{i+1}[$ se prolonge en une fonction de classe C^1 sur $[a_i, a_{i+1}]$ [15]. Une fonction définie sur $\mathbb{R}$ est dite continue et C^1 par morceaux si elle l'est sur tout segment contenu dans cet intervalle. Ainsi, l'ensemble des points où f n'est pas dérivable est au plus dénombrable, donc de mesure de Lebesgue nulle. Il est facile de voir qu'une telle fonction est absolument continue.

Proposition 14 (Régularité d'un modèle de translation)
Si la densité f est continue sur $\mathbb{R}$ et C^1 par morceaux, 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. En notant $\mathcal{D}$ l'ensemble au plus dénombrable de points où f n'est pas dérivable, les deux premiers points de la Définition 26 se vérifient facilement :

- pour tout x, la fonction $\theta \mapsto f_\theta(x) = f(x-\theta)$ étant continue et C^1 par morceaux, elle est absolument continue sur Θ ;
- pour tout θ_0, notons $N_0 = \theta_0 + \mathcal{D}$, alors N_0 est négligeable pour la mesure de Lebesgue et pour tout $x \in E_0 = \mathbb{R} \setminus N_0$, la fonction $\theta \mapsto f'_\theta(x) = -f'(x-\theta)$ est continue au point θ_0.

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 26 que nous avons adoptée et celle, plus classique mais plus restrictive, du Lemme 4.

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 2.13) :

$$\forall x \neq 0 \qquad f'(x)^2 = \frac{1}{4}e^{-2|x|} \implies \frac{f'(x)^2}{f(x)} = \frac{1}{2}e^{-|x|} \implies \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)} dx = 1.$$

   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 4 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 2.13)

$$f(x) = \frac{1+\cos x}{2\pi}\mathbb{1}_{[-\pi,\pi]}(x) \implies f'(x) = \frac{-\sin x}{2\pi}\mathbb{1}_{[-\pi,\pi]}(x) \implies \frac{f'(x)^2}{f(x)} = \frac{1-\cos x}{2\pi}\mathbb{1}_{[-\pi,\pi]}(x)$$

15. La fonction $f : [-1,1] \to \mathbb{R}$ définie par $f(x) = x^2 \sin(1/x)\mathbf{1}_{x\neq 0}$ est un exemple de fonction dérivable sur $[-1,1]$ mais non C^1 par morceaux car $f'(x)$ n'admet pas de limite à droite en 0 (ni à gauche du reste). Elle est cependant absolument continue puisqu'elle est 3-lipschitzienne.

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 4, 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 $f_\theta(x) = \mathbb{1}_{[0,1]}(x-\theta)$, on voit que, pour tout réel x, la fonction $\theta \mapsto f_\theta(x)$ présente deux discontinuités, en $x-1$ et en x. Le premier point de la Définition 26 n'est pas vérifié et ce modèle de translation 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.

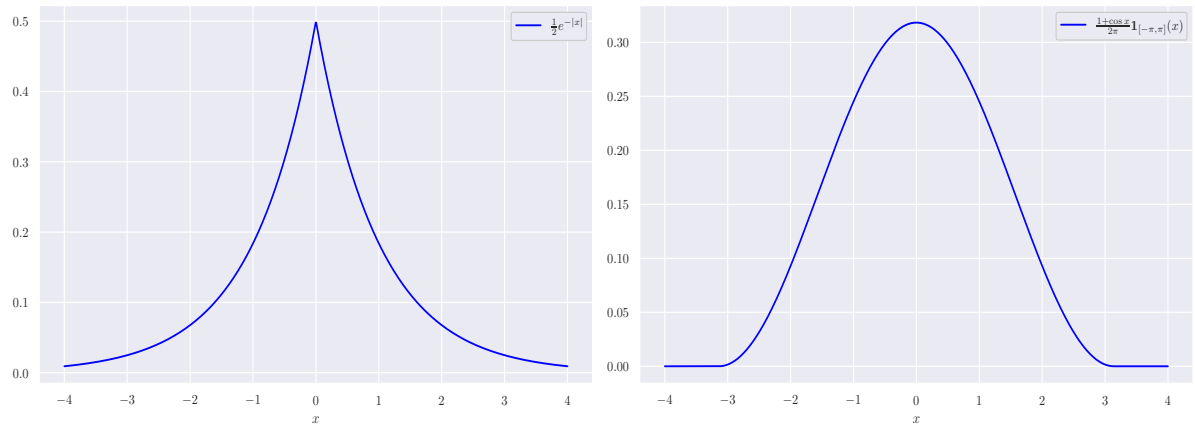

FIGURE 2.13 – Loi de Laplace (à gauche) et loi "exotique" (à droite).

2.3.3 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 15 (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 de classe C^1, 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 : Puisque $(a+b)^2 \leq 2(a^2+b^2)$ pour tous réels a et b, il vient

$$\hat{\theta}(\mathbf{X})^2 \leq 2\left\{(\hat{\theta}(\mathbf{X})-\theta)^2+\theta^2\right\} \implies \mathbb{E}_\theta[\hat{\theta}(\mathbf{X})^2] \leq 2\left\{\mathbb{E}_\theta\left[(\hat{\theta}(\mathbf{X})-\theta)^2\right]+\theta^2\right\}.$$

Les deux membres de droite étant localement bornés, il en va de même pour celui de gauche. On peut donc appliquer la Proposition 10 à la statistique $\hat{\theta}(\mathbf{X})$, ce qui assure que la fonction $\theta \mapsto \mathbb{E}_\theta[\hat{\theta}(\mathbf{X})]$ est de classe C^1 sur Θ, de dérivée

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\theta}(\mathbf{X})] = \mathbb{E}_\theta\left[\hat{\theta}(\mathbf{X})\ell'_\theta(\mathbf{X})\right].$$

Or on sait que le score est centré, i.e. $\mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})\right] = 0$, donc l'équation précédente s'écrit encore

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\theta}(\mathbf{X})] = \mathbb{E}_\theta\left[(\hat{\theta}(\mathbf{X}) - \mathbb{E}_\theta[\hat{\theta}(\mathbf{X})])\ell'_\theta(\mathbf{X})\right].$$

L'inégalité de Cauchy-Schwarz donne alors

$$\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\theta}(\mathbf{X})]\right)^2 \leq \mathrm{Var}_\theta(\hat{\theta}(\mathbf{X})) \times I(\theta). \tag{2.12}$$

Il reste à voir que, pour le membre de gauche, $\mathbb{E}_\theta[\hat{\theta}(\mathbf{X})] = b(\theta) + \theta$. La fonction $\theta \mapsto \mathbb{E}_\theta[\hat{\theta}(\mathbf{X})]$ étant de classe C^1, le biais l'est aussi et

$$\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\theta}(\mathbf{X})]\right)^2 = (1 + b'(\theta))^2.$$

On peut de plus appliquer au membre de droite la décomposition classique du risque quadratique :

$$\mathrm{Var}_\theta(\hat{\theta}(\mathbf{X})) = \mathbb{E}_\theta\left[\left(\hat{\theta}(\mathbf{X}) - \theta\right)^2\right] - b(\theta)^2.$$

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 (2.12). On voit que si $I(\theta_0) = 0$, tout s'écroule et on perd toute information sur la variance de $\hat{\theta}(\mathbf{X})$ en θ_0.

Donnons maintenant la version la plus connue de l'inégalité précédente : elle est due à Fréchet, Darmois, Cramér et Rao, mais l'usage n'a conservé que les deux derniers auteurs.

Corollaire 4 (Borne de Cramér-Rao)
Si $\hat{\theta}(\mathbf{X})$ un estimateur sans biais de θ dont la variance est localement bornée 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

$$(X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

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 2.2.2)

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

Il est clairement non biaisé et de variance [16]

$$\mathrm{Var}(\hat{\sigma}^2) = \frac{\mathrm{Var}(X_1^2)}{n} = \frac{\mathbb{E}[X_1^4] - (\mathbb{E}[X_1^2])^2}{n} = \frac{2\sigma^4}{n},$$

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 2.3.1, les estimateurs sans biais, lorsqu'ils existent, ne sont pas nécessairement les plus intéressants en terme d'erreur quadratique. La seconde vient de ce qu'un estimateur efficace ne peut exister que dans des conditions très particulières et clairement identifiées (estimateur "linéaire" dans un modèle exponentiel). En fait, la plupart des problèmes d'estimation n'admettent pas d'estimateur efficace.

Exemple : Reprenons le cas des lois exponentielles $(\mathcal{E}(\lambda))_{\lambda>0}$. Le calcul de l'information de Fisher a déjà été fait : $I_1(\lambda) = 1/\lambda^2$. Lorsque

$$\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \mathcal{E}(\lambda),$$

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

$$\hat{\lambda}_n = \hat{\lambda}_n(\mathbf{X}) = \frac{n-1}{n\bar{X}_n}.$$

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.

16. Rappelons que si $X \sim \mathcal{N}(0,1)$, alors $\mathbb{E}[X^4] = 3$, cas particulier de la formule générale : $\mathbb{E}[X^{2n}] = (2n)!/(2^n n!)$.

Remarque : Avant de passer à l'efficacité asymptotique, revenons aux lois exponentielles, que nous définissons cette fois pour tout $\theta > 0$ par [17]

$$f_\theta(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \mathbb{1}_{x \geq 0}.$$

A partir d'un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. 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.

2.3.4 Efficacité asymptotique

La borne inférieure donnée par l'Inégalité de l'Information vue en Proposition 15 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.

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 10 (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 $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. selon f_θ, alors pour tout estimateur $\hat{\theta}_n(\mathbf{X})$, on a

$$\sup_{\theta \in J} R(\hat{\theta}_n(\mathbf{X}), \theta) \geq \frac{1}{n\bar{I}} \times \left(\frac{1}{1 + \frac{1}{r\sqrt{n\bar{I}}}}\right)^2.$$

17. C'est d'ailleurs la définition donnée dans beaucoup d'ouvrages et considérée par certains logiciels.

Exemple : Pour l'estimation de la moyenne dans le modèle $(\mathcal{N}(\theta,1))_{\theta\in\mathbb{R}}$, nous avons vu que l'information est constante égale à $I(\theta)=1$, 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 $\hat{\theta}_n(\mathbf{X})$,

$$\sup_{-r\leq\theta\leq r}\mathbb{E}_\theta\left[\left(\hat{\theta}_n(\mathbf{X})-\theta\right)^2\right]\geq\frac{1}{n}\times\left(\frac{1}{1+\frac{1}{r\sqrt{n}}}\right)^2.$$

En particulier, on voit que l'estimateur trivial $\tilde{\theta}(\mathbf{X})=\tilde{\theta}_n(\mathbf{X})=0$ proposé ci-dessus n'est plus du tout optimal puisque

$$\sup_{-r\leq\theta\leq r}\mathbb{E}_\theta\left[\left(\tilde{\theta}_n(\mathbf{X})-\theta\right)^2\right]=r^2,$$

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 (cf. preuve de la Proposition 15), la fonction 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 10 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,

$$\sup_{\theta\in J} \mathbb{E}_\theta \left[\left(\hat{\varphi}_n(\mathbf{X}) - \varphi(\theta)\right)^2 \right] \geq \frac{1}{n\bar{I}_\varphi} \times \left(\frac{1}{1 + \frac{2}{\Delta(\varphi)\sqrt{n\bar{I}_\varphi}}} \right)^2 .$$

Exemple : Revenons au résultat du Théorème 10. 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

$$\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta \left[\left(\hat{\theta}_n(\mathbf{X}) - \theta\right)^2 \right] \geq \frac{1}{nI}.$$

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 !

