Problèmes inverses — partie 2
Estimation statistique par maximum de vraisemblance

Nicolas Champagnat

26 mai 2020

Table des matières

Version provisoire, merci de me signaler les erreurs et les fautes de frappe !

1 Introduction

Ce cours est la seconde partie d'un cours portant sur les **problèmes inverses**. La première partie, par Takéo Takahashi, portait sur la théorie des problèmes inverses du point de vue des EDP (équations aux dérivées partielles). Il s'agit de retrouver les paramètres de l'EDP à partir d'une observation (complète ou partielle) d'une ou plusieurs solution de l'EDP. C'est la compréhension classique du terme « problèmes inverses ». Cependant, la question de reconstruction de paramètres à partir d'observation se pose également en statistiques. Il s'agit d'**inférence statistique**, qui peut être vue comme un problème inverse.

Ce cours se concentre sur les méthodes d'**estimation paramétrique** basées sur la méthode du **maximum de vraisemblance**. Il s'organise comme suit :

1. Contexte et rappels
2. Généralités sur l'estimation paramétrique
3. Estimation par maximum de vraisemblance
4. Algorithme EM

La troisième partie est la plus développée et constitue le sujet principal de ce cours. Il couvrira notamment les questions de consistance (information de Kullback-Leibler), de normalité asymptotique (information de Fisher), efficacité asymptotique (borne de Cramer-Rao) et test de Wald.

Ce cours est inspiré du plusieurs sources, notamment

[1] B. Cadre et C. Vial. *Statistique Mathématique. Cours & Exercices Corrigés*. Ellipse, 2012.

[2] V. Rivoirard et G. Stoltz. *Statistique Mathématique en Action*. Vuibert, 2012.

Les preuves difficiles sont écrites en petits caractères et peuvent être ignorées en première lecture.

2 Contexte, « rappels » et propriétés de base en probabilités et statistiques

On supposera connues les notions de base en probabilités et statistiques. Cette section a pour but de les préciser et rappeler brièvement. En particulier, les preuves sont omises.

2.1 Rappels de probabilités

2.1.1 Mesure et intégration

On munit $\mathbb{R}^k$ de sa tribu borélienne $\mathcal{B}(\mathbb{R}^k)$ (c'est-à-dire la tribu engendrée par les ouverts de $\mathbb{R}^k$). Pour μ mesure positive sur $\mathbb{R}^k$, rappelons la notation

$$\mathbb{L}^p(\mu) = \left\{ f : \mathbb{R}^k \to R \text{ tel que } \int_{\mathbb{R}^k} |f|^p \, d\mu < +\infty \right\}.$$

La notion de mesure produit sera fondamentale dans la suite.

Définition 2.1 (mesure produit) — *Soit μ_1 et μ_2 deux mesures de probabilité sur $\mathbb{R}^k$ et $\mathbb{R}^d$, respectivement. On définit et on note $\mu_1 \otimes \mu_2$ la **mesure produit** de μ_1 et μ_2 comme l'unique mesure de probabilité sur $\mathbb{R}^{k+d}$ telle que*

$$\mu_1 \otimes \mu_2(A \times B) = \mu_1(A)\mu_2(B), \quad \forall A \in \mathcal{B}(\mathbb{R}^k) \text{ et } B \in \mathcal{B}(\mathbb{R}^d).$$

— *Soit μ une mesure de probabilité sur $\mathbb{R}^k$ et $n \in \mathbb{N}$. On note*

$$\mu^{\otimes n} = \underbrace{\mu \otimes \ldots \otimes \mu}_{n \text{ fois}}.$$

Rappelons quatre résultats de base de la théorie de l'intégration.

Théorème 2.2 (convergence dominée) *Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions boréliennes de $\mathbb{R}^k$ dans $\mathbb{R}$ qui converge μ-presque partout pour une certaine mesure μ sur $\mathbb{R}^k$. S'il existe $g \in \mathbb{L}^1(\mu)$ telle que $|f_n| \leq g$ μ-presque partout pour tout $n \in \mathbb{N}$, alors*

$$\int_{\mathbb{R}^k} \lim_{n\to+\infty} f_n \, d\mu = \lim_{n\to+\infty} \int_{\mathbb{R}^k} f_n \, d\mu.$$

Théorème 2.3 (continuité sous le signe somme) *Soit $f : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$ tel que $f(x, \cdot) \in \mathbb{L}^1(\mu)$ pour tout $x \in \mathbb{R}$, $f(\cdot, y)$ est continue pour μ-presque tout $y \in \mathbb{R}^k$ et, pour tout $x \in \mathbb{R}$, il existe $g \in \mathbb{L}^1(\mu)$ et $V \subset \mathbb{R}$ un voisinage de x tel que*

$$\sup_{z\in V} |f(z, \cdot)| \leq g.$$

Alors l'application

$$x \mapsto \int_{\mathbb{R}^k} f(x,y)\,\mu(dy)$$

est continue sur $\mathbb{R}$.

Théorème 2.4 (dérivation sous le signe somme) *Soit* $f : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$ *tel que* $f(x,\cdot) \in \mathbb{L}^1(\mu)$ *pour tout* $x \in \mathbb{R}$, $f(\cdot,y) \in \mathcal{C}^1$ *pour* μ*-presque tout* $y \in \mathbb{R}^k$ *et, pour tout* $x \in \mathbb{R}$, *il existe* $g \in \mathbb{L}^1(\mu)$ *et* $V \subset \mathbb{R}$ *un voisinage de* x *tel que*

$$\sup_{z \in V} \left| \frac{\partial}{\partial x} f(z,\cdot) \right| \leq g.$$

Alors l'application $x \mapsto \int_{\mathbb{R}^k} f(x,y)\,\mu(dy)$ *est* $\mathcal{C}^1$ *et, pour tout* $x \in \mathbb{R}$,

$$\frac{d}{dx} \int_{\mathbb{R}^k} f(x,y)\,\mu(dy) = \int_{\mathbb{R}^k} \frac{\partial}{\partial x} f(x,y)\,\mu(dy).$$

Définition 2.5 *On dit qu'une mesure* ν *sur* $\mathbb{R}^k$ *est* ***absolument continue*** *par rapport à la mesure* μ *sur* $\mathbb{R}^k$, *et on note* $\nu \ll \mu$, *ssi pour tout* $A \in \mathcal{B}(\mathbb{R}^k)$ *tel que* $\mu(A) = 0$, *on a* $\nu(A) = 0$.

Théorème 2.6 (Radon-Nikodym) *Soit* ν *mesure sur* $\mathbb{R}^k$ *telle que* $\nu \ll \mu$. *Alors il existe* $f : \mathbb{R}^k \to \mathbb{R}$ *mesurable telle que* $\nu(A) = \int_A f\,d\mu$ *pour tout* $A \in \mathcal{B}(\mathbb{R}^k)$. *La fonction* f *est unique* μ*-presque partout et* $f \in \mathbb{L}^1(\mu)$ *si* ν *est finie (c'est-à-dire* $\nu(\mathbb{R}^k) < +\infty$*). On appelle* f *la* ***densité*** *de* ν *par rapport à* μ. *Si* μ *est la mesure de Lebesgue sur* $\mathbb{R}^k$, f *est simplement appelée* ***densité de*** ν.

Enfin, rappelons la notion de convergence étroite (ou faible au sens des mesures).

Définition 2.7 (convergence étroite) *Une suite* $(\mu_n)_{n\in\mathbb{N}}$ *de mesures de probabilité sur* $\mathbb{R}^k$ ***converge étroitement*** *vers la mesure de probabilité* μ *sur* $\mathbb{R}^k$, *ce qu'on note* $\mu_n \Rightarrow \mu$, *ssi pour tout* $f : \mathbb{R}^k \to \mathbb{R}$ *continue bornée,*

$$\lim_{n\to+\infty} \int_{\mathbb{R}^k} f\,d\mu_n = \int_{\mathbb{R}^k} f\,d\mu.$$

Il existe diverses caractérisations de la convergence étroite. La suivante est utile.

Théorème 2.8 (Portmanteau) $\mu_n \Rightarrow \mu$ *ssi* $\liminf_{n\to+\infty} \mu_n(O) \geq \mu(O)$ *pour tout ouvert* $O \subset \mathbb{R}^k$.
De plus, si la mesure limite μ *est absolument continue par rapport à la mesure de Lebesgue, alors* $\mu_n \Rightarrow \mu$ *ssi* $\lim_{n\to+\infty} \mu_n(O) = \mu(O)$ *pour tout ouvert* $O \subset \mathbb{R}^k$, *ssi* $\lim_{n\to+\infty} \mu_n(F) = \mu(F)$ *pour tout fermé* $F \subset \mathbb{R}^k$.

2.1.2 Variables aléatoires, loi, indépendance

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.