Inversement, on peut s'intéresser au comportement local de cette inégalité. Pour ce faire, considérons maintenant $J = J_n = [\theta_0 - r_n, \theta_0 + r_n]$ avec (r_n) une suite tendant 0, alors la continuité de la fonction I implique que

$$\bar{I} = \bar{I}_n = \sup_{\theta_0 - r_n \leq \theta \leq \theta_0 + r_n} I(\theta) \xrightarrow[n\to\infty]{} I(\theta_0).$$

Ainsi, pour toute suite (r_n) de limite nulle, on a à la fois $\bar{I}_n$ qui tend vers $I(\theta_0)$ et

$$\sup_{\theta_0 - r_n \leq \theta \leq \theta_0 + r_n} nR(\hat{\theta}_n, \theta) \geq \frac{1}{\bar{I}_n} \times \left(\frac{1}{1 + \frac{1}{r_n\sqrt{n\bar{I}_n}}} \right)^2 .$$

Cette minoration est en particulier vérifiée dans le pire des cas pour le minorant, c'est-à-dire lorsque celui-ci est de limite la plus grande possible : il suffit pour ça de choisir r_n de sorte que $r_n\sqrt{n}$ tende vers l'infini (par exemple $r_n = n^{-1/4}$), ce qui donne

$$\liminf_{n\to\infty} \sup_{\theta_0 - r_n \leq \theta \leq \theta_0 + r_n} nR(\hat{\theta}_n, \theta) \geq \frac{1}{I(\theta_0)}.$$

Autrement dit, le risque d'un estimateur $\hat{\theta}_n$ de θ ne peut être asymptotiquement meilleur que $1/(nI(\theta_0))$ au voisinage de θ_0. Ceci laisse à penser que pour un estimateur $\hat{\theta}_n$ asymptotiquement normal de θ_0, c'est-à-dire tel que

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2(\theta_0)),$$

la plus petite valeur possible pour $\sigma^2(\theta_0)$ serait $1/I(\theta_0)$. En généralisant comme toujours via une fonction φ, la borne serait en $\varphi'(\theta_0)^2/(nI(\theta_0))$. La définition de l'efficacité asymptotique part de ce constat.

Définition 27 (Efficacité asymptotique)
Soit un modèle régulier $(f_\theta)_{\theta\in\Theta}$ d'information de Fisher $I(\theta) = I_1(\theta)$. Un estimateur $\hat{\theta}_n(\mathbf{X})$ de θ est dit asymptotiquement efficace si, lorsque $\mathbf{X} = (X_1, \ldots, X_n)$ est un échantillon i.i.d. selon f_θ, on a

$$\sqrt{n}\left(\hat{\theta}_n(\mathbf{X}) - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2(\theta)) \qquad \text{avec} \qquad \sigma^2(\theta) \leq \frac{1}{I(\theta)}$$

pour tout θ tel que $I(\theta) > 0$.

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 2.3.2, 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. Sous les mêmes hypothèses, on peut généraliser la Définition 27 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. Cet estimateur est dit asymptotiquement efficace si

$$\sqrt{n}\left(\hat{\varphi}_n(\mathbf{X}) - \varphi(\theta)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2(\theta)) \qquad \text{avec} \qquad \sigma^2(\theta) \leq \frac{\varphi'(\theta)^2}{I(\theta)}$$

   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 méthode Delta assure que $\hat{\varphi}_n(\mathbf{X}) = \varphi(\hat{\theta}_n(\mathbf{X}))$ est un estimateur asymptotiquement efficace de $\varphi(\theta)$.

3. **Estimateur de Hodges** : on considère le modèle de translation gaussien $\mathcal{N}(\theta, 1)$, d'information de Fisher constante $I(\theta) = 1$. Si $\mathbf{X} = (X_1, \ldots, X_n)$ est un échantillon i.i.d. selon cette loi, alors par les propriétés classiques des variables gaussiennes, l'estimateur $\bar{X}_n$ vérifie pour tout n

$$\sqrt{n}\left(\bar{X}_n - \theta\right) \sim \mathcal{N}(0, 1) \qquad \text{avec} \qquad 1 = \frac{1}{I(\theta)},$$

   donc c'est un estimateur asymptotiquement efficace. Etant donné que $\mathbb{E}_\theta[\bar{X}_n] = \theta$ et $\mathrm{Var}_\theta(\bar{X}_n) = 1/n$, il est d'ailleurs également efficace. L'estimateur de Hodges $\hat{\theta}_n$ s'obtient en annulant ce premier estimateur lorsqu'il est proche de 0, à savoir

$$\hat{\theta}_n = \bar{X}_n \mathbb{1}_{|\bar{X}_n| \geq n^{-1/4}}.$$

   Autrement dit, si la moyenne empirique est proche de 0 alors on estime θ par 0, sinon on garde la moyenne empirique. Etudions la normalité asymptotique de cet estimateur.
   — Si $\theta = 0$, alors $\bar{X}_n \sim \mathcal{N}(0, 1/n)$ et pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\sqrt{n}\hat{\theta}_n\right| \geq \varepsilon\right) = \mathbb{P}\left(\left|\sqrt{n}\bar{X}_n \mathbb{1}_{|\bar{X}_n| \geq n^{-1/4}}\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\bar{X}_n\right| \geq n^{-1/4}\right) = \mathbb{P}\left(\left|n^{1/4}\bar{X}_n\right| \geq 1\right),$$

   or, par le Lemme de Slutsky [18],

$$n^{1/4}\bar{X}_n = n^{-1/4} \times (\sqrt{n}\bar{X}_n) \xrightarrow[n\to\infty]{d} 0 \times \mathcal{N}(0, 1) = 0,$$

18. On peut aussi conclure par l'inégalité de Markov puisque $\mathbb{P}\left(\left|n^{1/4}\bar{X}_n\right| \geq 1\right) \leq \mathbb{E}[n^{1/2}(\bar{X}_n)^2] = n^{-1/2}$.

d'où l'on déduit que $n^{1/4}\bar{X}_n$ tend en probabilité vers 0, et idem pour $\sqrt{n}\hat{\theta}_n$. Ainsi, lorsque $\theta = 0$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,0) \qquad \text{avec} \qquad 0 < 1 = \frac{1}{I(0)}.$$

— Lorsque $\theta \neq 0$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \sqrt{n}\left(\bar{X}_n \mathbb{1}_{|\bar{X}_n| \geq n^{-1/4}} - \theta\right) = \sqrt{n}\left(\bar{X}_n - \theta\right) - \sqrt{n}\bar{X}_n \mathbb{1}_{|\bar{X}_n| < n^{-1/4}}.$$

Cette fois, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\sqrt{n}\bar{X}_n \mathbb{1}_{|\bar{X}_n| < n^{-1/4}}\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\bar{X}_n\right| < n^{-1/4}\right) = \mathbb{P}\left(n^{1/4}\left|\bar{X}_n\right| < 1\right),$$

or par la Loi des Grands Nombres et le théorème de continuité, $\left|\bar{X}_n\right|$ tend presque sûrement vers $|\theta|$, donc $n^{1/4}\left|\bar{X}_n\right|$ tend presque sûrement vers $+\infty$ et le théorème de convergence dominée permet de conclure :

$$\mathbb{P}\left(n^{1/4}\left|\bar{X}_n\right| < 1\right) = \mathbb{E}\left[\mathbb{1}_{n^{1/4}|\bar{X}_n|<1}\right] \xrightarrow[n\to\infty]{} 0,$$

c'est-à-dire que

$$\sqrt{n}\bar{X}_n \mathbb{1}_{|\bar{X}_n| \leq n^{-1/4}} \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

et par Slutsky

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1) \qquad \text{avec} \qquad 1 = \frac{1}{I(\theta)}.$$

Autrement dit, l'estimateur de Hodges a une variance asymptotique en $1/I(\theta)$ pour tout $\theta \neq 0$ et une variance asymptotique strictement plus petite pour $\theta = 0$: on dit qu'il est super-efficace. On peut toutefois montrer que pour tout n, l'erreur quadratique moyenne $R(\hat{\theta}_n, \theta)$ est détériorée localement autour de 0 par rapport à celle de la moyenne empirique $R(\bar{X}_n, \theta) = 1/n$. Précisément, il existe une constante $c > 0$ indépendante de n et de θ telle que $\sup_{|\theta| \leq n^{-1/4}} nR(\hat{\theta}_n, \theta) \geq c\sqrt{n}$. Ce comportement, parfois appelé phénomène de Hodges, est illustré Figure 2.14.

4. Un résultat (difficile) dû à Le Cam assure néanmoins qu'on ne peut avoir super-efficacité, i.e. $\sigma^2(\theta) < 1/I(\theta)$, que sur un ensemble Θ_0 de mesure de Lebesgue nulle. **On retient** : en général, un estimateur asymptotiquement efficace de θ est donc un estimateur asymptotiquement normal de variance asymptotique $1/I(\theta)$ pour tout θ tel que $I(\theta) > 0$.

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 $\mathbf{X} = (X_1, \ldots, X_n)$ est un échantillon i.i.d. selon f_λ, le Théorème Central Limite nous dit que

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1/\lambda^2).$$

La méthode Delta donne alors

$$\sqrt{n}\left(\tilde{\lambda}_n(\mathbf{X}) - \lambda\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \lambda^2) \qquad \text{avec} \qquad \lambda^2 = \frac{1}{I(\lambda)},$$

donc cet estimateur est asymptotiquement efficace.