Définition 2.9 (variable aléatoire, loi) — *On dit que X est une **variable aléatoire** (abrégé en* v.a.*) sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}^k$ si $X : \Omega \to \mathbb{R}^k$ est mesurable.*

— *La **loi de la v.a.** X est la mesure de probabilité sur $\mathbb{R}^k$ notée $\mathbb{P}_X$ définie pour tout $A \in \mathcal{B}(\mathbb{R}^k)$ par*

$$\mathbb{P}_X(A) = \mathbb{P}\{\omega \in \Omega \text{ tels que } X(\omega) \in A\}.$$

Définition 2.10 (espérance, variance) *Soit X une v.a. à valeurs dans $\mathbb{R}^k$.*

— *Si $p \in [1, +I)$, on note $X \in \mathbb{L}^p = \mathbb{L}^p(\mathbb{P})$ si $\int_\Omega |X(\omega)|^p \mathbb{P}(d\omega) < +\infty$.*

— *Si $g : \mathbb{R}^k \to \mathbb{R}^d$ est telle que $g(X) \in \mathbb{L}^1$, on appelle **espérance** de $g(X)$, et on note $\mathbb{E}(g(X)) = \int_\Omega g(X)\, d\mathbb{P}$.*

— *Si $X \in \mathbb{L}^2$, on appelle **matrice de variance-covariance** de X (ou simplement **variance** en dimension $k = 1$) et note*

$$\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}(X))(X - \mathbb{E}(X)^T),$$

où M^T est la matrice transposée de M et où nous utilisons, ici et dans toute la suite, la convention que les vecteurs de $\mathbb{R}^k$ sont des vecteurs colonne.

Rappelons également qu'on appelle **écart-type** d'une v.a. réelle la racine carrée de sa variance.

Définition 2.11 (indépendance) — *Si X et Y sont des v.a. à valeurs dans $\mathbb{R}^k$ et $\mathbb{R}^d$ respectivement, on dit que X et Y sont **indépendantes** ssi $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y$, c'est-à-dire si*

$$\mathbb{P}((X,Y) \in A \times B) = \mathbb{P}(X \in A)\mathbb{P}(X \in B), \quad \forall A \in \mathcal{B}(\mathbb{R}^k),\ B \in \mathcal{B}(\mathbb{R}^d),$$

ou, de façon équivalente, si pour toutes fonctions mesurables bornées $f : \mathbb{R}^k \to \mathbb{R}$ et $g : \mathbb{R}^k \to \mathbb{R}$,

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)).$$

— *Les v.a. $X_1, \ldots, X_n$ à valeurs dans $\mathbb{R}^{k_1}, \ldots, \mathbb{R}^{k_n}$ respectivement sont dites **indépendantes** ssi $\mathbb{P}_{(X_1,\ldots,X_n)} = \mathbb{P}_{X_1} \otimes \ldots \otimes \mathbb{P}_{X_n}$.*

— *Les v.a. $X_1, \ldots, X_n$ à valeurs dans $\mathbb{R}^k$ sont dites **indépendantes et identiquement distribuées** (abrégé en* i.i.d.*) de loi μ ssi la loi du vecteur $(X_1, \ldots, X_n)$ est $\mu^{\otimes n}$. On a alors en particulier que $\mu = \mathbb{P}_{X_i}$ pour tout i.*

Rappelons une première conséquence de l'indépendance.

Proposition 2.12 *Si les v.a. $X_1, \ldots, X_n$ à valeurs dans $\mathbb{R}^k$ sont indépendantes et $\mathbb{L}^2$, alors*

$$\mathbb{V}(X_1 + \ldots + X_n) = \sum_{i=1}^{n} \mathbb{V}(X_i).$$

2.1.3 Quelques inégalités

On rappelle quelques inégalités utiles sur les v.a.

Inégalité de Markov Si $X \in \mathbb{L}^1$, pour tout $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t}.$$

Inégalité de Tchebytchev Si $X \in \mathbb{L}^2$, pour tout $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{V}|X|}{t^2}.$$

Inégalité de Cauchy-Schwarz Si $X, Y \in \mathbb{L}^2$,

$$(\mathbb{E}XY)^2 \leq (\mathbb{E}X^2)(\mathbb{E}Y^2).$$

De plus, l'inégalité précédente est une égalité ssi $\exists C \in \mathbb{R}$ telle que $X = CY$ presque sûrement (c'est-à-dire avec $\mathbb{P}$-probabilité 1).

Inégalité de Jensen Si $X \in \mathbb{L}^1$ est une v.a. à valeurs réelles et $\psi : \mathbb{R} \to \mathbb{R}$ est convexe, alors

$$\psi(\mathbb{E}(X)) \leq \mathbb{E}\psi(X).$$

Si de plus ψ est strictement convexe, l'inégalité précédente est une égalité ssi $\exists C \in \mathbb{R}$ telle que $X = C$ presque sûrement.