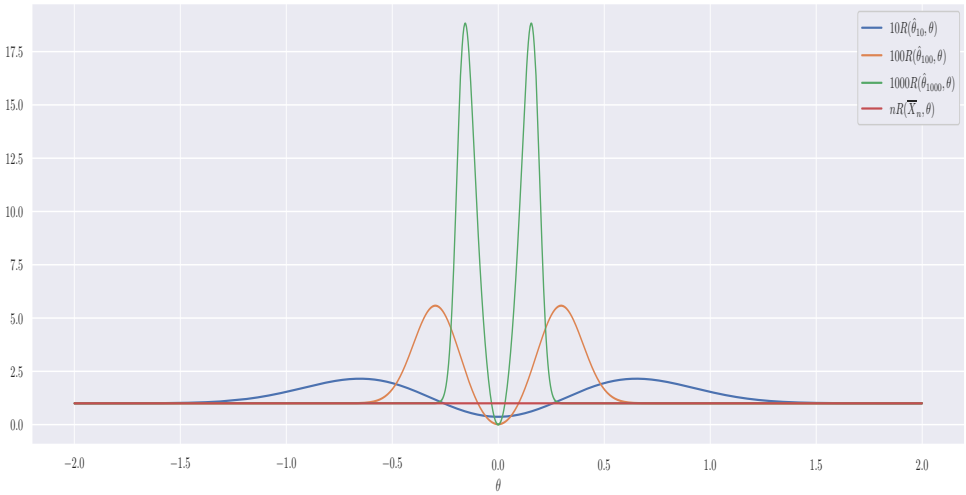

FIGURE 2.14 – Phénomène de Hodges : risques quadratiques normalisés $\theta \mapsto nR(\bar{X}_n, \theta) = 1$ et $\theta \mapsto nR(\hat{\theta}_n, \theta)$ pour $n = 10$, $n = 100$, et $n = 1000$.

2. Voyons ce qui peut se passer lorsque l'information de Fisher s'annule. On effectue un changement de paramètre dans l'exemple des gaussiennes translatées en considérant le modèle $(\mathcal{N}(\eta^3, 1))_{\eta\in\mathbb{R}}$, où $\eta \in \mathbb{R}$ est le paramètre inconnu que l'on cherche à estimer. Avec les notations de la Proposition 12, on a la bijection $\theta = \eta^3 = \psi(\eta)$ avec ψ de classe C^1. Ce modèle est donc régulier, d'information de Fisher

$$J(\eta) = \psi'(\eta)^2 I(\psi(\eta)) = 9\eta^4,$$

qui est strictement positive si et seulement si η est non nul. L'estimateur naturel (moments ou EMV) est

$$\hat{\eta}_n(\mathbf{X}) = \bar{X}_n^{1/3} = \left(\frac{X_1 + \cdots + X_n}{n}\right)^{1/3}.$$

On sait que

$$\sqrt{n}\left(\bar{X}_n - \eta^3\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1)$$

donc la méthode Delta telle qu'énoncée en Proposition 6 assure que, si $\eta \neq 0$,

$$\sqrt{n}\left(\hat{\eta}_n(\mathbf{X}) - \eta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1/(9\eta^4)) \qquad \text{avec} \qquad \frac{1}{9\eta^4} = \frac{1}{J(\eta)},$$

ce qui prouve l'efficacité asymptotique. Si $\eta = 0$, alors $\sqrt{n}\bar{X}_n \sim \mathcal{N}(0, 1)$. Soit Z une variable aléatoire distribuée selon une loi normale centrée réduite, alors on a les égalités en loi suivantes :

$$\sqrt{n}\bar{X}_n \overset{\text{loi}}{=} Z \implies n^{1/6}\hat{\eta}_n(\mathbf{X}) \overset{\text{loi}}{=} Z^{1/3}.$$

Ou encore, de façon équivalente : notons Y une variable telle que $Y^3 \sim \mathcal{N}(0, 1)$, alors

$$n^{1/6}\left(\hat{\eta}_n(\mathbf{X}) - 0\right) \overset{\text{loi}}{=} Y.$$

La variable Y n'est pas gaussienne : elle est bimodale (voir Figure 2.15), 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 en loi, 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[19]. Cependant, l'estimateur $\hat{\eta}_n$ est asymptotiquement efficace puisque, pour tout η tel que $J(\eta) \neq 0$, il est asymptotiquement normal de variance limite $1/J(\eta)$. Cet exemple permet simplement de constater que, en un point où l'information de Fisher s'annule, le comportement d'un estimateur asymptotiquement efficace peut être complètement différent de ce qui se passe partout ailleurs : ici, lorsque $J(\eta) = 0$, i.e. lorsque $\eta = 0$, la vitesse n'est plus en $1/\sqrt{n}$ et la loi limite n'est plus gaussienne.

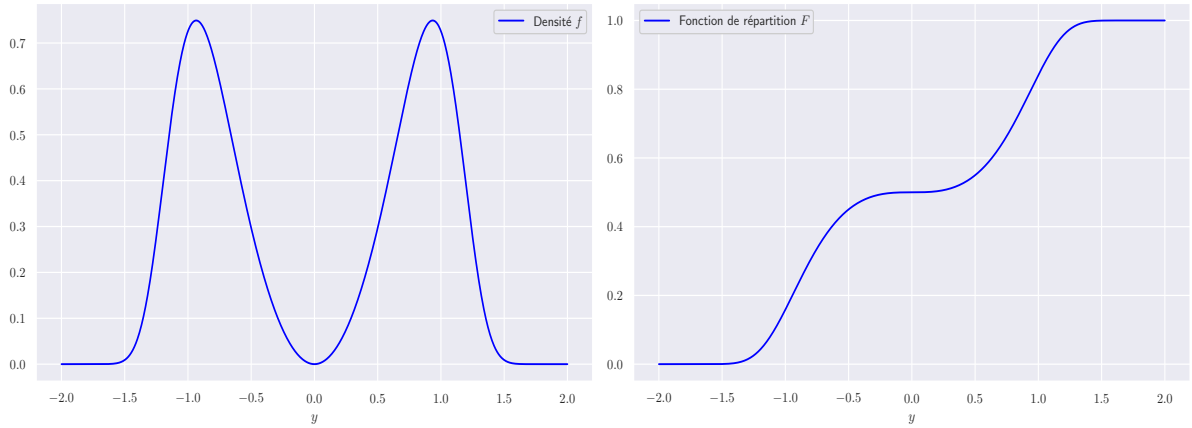

FIGURE 2.15 – Fonction de répartition et densité de la variable Y telle que $Y^3 \sim \mathcal{N}(0,1)$.

Modulo une hypothèse de domination, on peut montrer (mais nous l'admettrons[20]) un résultat général assurant l'efficacité asymptotique de l'estimateur du maximum de vraisemblance dans un modèle régulier.

Théorème 11 (EMV et efficacité asymptotique)
Soit un modèle régulier $(f_\theta)_{\theta\in\Theta}$ d'information de Fisher $I(\theta)$, soit $\theta^* \in \Theta$ vérifiant $I(\theta^*) > 0$ et $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon i.i.d. selon f_{θ^*}. S'il existe une suite $(\hat{\theta}_n(\mathbf{X}))_{n\geq n_0}$ d'estimateurs du maximum de vraisemblance consistante pour θ^* ainsi qu'un réel $h > 0$ tel que

$$\mathbb{E}_{\theta^*}\left[\sup_{\theta^*-h\leq\theta\leq\theta^*+h} \ell'_\theta(X_1)^2\right] < \infty, \tag{2.13}$$

alors

$$\sqrt{n}\left(\hat{\theta}_n(\mathbf{X}) - \theta^*\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1/I(\theta^*)),$$

c'est-à-dire qu'on a efficacité asymptotique.

Exemple : Dans le modèle régulier déjà mentionné où $X \sim \mathcal{E}(\theta)$, nous avons vu que $I(\theta) = I_1(\theta) = 1/\theta^2 > 0$ pour tout $\theta > 0$. Considérons $\theta^* > 0$ fixé et un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. selon la loi $\mathcal{E}(\theta^*)$. L'estimateur du maximum de vraisemblance est $\hat{\theta}_n = 1/\bar{X}_n$ et la loi des grands nombres montre qu'il est consistant. Nous avons même vérifié qu'il est en fait asymptotiquement efficace.

19. Pour voir que $\sqrt{n}(\hat{\eta}_n(\mathbf{X}) - \eta) = \sqrt{n}\hat{\eta}_n(\mathbf{X})$ ne converge pas en loi, il suffit de noter que, pour tout réel t, on a $\mathbb{P}(\sqrt{n}\hat{\eta}_n(\mathbf{X}) \leq t) = \mathbb{P}(\mathcal{N}(0,1) \leq t^3/n) = \Phi(t^3/n) \to 1/2$ quand $n \to \infty$, ce qui exclut l'existence d'une fonction de répartition F telle que la limite précédente coïncide avec F en tout point de continuité de celle-ci.
20. Voir [6] pour une preuve.

On peut retrouver ce dernier point grâce au résultat général précédent. En effet, prenons $h = \theta^*/2$, alors un calcul déjà fait donne $\ell'_\theta(X_1) = (1/\theta - X_1)$ donc, pour tout $\theta \in [\theta^*/2; 3\theta^*/2]$,

$$\ell'_\theta(X_1)^2 = \left(\frac{1}{\theta} - X_1\right)^2 = \left|\frac{1}{\theta} - X_1\right|^2 \leq \left(\frac{1}{\theta} + X_1\right)^2 \leq \left(\frac{2}{\theta^*} + X_1\right)^2,$$

et puisque $X_1 \sim \mathcal{E}(\theta^*)$ admet un moment d'ordre 2, on a bien

$$\mathbb{E}_{\theta^*}\left[\sup_{\theta^*-h\leq\theta\leq\theta^*+h} \ell'_\theta(X_1)^2\right] \leq \mathbb{E}_{\theta^*}\left[\left(\frac{2}{\theta^*} + X_1\right)^2\right] < \infty,$$

donc d'après le Théorème 11 l'EMV $\hat{\theta}_n = 1/\bar{X}_n$ est asymptotiquement efficace. Sur cet exemple élémentaire, on constate néanmoins que la vérification directe par le TCL et la méthode Delta permettent de conclure plus rapidement.

Bilan : Pour reprendre la question posée en début de section : "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 d'un EMV consistant, hypothèse de domination (2.13), non-nullité de l'information de Fisher). Encore faut-il pouvoir le calculer, ce qui n'est pas toujours chose facile. De plus, comme nous l'avons vu, l'EMV souffre d'un manque de robustesse aux données aberrantes ou à une mauvaise spécification du modèle.

Notant θ^* la vraie valeur du paramètre, le Théorème 11 signifie que plus l'information de Fisher en ce point est grande, plus on peut estimer précisément θ^*, en particulier par l'estimateur du maximum de vraisemblance. Dit autrement, plus $I(\theta^*)$ est grande, plus l'information moyenne apportée par une donnée est importante : on peut par exemple écrire

$$\mathbb{P}_{\theta_0}\left(\hat{\theta}_n - \frac{2}{\sqrt{nI(\theta^*)}} \leq \theta^* \leq \hat{\theta}_n + \frac{2}{\sqrt{nI(\theta^*)}}\right) \xrightarrow[n\to\infty]{} 0.95.$$

On notera au passage que ceci ne correspond pas à un intervalle de confiance asymptotique à 95% : puisqu'on ne connaît pas θ^*, en général on ne connaît pas non plus $I(\theta^*)$. Néanmoins, puisque la fonction I est continue, si l'on dispose d'une formule explicite pour celle-ci, il suffit de remplacer $I(\theta^*)$ par $I(\hat{\theta}_n)$ pour en déduire un intervalle de confiance asymptotique.

Par ailleurs, on peut donner une interprétation graphique de l'information de Fisher grâce au lien avec la théorie de l'information [21]. 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 f par rapport à g 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 :

$$-D(f \parallel g) = \int \log\left(\frac{g(x)}{f(x)}\right) f(x)dx \leq \log\left(\int g(x)dx\right) = 0.$$

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 statistique, supposons que θ^* soit la vraie valeur du paramètre, alors pour une autre valeur θ, la divergence de f_θ à f_{θ^*} peut encore s'écrire

$$D(f_{\theta^*} \parallel f_\theta) = -\mathbb{E}_{\theta^*}[\ell_\theta(X) - \ell_{\theta^*}(X)].$$

FIGURE 2.16 – Divergence et information de Fisher, avec $I(\theta^*)$ plus grande à droite qu'à gauche.

Sous les hypothèses de régularité ad hoc, on a donc au voisinage de θ^*

$$\ell_\theta(X) \approx \ell_{\theta^*}(X) + \ell'_{\theta^*}(X)(\theta - \theta^*) + \tfrac{1}{2}\ell''_{\theta^*}(X)(\theta - \theta^*)^2.$$

Passant à l'espérance, puisque le score est centré, on en déduit que

$$D(f_{\theta^*} \parallel f_\theta) \approx \tfrac{1}{2} I(\theta^*)(\theta - \theta^*)^2.$$

Autrement dit, l'information de Fisher en θ^* correspond à la courbure de la divergence de Kullback-Leibler au voisinage de θ^*. Plus cette courbure est importante, plus il est facile de discriminer entre la vraie valeur θ^* et une valeur voisine, et inversement. La Figure 2.16 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

$$D(f_{\theta^*} \parallel f_\theta) = \mathbb{E}_{\theta^*}[\ell_{\theta^*}(X)] - \mathbb{E}_{\theta^*}[\ell_\theta(X)],$$

c'est-à-dire qui maximise la fonction $\theta \mapsto \mathbb{E}_{\theta^*}[\ell_\theta(X)]$, 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 i.i.d. de densité f_{θ^*}, alors

$$\frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i) \xrightarrow[n\to\infty]{p.s.} \mathbb{E}_{\theta^*}[\ell_\theta(X)].$$

Or maximiser le terme de gauche, c'est justement ce que fait l'estimation au maximum de vraisemblance.

21. Voir [5] pour une introduction à ce domaine, en particulier le chapitre *Information Theory and Statistics*.

Chapitre 3

Le modèle linéaire gaussien

Introduction

Le principe de la régression est de modéliser une variable y, dite variable à expliquer ou variable réponse, comme une fonction de p variables [1] $\mathbf{x} = [x_1, \ldots, x_p]'$, dites variables explicatives :

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

On dispose de n de couples $(\mathbf{x}_i, y_i)_{1 \leq i \leq n}$ et le but est de retrouver la fonction g. Le modèle le plus simple est celui d'une relation linéaire, c'est-à-dire qu'on suppose l'existence d'un vecteur de paramètres $\beta = [\beta_1, \ldots, \beta_p]'$ 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 des données à disposition, 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 7. Les exemples d'applications de la régression linéaire foisonnent, on se contente ici d'en mentionner quelques-uns :

1. **Concentration de l'ozone** : dans ce domaine, on cherche à expliquer le maximum journalier de la concentration en ozone, notée O_3 (en μg/m^3), en fonction de la température à midi T. Le nuage de points de la Figure 3.1 (à gauche) 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 [2] 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.

1. Dans tout ce chapitre, le symbole $'$ correspond à la transposition.
2. Celle-ci prend des valeurs entières de 0 à 8, pour un ciel allant de très dégagé à très couvert.

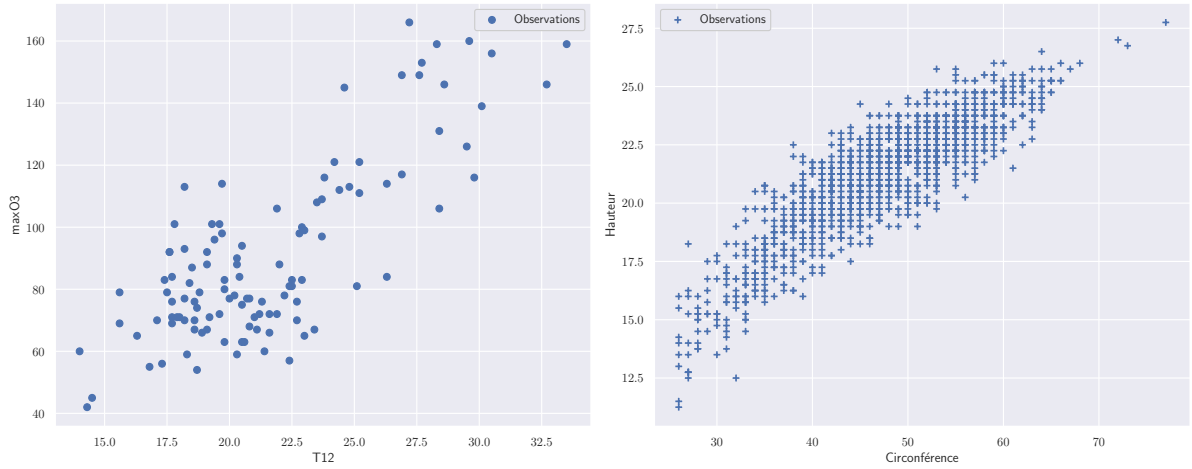

FIGURE 3.1 – Nuages de points pour l'ozone et les eucalyptus.

2. **Hauteur d'un eucalyptus** : la Figure 3.1 (à droite) 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 !
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
$$\log P = \beta_1 + \beta_2 \log K + \beta_3 \log T + \varepsilon.$$
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...

3.1 Régression linéaire multiple

3.1.1 Modélisation

Nous supposons que les données collectées suivent le modèle suivant :
$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{3.1}$$
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 (3.1), 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 une écriture matricielle pour (3.1), nous obtenons la définition suivante :

Définition 28 (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

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

L'hypothèse $(\mathcal{H}_1)$ assure que le modèle est identifiable, nous y reviendrons en Section 3.2 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 carrée $X'X$ inversible. Supposons en effet X de rang p : puisque $\mathrm{rg}(X) \leq \min(n, p)$, ceci implique bien $p \leq n$. De plus, s'il existait un vecteur α de $\mathbb{R}^p$ tel que $(X'X)\alpha = 0$, on aurait $\|X\alpha\|^2 = \alpha'(X'X)\alpha = 0$, donc $X\alpha = 0$, d'où $\alpha = 0$ puisque $\mathrm{rg}(X) = p$. La réciproque est claire : si $X'X$ est inversible, alors une matrice et sa transposée ayant le même rang, il vient

$$p = \mathrm{rg}(X'X) \leq \min(\mathrm{rg}(X'), \mathrm{rg}(X)) = \mathrm{rg}(X) \leq \min(n, p) \;\Rightarrow\; \mathrm{rg}(X) = p \leq n.$$

Concrètement, si $\mathrm{rg}(X) < p$, ceci signifie que (au moins) l'une des colonnes de la matrice X du plan d'expérience est combinaison linéaire des autres, c'est-à-dire que la variable correspondant à cette colonne n'apporte (linéairement) aucune information supplémentaire : elle est donc inutile.

Remarque : La matrice $X'X$ est symétrique et on vient de voir que, sous l'hypothèse $(\mathcal{H}_1)$, pour tout $\alpha \in \mathbb{R}^p$ non nul, on a $\alpha'(X'X)\alpha = \|X\alpha\|^2 > 0$. Autrement dit, la matrice $X'X$ est symétrique 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 3.1, 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 (3.1) s'écrit encore

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

et de façon matricielle on peut aussi écrire

$$Y = X\beta + \boldsymbol{\varepsilon} = \beta_1 X_1 + \cdots + \beta_p X_p + \boldsymbol{\varepsilon}.$$

3.1.2 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, comme nous le verrons en Section 3.2.4.

Définition 29 (Estimateur des Moindres Carrés)
L'estimateur des moindres carrés $\hat{\beta}$ est défini comme suit :

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

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 $\mathbb{R}^p$:

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

Selon le modèle de la Définition 28, 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 $\mathcal{M}_X$ le plus proche de Y au sens de la norme euclidienne. Cet élément, unique puisque $\mathcal{M}_X$ est un convexe fermé de $\mathbb{R}^n$, est par définition le projeté orthogonal de Y sur $\mathcal{M}_X$ (voir Figure 3.2). 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 16 (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 :

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

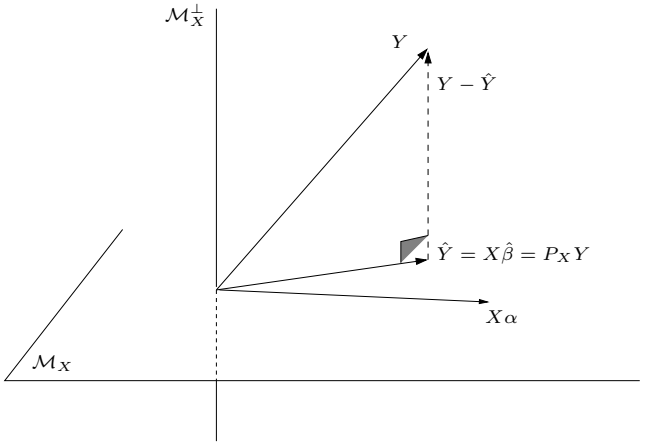

FIGURE 3.2 – Interprétation de $\hat{Y} = X\hat{\beta}$ comme projeté orthogonal de Y sur $\mathcal{M}_X$.

Preuve : On peut montrer ce résultat de plusieurs façons.

1. Par projection : il suffit de 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 :

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

   Ces p équations se regroupent en une seule : $X'(Y - X\hat{\beta}) = 0$, d'où l'on déduit bien l'expression de $\hat{\beta} = (X'X)^{-1}X'Y$. Puisque par définition $\hat{Y} = P_X Y = X\hat{\beta} = X(X'X)^{-1}X'Y$ et comme cette relation est valable pour tout $Y \in \mathbb{R}^n$, on en déduit que $P_X = X(X'X)^{-1}X'$.