2.1.4 Convergence stochastique

Soit $(Z_n)_{n\in\mathbb{N}}$ une suite de v.a. à valeurs dans $\mathbb{R}^k$ sur une suite d'espaces de probabilités $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$.

Définition 2.13 (convergence en probabilité, en loi) — *On dit que Z_n **converge en probabilité** vers une constante $z \in \mathbb{R}^k$, et on note $Z_n \xrightarrow{\mathbb{P}_n} z$, si $\forall \varepsilon > 0$, $\lim \mathbb{P}_n(|Z_n - z| > \varepsilon) = 0$.*

— *On dit que Z_n **converge en loi** vers la probabilité μ (resp. la v.a. Z de loi μ), et on note $Z_n \xrightarrow{loi} \mu$ (resp. $Z_n \xrightarrow{loi} Z$) si la suite des lois de Z_n converge étroitement vers μ, c'est-à-dire si, pour toute fonction $f : \mathbb{R}^k \to \mathbb{R}$ continue bornée, $\mathbb{E}_n f(X_n) \to \int_{\mathbb{R}^k} f \, d\mu = EE(Z)$.*

Si de plus tous les espaces de probabilité $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ sont identiques, égaux à $(\Omega, \mathcal{F}, \mathbb{P})$, on a la définition suivantes.

Définition 2.14 (convergence presque sûre) *On dit que Z_n* ***converge presque sûrement*** *vers la v.a. Z sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans RR^k, et on note $Z_n \to Z$ p.s., si $\mathbb{P}(\lim_{n\to+\infty} Z_n = Z) = 1$.*

Les résultats suivants relient les différentes notions de convergence.

Proposition 2.15
- *Toutes ces notions de convergence sont préservées par la composition par une fonction continue (c'est-à-dire que Z_n converge vers Z dans l'un des sens ci-dessus, alors $f(Z_n)$ converge vers $f(Z)$ dans le même sens pout tout f continue).*
- *La convergence presque sûre de Z_n vers Z implique la convergence en probabilité de $Z_n - Z$ vers 0, qui implique elle-même la convergence en loi de Z_n vers Z.*
- *Si $Z = C$ p.s. pour une certaine constante C, alors la convergence en loi de Z_n vers Z est équivalente à la convergence en probabilité de Z_n vers Z.*

Voici le premier théorème fondamental des probabilités, qui nous sera d'une grande utilité dans la suite.

Théorème 2.16 (Loi forte des grands nombres (LGN)) *Soit $(X_i)_{\in\mathbb{N}}$ une suite de v.a. i.i.d. sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}^k$ et de même loi μ telle que $\int_{\mathbb{R}^k} |x| \, \mu(dx) < +\infty$. Alors*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}(X_1) = \int_{\mathbb{R}^k} x \, \mu(dx)$$

et la convergence a même lieu presque sûrement.

Le lemme suivant s'intéresse à la convergence de couples de v.a.

Lemme 2.17 (Slutsky) *Soit $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ deux suites de v.a. à valeurs dans $\mathbb{R}^k$ et $\mathbb{R}^d$, respectivement, telles que X_n converge en loi vers une v.a. $X \in \mathbb{R}^k$ et Y_n converge en probabilité vers une constante $y \in \mathbb{R}^d$. Alors (X_n, Y_n) converge en loi vers (X, y).*

Voici le second théorème fondamental des probabilités.

Théorème 2.18 (théorème central limite (TCL)) *Considérons $(X_i)_{\in\mathbb{N}}$ une suite de v.a. i.i.d. sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}^k$ et de même loi μ telle que $\int_{\mathbb{R}^k} |x|^2 \, \mu(dx) < +\infty$. On note*

$$m = \int_{\mathbb{R}^k} x \, \mu(dx) \in \mathbb{R}^k \quad et \quad \Sigma = \mathbb{V}(\mu) = \int_{\mathbb{R}^k} (x-m)(x-m)^T \, \mu(dx)$$

sa matrice de variance-covariance. Alors

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - m\right) \xrightarrow{loi} \mathcal{N}_k(0,\Sigma),$$

où $\mathcal{N}_k(0,\Sigma)$ désigne la loi gaussienne de dimension k (cf. section 2.1.5 ci-dessous), de moyenne nulle et de matrice de variance-covariance Σ.

2.1.5 Vecteurs gaussiens

Définition 2.19 (vecteurs gaussiens) *On dit qu'un v.a. X à valeurs dans $\mathbb{R}^d$ est un **vecteur gaussien** s'il exite un vecteur $m \in \mathbb{R}^d$ et un matrice $d\times d$ symétrique positive Σ tels que*

$$\mathbb{E}e^{i\langle u,X\rangle} = \exp\left(i\langle u,m\rangle - \frac{1}{2}u^T\Sigma u\right), \quad \forall u \in \mathbb{R}^d,$$

où $\langle\cdot,\cdot\rangle$ désigne le produit scalaire usuel dans $\mathbb{R}^d$. On note $\mathcal{N}_d(m,\Sigma)$ la loi de X.

Lorsque $d = 1$, la matrice Σ n'a qu'un seul élément $\sigma^2 \in \mathbb{R}_+$, et on note $\mathcal{N}(m,\sigma^2)$ au lieu $\mathcal{N}_1(m,\Sigma)$.

Remarquons que, si $X \sim \mathcal{N}_d(m,\Sigma)$, alors, pour tout $u \in \mathbb{R}^d$, $\langle u, X\rangle$ suit la loi gaussienne en dimension 1 de moyenne $\langle u,m\rangle$ et de variance $u^T\Sigma u$. De plus

$$\mathbb{E}(X) = m \quad \text{et} \quad \mathbb{V}(X) = \Sigma.$$

Si la matrice Σ est inversible, alors X a pour densité (par rapport à la mesure de Lebesgue)

$$\frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right), \quad \forall x \in \mathbb{R}^d.$$

Dans le cas de la dimension $d = 1$, on retrouve la formule classique de la densité de la loi gaussienne $\mathcal{N}(m,\sigma^2)$ avec $m \in \mathbb{R}$ et $\sigma > 0$:

$$\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R}.$$

Voici quelques propriétés fondamentales des vecteurs gaussiens.

Proposition 2.20 (caractérisation d'un vecteur gaussien) *Un vecteur aléatoire X est gaussien ssi toute combinaison linéaire de ses composantes (c'est-à-dire toute v.a. de la forme $\langle u, X\rangle$ pour un vecteur u) est une v.a. réelle gaussienne.*

Proposition 2.21 *Si A est une matrice réelle $k \times d$, $b \in \mathbb{R}^k$ et $X \sim \mathcal{N}_d(m, \Sigma)$,*

$$AX + b \sim \mathcal{N}_k(Am + b, A\Sigma A^T).$$

Proposition 2.22 (caractérisation de l'indépendance) *Soit X un vecteur gaussien. Les composantes de X sont des v.a. réelles indépendantes ssi la matrice de variance-convariance de X est diagonale.*

2.2 Rappels de statistiques

Nous allons ici rappeler le vocabulaire usuel pour définir un problème statistique. Nous allons également introduire la notions de modèle statistique et donner le principe fondamental de la statistique.

2.2.1 Vocabulaire

Pour une étude statistique donnée, on appelle

Population l'ensemble des objets sur lesquels l'étude porte (par exemple l'ensemble des patients dans un essai clinique, ou la population française dans un sondage, ou l'ensemble des poissons d'un lac...) ;

Individu chaque objet dont est constituée la population (par ex. un patient dans l'essai clinique, ou un français, ou un poisson du lac...) ;

Caractère ou variable toute propriété des individus sur laquelle porte l'étude ; un caractère peut être **quantitatif** s'il prend ses valeurs dans $\mathbb{R}^k$ pour un certain $k \geq 1$ (par ex. la concentration d'insuline dans le sang de patients diabétiques, ou la longueur des nageoires des poissons d'un lac, ou les rendements du cours d'une action en bourse...), ou **qualitatif (ou catégoriel)** s'il prend ses valeurs dans un ensemble discret, c'est-à-dire fini ou dénombrable (par ex. $\{\text{malade}, \text{sain}\}$, ou $\{\text{vote A}, \text{vote B}\}$, ou $\{\text{pile}, \text{face}\}$...) ;

Échantillon un ensemble d'individus sélectionnés selon un certain processus parmi la population (par ex. tous les patients impliqués dans l'essai clinique, ou 1000 français choisis aléatoirement pour être sondés, ou les n premiers poissons pêchés dans le lac) ; si l'échantillon est de taille n, on parle de **n-échantillon** ;

Processus de sélection l'échantillon est obtenu avec un certain processus de sélection des individus dans la population. Il peut par exemple être obtenu en considérant tous les individus de la population (par ex. dans un essai clinique), ou en effectuant un tirage aléatoire sans remise dans la population (par ex. pour un sondage, sans remise signifiant ici qu'un individu ne peut être sondé qu'une seule fois), ou en effectuant un tirage aléatoire avec remise dans la population (par ex. pour les poissons du lac, si on suppose que les poissons pêchés

sont relâchés après avoir mesuré la longueur de leurs nageoires). Une sélection aléatoire permet de supposer une propiété d'indépendance des caractères observés dans l'échantillon. D'autres processus de sélection sont également possibles, par exemple si la concentration d'insuline est mesurée à plusieurs dates rapprochées chez un seul patient. Dans ce cas il n'est pas raisonnable de supposer les mesures indépendantes.