2. 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 $\hat{\beta} = (X'X)^{-1}X'Y$ et par le même raisonnement que ci-dessus il s'ensuit que $P_X = X(X'X)^{-1}X'$.

■

Remarques :

1. Puisque $Y = X\beta + \boldsymbol{\varepsilon}$, l'estimateur $\hat{\beta}$ s'écrit encore

$$\hat{\beta} = \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon}. \tag{3.2}$$

   Vu que β et $\boldsymbol{\varepsilon}$ sont inconnus, cette expression ne permet en rien de calculer $\hat{\beta}$. Néanmoins, elle va s'avérer utile pour établir certaines propriétés de cet estimateur : en particulier, elle montre que $\hat{\beta}$ est une transformation affine du vecteur aléatoire $\boldsymbol{\varepsilon}$.

2. Dire que la matrice X n'est pas de rang p signifie que le sous-espace $\mathcal{M}_X$ engendré par ses colonnes est strictement inférieur à p, ou encore que le noyau de l'application linéaire $\alpha \in \mathbb{R}^p \mapsto X\alpha \in \mathbb{R}^n$ n'est pas réduit à 0. La projection $\hat{Y}$ sur $\mathcal{M}_X$ reste bien définie, mais on perd l'unicité de l'estimateur des moindres carrés puisque si $\hat{\beta}$ permet d'atteindre le minimum, celui-ci est encore atteint pour tout vecteur de la forme $\hat{\beta}+\alpha$ avec α appartenant au noyau de X.

Exemples.

1. La droite des moindres carrés pour le modèle expliquant le maximum journalier de l'ozone en fonction de la température à midi est superposée au nuage de points en Figure 3.3 à gauche.
2. Pour 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 3.3 à droite.

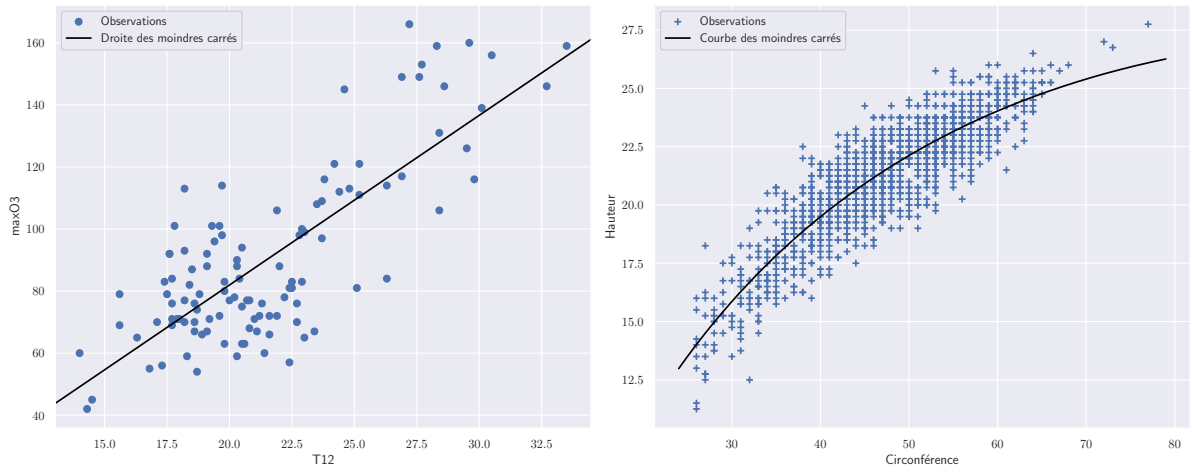

FIGURE 3.3 – Droite et courbe des moindres carrés pour l'ozone et les eucalyptus.

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

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

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 $p \times p$. Elle est symétrique semi-définie positive, mais pas nécessairement définie positive. De plus, pour pour toute matrice A de taille $m \times p$ et tout vecteur b de dimension m déterministes, on a

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

Ces propriétés élémentaires seront très souvent appliquées dans la suite, et en particulier dans le résultat suivant.

Proposition 17 (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

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

Preuve : D'après (3.2), $\hat{\beta}$ est une transformation affine du vecteur aléatoire $\boldsymbol{\varepsilon}$. Puisque $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$, il vient

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

Pour la covariance, vu que $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, on procède de même :

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}(\beta + (X'X)^{-1}X'\boldsymbol{\varepsilon}) = (X'X)^{-1}X'\mathrm{Cov}(\boldsymbol{\varepsilon})X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

■

Comme $Y = X\beta + \boldsymbol{\varepsilon}$ et $X\beta \in \mathcal{M}_X$, il est clair que $P_{X^\perp}Y = P_{X^\perp}\boldsymbol{\varepsilon}$. Ceci donne plusieurs formulations pour le vecteur des résidus que nous définissons maintenant (voir Figure 3.4) et qui va nous permettre d'estimer σ^2.

Définition 30 (Résidus)
On appelle vecteur des résidus le vecteur aléatoire de taille n défini par

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

On appelle Somme des Carrés Résiduelle le carré de la norme euclidienne de ce vecteur :

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

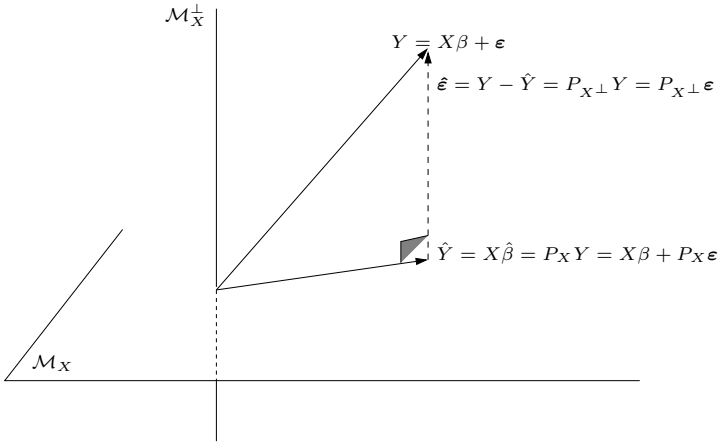

FIGURE 3.4 – Interprétation de $\hat{\boldsymbol{\varepsilon}}$ comme projeté orthogonal de Y sur $\mathcal{M}_X^\perp$.

Noter que dans la définition précédente, la dernière expression $\hat{\boldsymbol{\varepsilon}} = P_{X^\perp}\boldsymbol{\varepsilon}$ ne permet pas, contrairement aux autres, de calculer les résidus puisque le vecteur des erreurs $\boldsymbol{\varepsilon}$ est inconnu. A nouveau, cette formule est néanmoins utile dans certains cas. Par ailleurs, si $\hat{\beta}$ estime bien β, alors d'une certaine façon les résidus $\hat{\boldsymbol{\varepsilon}} = Y - X\hat{\beta}$ estiment bien les erreurs $\boldsymbol{\varepsilon} = Y - X\beta$, donc un estimateur "naturel" de la variance résiduelle σ^2 est donné par :

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

En fait, comme on va le voir, cet estimateur est biaisé. Ce biais est néanmoins facilement corrigeable, comme le montre le résultat suivant.

Proposition 18 (Estimateur de la variance)
La statistique

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

est un estimateur sans biais de σ^2.

Remarque : Ceci suppose bien entendu qu'on a en fait $p < n$. Ceci n'a rien d'étonnant : si $p = n$ avec $\mathrm{rg}(X) = p$, alors $Y \in \mathcal{M}_X$ donc $Y = \hat{Y} = X\hat{\beta}$ et $\hat{\boldsymbol{\varepsilon}} = 0$. Du point de vue des données, tout se passe comme s'il n'y avait pas de terme d'erreur $\boldsymbol{\varepsilon}$ dans le modèle initial $Y = X\beta + \boldsymbol{\varepsilon}$. Cette situation ne nous intéressera pas.

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 :

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

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 sans biais de la matrice de 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

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

Attention ! 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 $(X'X)$".

Exercice : On considère le modèle $Y_i = \beta_1 + \varepsilon_i$ avec les ε_i i.i.d. centrées de même variance σ^2 et on applique la méthode précédente pour estimer β_1 et σ^2. Vérifier que $\hat{\beta}_1 = \bar{Y}_n$ (moyenne empirique des observations Y_i) et que $\hat{\sigma}^2$ est l'estimateur sans biais de la variance dans le modèle d'échantillonnage, à savoir :

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

3.2 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.

Remarque : Contrairement à tous les exemples des Chapitre 1 et 2, 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. Elles sont néanmoins indépendantes puisque les erreurs ε_i le sont.

3.2.1 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 $\Sigma_Y = \mathrm{Cov}(Y) = \mathbb{E}[(Y-\mu)(Y-\mu)']$ 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

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

Remarque : Si l'on reprend la Définition 6 d'une expérience statistique, l'objet aléatoire est ici le vecteur $Y = X\beta + \boldsymbol{\varepsilon}$ de $\mathbb{R}^n$, de loi normale $\mathcal{N}(X\beta, \sigma^2 I_n)$. En accord avec la Définition 8, 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 $(\mathcal{H}_1)$.

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 (i.e. symétrique définie positive), auquel cas :

$$f(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma_Y)}} e^{-\frac{1}{2}(y-\mu)'\Sigma_Y^{-1}(y-\mu)}. \tag{3.3}$$

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

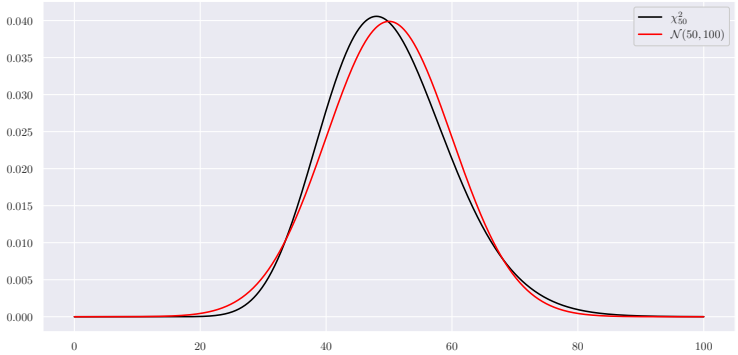

FIGURE 3.5 – Densités d'un χ^2_{50} et d'une $\mathcal{N}(50, 100)$.

Définition 31 (Lois du khi-deux, de Student et de Fisher)
Soit $X_1, \ldots, X_d$ des variables aléatoires i.i.d. 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* $S \sim \chi^2_d$, *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* $F \sim \mathcal{F}(d_1, d_2)$.

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 3.5 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 par ailleurs d'aucune utilité dans ce qui suit.