Observations ou données le vecteur $(x_1, \ldots, x_n)$ des caractères observés dans le n-échantillon (par ex. le vecteur des concentrations d'insuline de tous les patients impliqués dans l'essai clinique, ou bien l'opinion des 1000 français participant au sondage, ou bien la longueur des nageoires des n premiers poissons pêchés dans le lac).

Remarquons qu'on donne également dans la théorie de l'estimation paramétrique le nom d'**échantillon** à une variable aléatoire dont les observations sont une réalisation (cf. section 3.1). A partir de la section 3 de ce document, le mot échantillon sera toujours entendu avec ce second sens.

Ce cours porte sur **l'estimation paramétrique**, qui consiste à évaluer des paramètres inconnus liés aux caractères considérés dans la population (par exemple leur moyenne ou leur écart-type) à partir des observations $x_1, \ldots, x_n$. On verra que cette question d'estimation peut être abordée de trois manières différentes :

- **estimation ponctuelle**, c'est-à-dire l'estimation directe de la valeur des paramètres inconnus ;
- **estimation par intervalles de confiance**, c'est-à-dire la définition de bornes entre lesquelles les paramètres inconnus se trouvent avec un certain niveau de confiance ;
- **tests statistiques sur les paramètres**, c'est-à-dire la validation ou le rejet d'hypothèses sur les paramètres inconnus avec un certain seuil de confiance.

2.2.2 Modélisation : un exemple

L'exemple développé dans cette section sera suivi et développé tout le long de ce cours. Il servira de ≪ fil rouge ≫ pour illustrer au fur et à mesure les notions et les résultats présentés.

Nous considérons l'exemple du jeu de pile-ou-face répété n fois avec la même pièce. Le but de l'étude statistique est d'évaluer les caractéristiques de la pièce. La population est ici l'ensemble des n lancers de la pièce. C'est également le n-échantillon. Les variables sont le résultat d'un lancer, qu'on notera 0 pour pile et 1 pour face. Les observations sont donc le vecteur $(x_1, \ldots, x_n) \in \{0, 1\}^n$ qui décrit la suite des résultats.

Afin de pouvoir estimer les caractéristiques de la pièce, il est nécessaire d'introduire une modélisation du jeu de pile-ou-face, c'est-à-dire de la loi des

observations. Cela signife que l'on suppose que le vecteur des observations $(x_1, \dots, x_n)$ est la réalisation d'une certaine v.a. $(X_1, \dots, X_n)$ sur un certain espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$, v.a. qui décrit le résultat aléatoire du jeu de pile-ou-face. Autrement dit,

$$(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega)) \text{ pour un certain } \omega \in \Omega.$$

Ici, il est naturel de supposer que les v.a. $X_1, \dots, X_n$ sont indépendantes (il n'y a pas de lien de dépendance entre les différents lancers de la pièce) et identiquement distribuées (la loi du résultat d'un tirage ne dépend pas du tirage considéré), ce que nous avons noté *i.i.d.* Puisque chaque v.a. X_i prend ses valeurs dans $\{0, 1\}$, elles suivent nécessairement une certaine loi de Bernoulli [1]. Le modèle statistique de cette étude est donc le vecteur aléatoire $(X_1, \dots, X_n)$ de loi $\text{Bern}(\theta)^{\otimes n}$ (cf. définition 2.1), où $\theta \in [0, 1]$ est le paramètre inconnu du modèle. Autrement dit, on suppose que les X_i sont i.i.d. (cf. définition 2.11) de loi $\text{Bern}(\theta)$ pour un certain $\theta \in [0, 1]$ inconnu.

Le **modèle statistique** associé à cette étude est donc donné par

$$\left(\{0, 1\}^n, \left\{\text{Bern}(\theta)^{\otimes n}\right\}_{\theta \in [0,1]}\right),$$

où $\{0, 1\}^n$ est **l'ensemble des valeurs possibles de la v.a.** X, θ est le **paramètre du modèle**, $[0, 1]$ **l'ensemble des valeurs du paramètre**, et $\text{Bern}(\theta)^{\otimes n}$ **la loi des observations lorsque le paramètre vaut** θ.

Le problème d'estimation paramétrique consiste ici, à partir des observations $(x_1, \dots, x_n) \in \{0, 1\}^n$, à identifier la valeur réelle θ_0 du paramètre inconnu θ. Pour cela, nous pouvons par exemple utiliser la loi des grands nombres (théorème 2.16) : supposons que le vecteur aléatoire $(X_1, \dots, X_n)$ suit la loi $\text{Bern}(\theta_0)^{\otimes n}$, alors

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n \to +\infty]{} \mathbb{E}(X_1) = \theta_0 \quad \text{p.s.} \tag{1}$$

Remarquons que la taille de l'échantillon ne peut pas a priori converger vers l'infini, mais la convergence précédente peut être considérée comme une bonne approximation lorsque n est grand. Ceci suggère en particulier que, pour la réalisation $(x_1, \dots, x_n)$ de $(X_1, \dots, X_n)$, on s'attend à ce que

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \approx \theta_0. \tag{2}$$

1. Rappelons que X suit une loi de Bernoulli de paramètre $p \in [0, 1]$, ce que l'on note $X \sim \text{Bern}(p)$, ssi X est à valeurs dans $\{0, 1\}$ et

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

Remarquons cependant que la convergence dans l'équation (1) a seulement lieu *presque sûrement*. Ainsi, rien ne garantit que le $\omega \in \Omega$ qui détermine les observations $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$ pourrait très bien être tel que la convergence dans (1) n'a pas lieu ou que la limite est différente de θ_0. Cependant, l'ensemble des tels ω a une probabilité nulle, ce qui signife que le fait que l'approximation (2) soit fausse est hautement improbable. Mais il faut donner un sens précis aux mots « fausse » et « hautement imporbable » dans la phrase précédente.

Il est donc nécessaire de quantifier l'approximation (2), afin par exemple de déterminer le nombre de lancers de pile-ou-face (c'est-à-dire la valeur de n) nécessaires afin que l'erreur dans l'approximation ci-dessus soit plus petite qu'un seuil donné. On calcule donc le **risque quadratique** à l'aide de la proposition 2.12 :

$$\mathbb{E}[(\bar{X}_n - \theta_0)^2] = \mathbb{V}(\bar{X}_n) = \frac{1}{n^2}\mathbb{V}(X_1 + \ldots + X_n) = \frac{1}{n}\mathbb{V}(X_1) = \frac{1}{n}\theta_0(1-\theta_0) \leq \frac{1}{4n},$$

où la dernière borne est indépendante de la valeur exacte de θ_0. D'après l'inégalité de Tchebytchev, pour tout $\varepsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \theta_0| \geq \varepsilon) \leq \frac{\mathbb{V}(\bar{X}_n)}{\varepsilon^2} \leq \frac{1}{4\varepsilon^2 n}.$$

Cette inégalité permet de construire un **intervalle de confiance** pour la valeur de θ_0 (cf. section 3.5 pour une définition précise).

Supposons par exemple que $n = 1000$ et que l'on veuille construire un intervalle de confiance à 95%. On résout l'équation $1/(4\varepsilon^2 n) = 1 - 95\% = 0,05$, ce qui donne $\varepsilon = 0,08$. On en déduit que

$$\mathbb{P}\left(\theta_0 \in [\bar{X}_n - 0,08\,,\, \bar{X}_n + 0,08]\right) \geq 0,95.$$

On en déduit que l'intervalle $[\bar{x}_n - 0,08\,,\, \bar{x}_n + 0,08]$, construit sur les observations $(x_1, \ldots, x_n)$, est un intervalle de confiance à 95% pour le paramètre inconnu θ.

2.2.3 Modèles statistiques

Formalisons la discussion de l'exemple précédent. On considérera ici la notion suivante de modèle statistique, correspondant au cas où les données sont obtenues à partir d'un n-échantillon.

Définition 2.23 *Un **modèle statistique** est un couple*

$$(\mathcal{H}^n, \mathcal{P}),$$

où $\mathcal{H} \in \mathcal{B}(\mathbb{R}^k)$ pour un certain $k \geq 1$ et $\mathcal{P}$ est une famille de mesures de probabilité sur $(\mathcal{H}^n, \mathcal{B}(\mathcal{H}^n))$

Remarque 2.24 (observations i.i.d.) *Dans le cas d'un processus de sélection par tirage aléatoire avec remise, ou bien par tirage aléatoire sans remise mais avec une population beaucoup plus grande que n, on suppose généralement que*

$$\mathcal{P} = \{\mathbb{Q}^{\otimes n}\}_{\mathbb{Q}\in\mathcal{Q}},$$

où $\mathcal{Q}$ est une famille de mesures de probabilités sur $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. Ceci correspond à l'hypothèse que les caractéristiques des individus de l'échantillon sont i.i.d. Dans l'exemple du jeu de pile ou face de la section précédente, on avait

$$\mathcal{H} = \{0, 1\} \quad et \quad \mathcal{Q} = \{Bern(\theta)\}_{\theta\in[0,1]}.$$

La nature de l'expérience statistique contraint la famille des modèles.