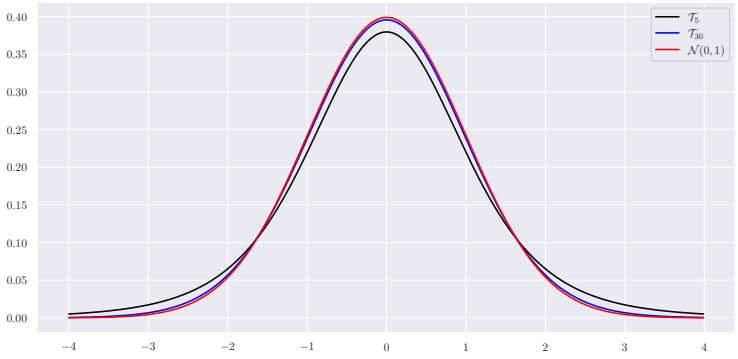

FIGURE 3.6 – Densités d'une $\mathcal{T}_5$, d'une $\mathcal{T}_{30}$ et d'une $\mathcal{N}(0,1)$.

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, en notant S_d au lieu de S et puisque $\mathbb{E}[S_d] = d$ et $\mathrm{Var}(S_d) = 2d$, l'inégalité de Tchebychev assure que la suite de variables aléatoires (S_d/d) tend vers 1 en probabilité : en effet, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_d}{d} - 1\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(S_d/d)}{\varepsilon^2} = \frac{2}{d\varepsilon^2} \xrightarrow[d\to\infty]{} 0.$$

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 3.6 pour $d = 10$ ddl. Par conséquent, lorsque d est grand, les quantiles d'une loi de Student $\mathcal{T}_d$ sont très proches de 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 S_2/d_2 tend vers 1 en probabilité. 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 3.7 pour $d_1 = 2$ et $d_2 = 10$.

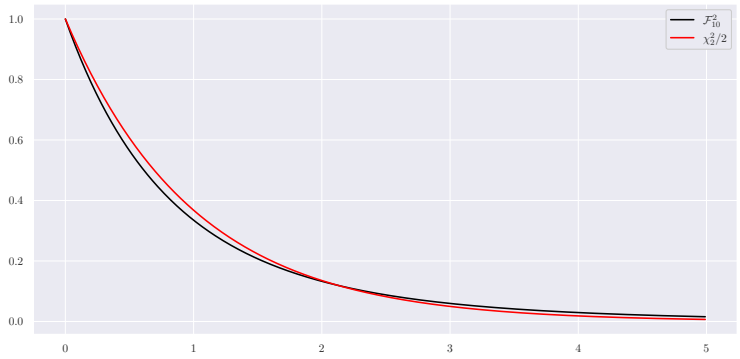

FIGURE 3.7 – Densités d'une $\mathcal{F}_{10}^2$ et d'un $\frac{\chi_2^2}{2}$.

Proposition 19 (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

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

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 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.

Rappel : Si X et Y sont deux vecteurs aléatoires de tailles respectives m et p dont toutes les composantes sont de carré intégrable, la covariance de (X, Y) est la matrice $m \times p$ définie par

$$\Sigma_{X,Y} = \mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])'] = \mathbb{E}[XY'] - \mathbb{E}[X]\mathbb{E}[Y]' = \mathrm{Cov}(Y,X)' = \Sigma_{Y,X}',$$

c'est-à-dire de terme générique $\Sigma_{X,Y}(i,j) = \mathrm{Cov}(X_i, Y_j)$. Dans le cas où le vecteur (X, Y) est gaussien, les vecteurs X et Y sont indépendants si et seulement si cette matrice est nulle.

Le Théorème de Cochran, très utile dans la suite, assure que la décomposition d'un vecteur gaussien à composantes indépendantes et de même variance sur des sous-espaces orthogonaux donne des vecteurs indépendants dont on peut expliciter les lois. Il peut ainsi être vu comme une version aléatoire du Théorème de Pythagore (voir Figure 3.8).

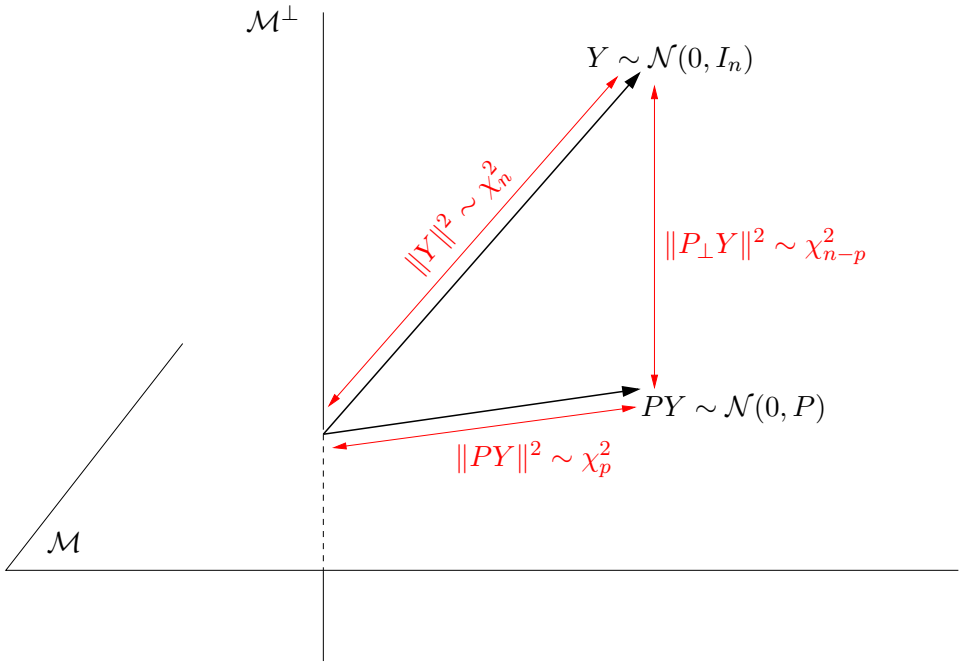

FIGURE 3.8 – Interprétation géométrique du Théorème de Cochran lorsque $Y \sim \mathcal{N}(0, I_n)$.

Théorème 12 (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) $\frac{\|P(Y-\mu)\|^2}{\sigma^2} \sim \chi^2_p$ et $\frac{\|P_\perp(Y-\mu)\|^2}{\sigma^2} \sim \chi^2_{n-p}$.

Preuve :

(i) Ce premier point est clair par stabilité des vecteurs gaussiens par transformation linéaire et puisque P et $P_\perp$ sont des projections.
(ii) Toujours par stabilité, le vecteur de taille $2n$ obtenu en empilant PY et $P_\perp Y$ est lui aussi gaussien. Pour prouver que PY et $P_\perp Y$ sont indépendants, il suffit donc de montrer que leur covariance est nulle. Or, puisque $P'_\perp = P_\perp$, on a tout simplement

$$\mathrm{Cov}(PY, P_\perp Y) = \mathbb{E}[PY(P_\perp Y)'] - \mathbb{E}[PY]\mathbb{E}[P_\perp Y]' = P\mathrm{Cov}(Y)P_\perp = \sigma^2 P P_\perp = 0.$$

(iii) D'après le premier point, $P(Y - \mu) \sim \mathcal{N}(0, \sigma^2 P)$. Par ailleurs, il existe une matrice orthogonale Q telle que $P = Q\Delta Q'$ où Δ est une matrice diagonale dont les p premiers éléments diagonaux valent 1 et les $(n-p)$ suivants valent 0. Soit maintenant $X = [X_1, \ldots, X_n]'$ vecteur aléatoire dont les p premières composantes sont des variables gaussiennes indépendantes centrées et réduites tandis que les $(n-p)$ dernières valent 0. Le vecteur X ainsi construit est gaussien, avec $X \sim \mathcal{N}(0, \Delta)$, donc σQX est aussi gaussien, avec $\sigma QX \sim \mathcal{N}(0, \sigma^2 P)$. Autrement dit, les vecteurs aléatoires σQX et $P(Y-\mu)$ ont même loi, donc les variables aléatoires $\sigma^2\|QX\|^2$ et $\|P(Y-\mu)\|^2$ aussi. Or

$$\|QX\|^2 = X'Q'QX = X'X = \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{p} X_i^2 \sim \chi^2_p.$$

■

Remarque : Si on projette un vecteur gaussien sur deux sous-espaces orthogonaux, les vecteurs aléatoires obtenus seront par définition orthogonaux, mais ils n'ont en général aucune raison d'être indépendants. Il suffit de considérer 3 gaussiennes i.i.d. standards (W_1, W_2, W_3) et, dans $\mathbb{R}^2$, le vecteur aléatoire $[X, Y]' = [W_1 + W_2, W_1 + W_3]'$. La projection sur l'axe des abscisses (respectivement des ordonnées) est le vecteur $V_1 = [X, 0]'$ (respectivement $V_2 = [0, Y]'$), or ces deux vecteurs ne sont pas indépendants puisque leur covariance n'est pas nulle :

$$\mathrm{Cov}(V_1, V_2) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

L'hypothèse cruciale dans le Théorème de Cochran est la forme de la matrice de covariance du vecteur initial, proportionnelle à l'identité, c'est-à-dire que ses composantes sont des variables gaussiennes i.i.d.

Nous allons voir en section suivante comment le résultat de Cochran s'applique dans notre cadre.

3.2.2 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 à partir de 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 : $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$;
(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) D'après (3.2), $\hat{\beta} = \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon}$, 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 sa moyenne et sa matrice de dispersion, lesquelles ont été établies en Proposition 17.

(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 (déterministe !) 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 (déterministe !) 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 déterministes de l'un et de l'autre.

(iii) Puisque $\hat{\boldsymbol{\varepsilon}} = P_{X^\perp}\boldsymbol{\varepsilon}$ avec $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, le théorème de Cochran assure 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 18, 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, on a

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

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 20 (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

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

Preuve :

(i) D'après la proposition précédente, on sait d'une part que

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 \left[(X'X)^{-1}\right]_{jj}),$$

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 19 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.

■

Remarque : La matrice $(X'X)$ étant symétrique définie positive, c'est aussi le cas pour son inverse $(X'X)^{-1}$. Or si S est symétrique définie positive, tous ses coefficients diagonaux sont strictement positifs puisque si e_j désigne le j^e vecteur de la base canonique, alors $S_{jj} = e_j' S e_j > 0$. Dès lors, la division par $\sqrt{[(X'X)^{-1}]_{jj}}$ dans la définition de T_j ne pose pas problème.

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 10 du Chapitre 1, car elles font intervenir les paramètres β et σ^2 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 20.

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 $i \in \{1, \ldots, n\}$, $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$. 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. L'hypothèse $(\mathcal{H}_1)$ correspond à supposer X de rang 2, ce qui revient à dire que $n \geq 2$ et que les x_i ne sont pas tous égaux. On a ensuite

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

et le point (ii) de la Proposition 20 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 $\hat{\beta}_j$, laquelle incite plutôt à étudier des domaines de confiance. Nous allons donc traiter les deux aspects, en considérant σ^2 inconnue, ce qui est généralement le cas en pratique.

Corollaire 5 (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 d'ordre $(1-\alpha/2)$ 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

$$\left\{\beta \in \mathbb{R}^p : \frac{1}{p\hat{\sigma}^2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \leq f^p_{n-p}(1-\alpha)\right\}. \tag{3.4}$$

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

Preuve : Il suffit d'appliquer le point (iii) des Propriétés 1 et les résultats de la Proposition 20. ■

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 $p = 2$. Un domaine de confiance de niveau $(1 - \alpha)$ pour (β_1, β_2) est défini par l'équation :

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

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 $X'X$, laquelle est bien définie positive grâce à $(\mathcal{H}_1)$. 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 3.9 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). Bien entendu, dans le cas de p variables explicatives, il faudrait considérer des intervalles de confiance de niveau $(1 - \alpha/p)$.

Remarque : De façon générale, on peut montrer que l'hyper-ellipsoïde de confiance de niveau $(1 - \alpha)$ obtenu grâce au Corollaire 5 est toujours de mesure de Lebesgue inférieure à celle de l'hyper-rectangle de confiance de niveau $(1 - \alpha)$ déduit des intervalles de confiance de niveau $(1 - \alpha/p)$. La preuve de ce résultat intuitivement clair n'est cependant pas complètement évidente.

3.2.3 Prévision

Une fois le modèle de régression construit, c'est-à-dire une fois les paramètres β et σ^2 estimés à partir des n observations $(\mathbf{x}_i', Y_i)_{1 \leq i \leq n}$, on peut bien entendu s'en servir pour faire de la prévision. Soit donc $\mathbf{x}_{n+1}' = [x_{n+1,1}, \cdots, x_{n+1,p}]$ une nouvelle valeur pour laquelle nous voudrions prédire Y_{n+1}. Cette variable réponse est définie par $Y_{n+1} = \mathbf{x}_{n+1}'\beta + \varepsilon_{n+1}$, avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ indépendant des $(\varepsilon_i)_{1 \leq i \leq n}$. La méthode naturelle est de prédire la valeur correspondante grâce au modèle ajusté, soit $\hat{Y}_{n+1}^{(p)} = \mathbf{x}_{n+1}'\hat{\beta}$. L'erreur de prévision est alors définie par

$$\hat{\varepsilon}_{n+1}^{(p)} = Y_{n+1} - \hat{Y}_{n+1}^{(p)} = \mathbf{x}_{n+1}'(\beta - \hat{\beta}) + \varepsilon_{n+1}.$$

Deux types d'erreurs vont alors entacher cette prévision : la première, incompressible, due à l'aléa de ε_{n+1}, l'autre à l'incertitude inhérente à l'estimateur $\hat{\beta}$, cette dernière décroissant typiquement avec le nombre n de données.

Attention ! La prévision $\hat{Y}_{n+1}^{(p)}$ et l'erreur de prévision $\hat{\varepsilon}_{n+1}^{(p)}$ ne jouent pas le même rôle que les valeurs ajustées $(\hat{Y}_i)_{1 \leq i \leq n}$ et les résidus $(\hat{\varepsilon}_i)_{1 \leq i \leq n}$, d'où la différence de notations.

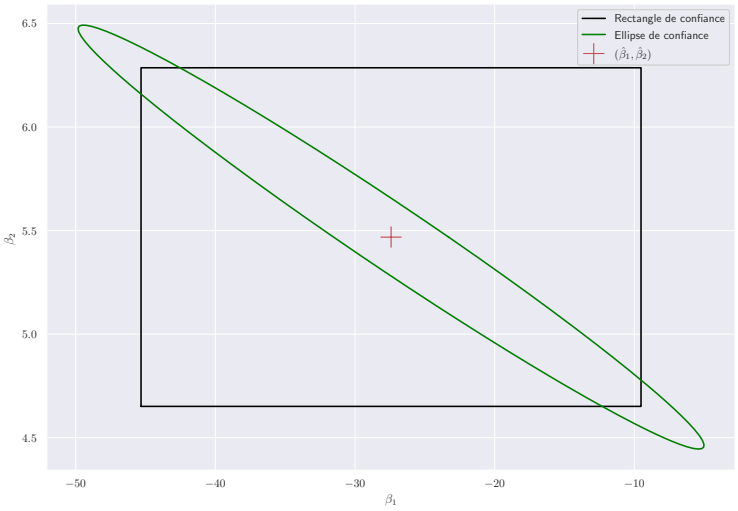

FIGURE 3.9 – Ellipse et rectangle de confiance à 95% pour $\beta = (\beta_1, \beta_2)$ sur l'exemple de l'ozone.

Proposition 21 (Erreur de prévision)
L'erreur de prévision $\hat{\varepsilon}_{n+1}^{(p)} = (Y_{n+1} - \hat{Y}_{n+1}^{(p)})$ *suit une loi normale, à savoir*

$$\hat{\varepsilon}_{n+1}^{(p)} \sim \mathcal{N}(0, \sigma^2(1 + \mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1})).$$

Preuve : Pour quantifier l'erreur de prévision $(Y_{n+1} - \hat{Y}_{n+1}^{(p)})$, on utilise la décomposition :

$$Y_{n+1} - \hat{Y}_{n+1}^{(p)} = \mathbf{x}'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1},$$

qui est la somme de deux variables gaussiennes indépendantes puisque $\hat{\beta}$ est construit à partir des $(\varepsilon_i)_{1 \leq i \leq n}$. On en déduit que $(Y_{n+1} - \hat{Y}_{n+1}^{(p)})$ est une variable gaussienne, dont il suffit de calculer moyenne et variance. Comme $\mathbb{E}[\varepsilon_{n+1}] = 0$ et puisque $\hat{\beta}$ est un estimateur sans biais de β, il est clair que

$$\mathbb{E}[\hat{\varepsilon}_{n+1}^{(p)}] = \mathbb{E}[\mathbf{x}'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}] = \mathbf{x}'_{n+1}(\beta - \mathbb{E}[\hat{\beta}]) + \mathbb{E}[\varepsilon_{n+1}] = 0.$$

Autrement dit, en moyenne, notre prévision ne se trompe pas. Calculons la variance de l'erreur de prévision. Puisque $\hat{\beta}$ dépend uniquement des variables aléatoires $(\varepsilon_i)_{1 \leq i \leq n}$, dont ε_{n+1} est indépendante, il vient :

$$\begin{aligned} \text{Var}\left(\hat{\varepsilon}_{n+1}^{(p)}\right) &= \text{Var}(\varepsilon_{n+1} + \mathbf{x}'_{n+1}(\beta - \hat{\beta})) = \sigma^2 + \mathbf{x}'_{n+1}\text{Cov}(\hat{\beta})\mathbf{x}_{n+1} \\ &= \sigma^2(1 + \mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1}). \end{aligned}$$

■

Nous retrouvons bien l'incertitude d'observation σ^2 à laquelle vient s'ajouter l'incertitude d'estimation. On peut montrer qu'en présence de la constante, cette incertitude est minimale au centre de gravité des variables explicatives, c'est-à-dire lorsque

$$\mathbf{x}'_{n+1} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p] = [1, \bar{x}_2, \ldots, \bar{x}_p],$$

et qu'elle vaut $\sigma^2(1 + 1/n)$. Ceci est facile à voir en régression linéaire simple : en effet, dans ce cas, en écrivant $\mathbf{x}'_{n+1} = [1, x]$, un calcul élémentaire montre que la variance de prédiction s'écrit encore

$$\text{Var}\left(\hat{\varepsilon}_{n+1}^{(p)}\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) \geq \sigma^2\left(1 + \frac{1}{n}\right),$$

avec égalité si et seulement si $x = \bar{x}$. Ainsi la variance augmente lorsque x_{n+1} s'éloigne du centre de gravité du nuage. Autrement dit, faire de la prévision lorsque x_{n+1} est "loin" de $\bar{x}$ est périlleux, puisque la variance de l'erreur de prévision peut être très grande ! Ceci s'explique intuitivement par le fait que plus une observation x_{n+1} est éloignée de la moyenne $\bar{x}$ et moins on a d'information sur elle.

Revenons au cadre de la Proposition 21. L'étape suivante consiste à préciser un intervalle de confiance pour $Y_{n+1} = \mathbf{x}'_{n+1}\beta + \varepsilon_{n+1}$. Comme d'habitude, le résultat de la Proposition 21 est inutilisable en l'état puisqu'il fait intervenir la variance σ^2, inconnue. Comme d'habitude, il suffit de la remplacer par son estimateur.

Proposition 22 (Intervalle de prédiction)
Un intervalle de confiance, dit intervalle de prédiction, de niveau $(1-\alpha)$ pour Y_{n+1} est donné par :

$$\left[\mathbf{x}'_{n+1}\hat{\beta} - t_\alpha\hat{\sigma}\sqrt{1+\mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1}}\ ;\ \mathbf{x}'_{n+1}\hat{\beta} + t_\alpha\hat{\sigma}\sqrt{1+\mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1}}\right],$$

où $t_\alpha = t_{n-p}(1-\alpha/2)$ est le quantile d'ordre $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-p}$.

Preuve : D'après ce qui a été dit auparavant, on a

$$\frac{Y_{n+1} - \hat{Y}^{(p)}_{n+1}}{\sigma\sqrt{1+\mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1}}} \sim \mathcal{N}(0,1).$$

En faisant intervenir $\hat{\sigma}$, il en découle naturellement

$$\frac{Y_{n+1} - \hat{Y}^{(p)}_{n+1}}{\hat{\sigma}\sqrt{1+\mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1}}} = \frac{\frac{Y_{n+1}-\hat{Y}^{(p)}_{n+1}}{\sigma\sqrt{1+\mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1}}}}{\frac{\hat{\sigma}}{\sigma}}.$$

Le numérateur suit une loi normale centrée réduite, le dénominateur est la racine d'un khi-deux à $(n-p)$ ddl divisé par $(n-p)$. Il reste à s'assurer que numérateur et dénominateur sont indépendants, or $Y_{n+1} - \hat{Y}^{(p)}_{n+1} = \mathbf{x}'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}$ et $\hat{\sigma}$ est indépendant à la fois de $\hat{\beta}$ (conséquence de Cochran, cf. Propriétés 1) et de ε_{n+1} (puisque $\hat{\sigma}$ ne dépend que des $(\varepsilon_i)_{1\leq i\leq n}$). On en conclut que

$$\frac{Y_{n+1} - \hat{Y}^{(p)}_{n+1}}{\hat{\sigma}\sqrt{1+\mathbf{x}'_{n+1}(X'X)^{-1}\mathbf{x}_{n+1}}} \sim \mathcal{T}_{n-p},$$

d'où se déduit l'intervalle de confiance de l'énoncé.

■

Dans le cadre de la régression linéaire simple mentionné ci-dessus, en notant $\hat{Y}^{(p)}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x$ la valeur prédite, ceci donne

$$\left[\hat{Y}^{(p)}_{n+1} - t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}; \hat{Y}^{(p)}_{n+1} + t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right].$$

on retrouve ainsi la remarque déjà faite : plus le point à prévoir admet pour abscisse x une valeur éloignée de $\bar{x}$, plus l'intervalle de confiance sera grand.

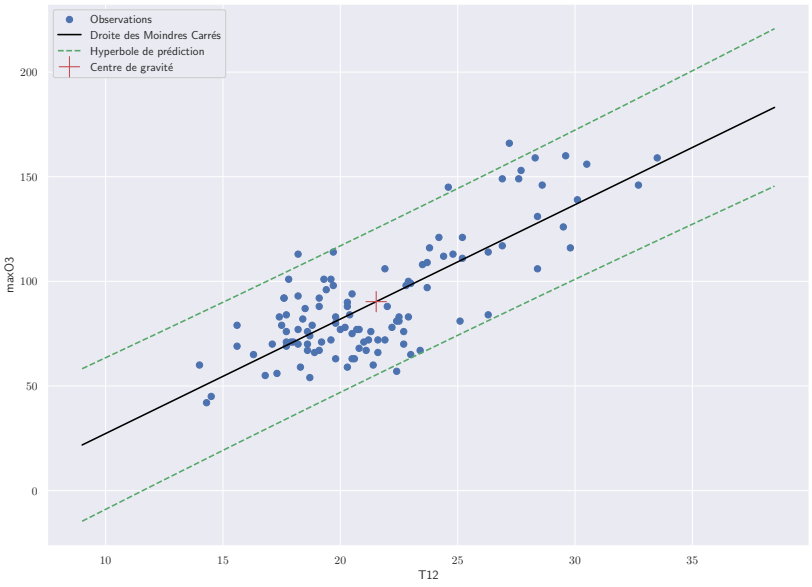

FIGURE 3.10 – Hyperbole de prédiction pour l'exemple de l'ozone.

Plus précisément, la courbe décrite par les limites de ces intervalles de confiance lorsque x_{n+1} varie est une hyperbole d'axes (non orthogonaux !) $y = \hat{\beta}_1 + \hat{\beta}_2 x$ (pour le nouvel axe des abscisses) et $x = \bar{x}$ (pour les ordonnées). Pour s'en persuader, il suffit d'effectuer le changement de variables

$$\begin{cases} X = x - \bar{x} \\ Y = y - (\hat{\beta}_1 + \hat{\beta}_2 x) \end{cases}$$

d'où il ressort qu'un point (X, Y) est dans la région de confiance ci-dessus si et seulement si

$$\frac{Y^2}{b^2} - \frac{X^2}{a^2} \leq 1,$$

avec

$$\begin{cases} a^2 = \left(1 + \frac{1}{n}\right) \sum (x_i - \bar{x})^2 \\ b^2 = \left(1 + \frac{1}{n}\right) (t_{n-2}(1 - \alpha/2)\hat{\sigma})^2 \end{cases}$$

ce qui définit bien l'intérieur d'une hyperbole. En particulier, le centre de cette hyperbole est tout bonnement le centre de gravité du nuage de points (voir Figure 3.10).

Remarque : La Figure 3.10 montre que la droite des moindres carrés passe par le centre de gravité G du nuage de points, c'est-à-dire que $\bar{Y}_n = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}$. Ceci est vrai de façon générale, dès lors que la constante fait partie du modèle, autrement dit typiquement lorsque la première colonne de la matrice X est composée de 1. En effet, dans ce cas, les vecteurs $Y - X\hat{\beta}$ et $\mathbf{1}$ sont orthogonaux, ce qui revient à dire que

$$\bar{Y}_n = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_1 + \cdots + \hat{\beta}_2 \bar{X}_n.$$

3.2.4 Estimateurs du Maximum de Vraisemblance

Dans le modèle gaussien, on peut faire le lien entre les estimateurs des moindres carrés $\hat{\beta}$ et $\hat{\sigma}^2$ et les estimateurs du maximum de vraisemblance. En Section 2.2.2, nous avons défini l'estimation au maximum de vraisemblance pour un paramètre θ réel. Ici le paramètre θ est le couple $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^\star$, mais le principe est rigoureusement le même : il s'agit de trouver le jeu de paramètres qui maximisent la vraisemblance des observations.

Rappelons que le vecteur Y des observations est gaussien : $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ avec $\sigma^2 I_n$ inversible. D'après la formule (3.3), il admet donc pour densité en un point y de $\mathbb{R}^n$

$$f(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\sigma^2 I_n)}} e^{-\frac{1}{2}(y-X\beta)'(\sigma^2 I_n)^{-1}(y-X\beta)} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right].$$

La vraisemblance de l'observation $Y = [Y_1, \ldots, Y_n]'$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ s'écrit donc

$$L_n(\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right].$$

D'où l'on déduit la log-vraisemblance

$$\ell_n(\beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\beta\|^2.$$

On cherche les estimateurs $\hat{\beta}_{mv}$ et $\hat{\sigma}^2_{mv}$ qui maximisent cette log-vraisemblance. Il est clair qu'il faut minimiser la quantité $\|Y - X\beta\|^2$, ce qui est justement le principe des moindres carrés ordinaires, donc

$$\hat{\beta}_{mv} = \hat{\beta} = (X'X)^{-1}X'Y.$$

Une fois ceci fait, on veut maximiser sur $\mathbb{R}_+^*$ une fonction de la forme $\varphi(x) = a + b\log x + \frac{c}{x}$, ce qui ne pose aucun souci en passant par la dérivée :

$$\frac{\partial \ell_n(\hat{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\hat{\beta}\|^2,$$

d'où il vient, si $Y \neq X\hat{\beta}$,

$$\hat{\sigma}^2_{mv} = \frac{\|Y - X\hat{\beta}\|^2}{n}.$$

Le cas très particulier où $Y = X\hat{\beta}$ revient à dire que $Y \in \mathcal{M}(X)$, auquel cas on convient de définir l'estimateur du maximum de vraisemblance par la même formule $\hat{\sigma}^2_{mv} = \|Y - X\hat{\beta}\|^2/n = 0$. Quoi qu'il en soit, si l'on compare à l'estimateur $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|^2/(n-p)$ obtenu précédemment, nous avons donc :

$$\hat{\sigma}^2_{mv} = \frac{n-p}{n}\hat{\sigma}^2.$$

On en déduit que l'estimateur $\hat{\sigma}^2_{mv}$ du maximum de vraisemblance est biaisé, mais d'autant moins que le nombre de variables explicatives est petit devant le nombre n d'observations.

Remarque : Historiquement, le premier estimateur étudié n'est pas celui des moindres carrés mais celui des "moindres déviations" (Least Absolute Deviations), introduit par Boscovich (1757) et analysé par Laplace (1793). En régression linéaire simple, il revient à chercher la droite qui minimise la somme des distances verticales (et non leurs carrés) entre celle-ci et les points de l'échantillon. On peut facilement l'interpréter en terme d'estimation au maximum de vraisemblance comme suit : considérons le même modèle que ci-dessus mais en supposant les erreurs de modélisation ε_i i.i.d. selon une loi de Laplace centrée et de de variance σ^2, c'est-à-dire qu'elles ont pour densité

$$f(t) = \frac{1}{\sqrt{2}\sigma}\exp\left(-\frac{\sqrt{2}}{\sigma}|t|\right).$$

Dans ce cas, les observations Y_i étant indépendantes et de densités $f_i(y_i) = f(y_i - \mathbf{x}_i'\beta)$, la vraisemblance s'écrit

$$L_n(\beta, \sigma) = \prod_{i=1}^n f(Y_i - \mathbf{x}_i'\beta) = \frac{1}{2^{n/2}\sigma^n}\exp\left(-\frac{\sqrt{2}}{\sigma}\sum_{i=1}^n |Y_i - \mathbf{x}_i'\beta|\right).$$

On voit que, dans ce modèle, l'estimateur $\hat{\beta}_{mv}$ du maximum de vraisemblance est la valeur de β qui minimise la quantité $\sum_{i=1}^{n} |Y_i - \mathbf{x}_i'\beta|$. Il présente l'avantage d'être plus robuste à d'éventuels outliers ou à une mauvaise spécification du modèle, mais l'inconvénient de ne pas être aussi simple que celui des moindres carrés : il n'a pas de formule explicite et nécessite de résoudre numériquement un problème d'optimisation. Face à ce constat, Legendre (1805) et Gauss (1823) ont proposé de privilégier l'estimateur des moindres carrés, lequel correspond donc, dans le cas d'erreurs gaussiennes, à l'estimateur du maximum de vraisemblance.

Bibliographie

[1] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics*. Prentice Hall, 1976.

[2] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons Inc., 3ème edition, 1995.

[3] Lucien Birgé. *Statistique mathématique*. Polycopié UPMC, 2014.

[4] Alexandr Alekseevich Borovkov. *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons Inc., 1991.

[6] Bernard Delyon. *Estimation paramétrique*. Format électronique, 2024.

[7] Benoît Cadre et Céline Vial. *Statistique mathématique - Master 1 et Agrégation*. Ellipses, 2012.

[8] Bernard Bercu et Djalil Chafaï. *Modélisation stochastique et simulation*. Dunod, 2007.

[9] Pierre-André Cornillon et Eric Matzner-Lober. *Régression avec R*. Springer, 2010.

[10] Vincent Rivoirard et Gilles Stoltz. *Statistique mathématique en action*. Vuibert, 2012.

[11] Jean Jacod et Philip Protter. *L'essentiel en théorie des probabilités*. Cassini, 2003.

[12] Andreï Kolmogorov et Sergeï Fomine. *Eléments de la théorie des fonctions et de l'analyse fonctionnelle*. Ellipses, Mir, 3ème edition, 1994.

[13] Dominique Fourdrinier. *Statistique inférentielle*. Dunod, 2002.

[14] Michel Lejeune. *Statistique - La théorie et ses applications*. Springer, 2005.

[15] Christian Robert. *Le choix bayésien*. Springer, 2010.

[16] Mark J. Schervish. *Theory of Statistics*. Springer-Verlag, 1995.

[17] Jun Shao. *Mathematical Statistics - Exercises and Solutions*. Springer, 2005.

[18] Larry Wasserman. *All of Statistics - A Concise Course in Statistical Inference*. Springer, 2004.

[19] Jan Wretman. A Simple Derivation of the Asymptotic Distribution of a Sample Quantile. *Scand. J. Statist.*, 5(2) :123–124, 1978.