Exemple : premier pile au jeu de pile-ou-face On considère un joueur de pile-ou-face qui répète n fois l'expérience suivante : il lance sa pièce autant de faois que nécessaire jusqu'à obtenir pile une fois, et il note le nombre de tirages effectués. Si on suppose (comme dans la section précédente) que les résultats des lancers de pile-ou-face sont i.i.d., on obtient le modèle statistique suivant :

$$\left((\mathbb{N}^*)^n, \left\{\mathrm{Geom}(\theta)^{\otimes n}\right\}_{\theta\in[0,1]}\right),$$

où $\mathrm{Geom}(p)$ est la loi géométrique de paramètre $p \in [0, 1]$, définie par $X \sim \mathrm{Geom}(p)$ si X est à valeurs dans $\mathbb{N}^*$ et $\mathbb{P}(X = k) = (1 - p)p^{k-1}$ pour tout $k \geq 1$.

On distingue plusieurs types de modèles statistiques en fonction de leur complexité.

Définition 2.25 *Le modèle statistique $(\mathcal{H}^n, \mathcal{P})$ est un **modèle statistique paramétré** par Θ si $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$. C'est un **modèle statistique paramétrique** si $\Theta \subset \mathbb{R}^d$ pour un certain $d \geq 1$. Sinon, le modèle est **non-paramétrique**.*

La construction d'un modèle statistique n'est satisfisante que s'il n'y a pas de redondance dans l'ensemble des lois du modèle. La notion d'identifiabilité formalise cette contrainte.

Définition 2.26 (identifiabilité) *Un modèle statistique $(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$ paramétré est **identifiable** si $\theta \mapsto \mathbb{P}_\theta$ est injective sur Θ.*

Exemple Le modèle stastistique gaussien

$$\left(\mathbb{R}^n, \{\mathcal{N}(m, \sigma^2)^{\otimes n}\}_{m\in\mathbb{R},\sigma\in\mathbb{R}}\right)$$

n'est pas identifiable car $\mathcal{N}(m, \sigma^2)$ est identique pour $\sigma = 1$ et $\sigma = -1$. En revanche, le modèle

$$\left(\mathbb{R}^n, \{\mathcal{N}(m, \sigma^2)^{\otimes n}\}_{m\in\mathbb{R}, \sigma\in[0,+\infty)}\right)$$

est identifiable.

2.2.4 Principe fondamental de la statistique

Le problème fondamental des statistiques est la reconstruction de la loi des observations à partir de $x_1, \ldots, x_n$. Il existe un principe général qui garantit la faisabilité théorique dans la limite $n \to +\infty$ de ce problème, dans le cas d'observations i.i.d.

Supposons que l'on dispose d'une suite infinie d'observations $x_1, x_2, \ldots$ réalisation d'une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ i.i.d. à valeurs dans $\mathbb{R}^k$. Le problème se reformule comme la reconstruction d'une loi inconnue $\mathbb{Q}$ sur $\mathbb{R}^k$ à partir de la suite $(X_n)_{n\in\mathbb{N}}$ i.i.d. de loi $\mathbb{Q}$.

On introduit la **mesure empirique**

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

où δ_x est la mesure de Dirac en x, c'est-à-dire la mesure que donne un masse 1 au point x et 0 ailleurs. La mesure empirique est un candidat prometteur pour reconstruire la loi inconnue $\mathbb{Q}$ puisque, en vertu de la loi des grands nombres (théorème 2.16), pour tout $A \in \mathcal{B}(\mathbb{R}^k)$,

$$\lim_{n\to+\infty} Q_n(A) = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(X_i) = \mathbb{P}(X_1 \in A) = \mathbb{Q}(A) \quad \text{p.s.}, \qquad (3)$$

où on rappelle que la notation $\mathbb{1}_A$ désigne la fonction indicatrice de A, telle que $\mathbb{1}_A(x) = 1$ si $x \in A$ et $\mathbb{1}_A(x) = 0$ sinon.

Cependant, on ne peut pas intervertir la limite p.s. et le $\forall A \in \mathcal{B}(\mathbb{R}^k)$. Il est donc nécessaire de préciser la convergence de Q_n vers $\mathbb{Q}$.

Théorème 2.27 (Varadarajan) *Avec probabilité 1, la suite de mesures Q_n converge étroitement vers $\mathbb{Q}$:*

$$Q_n \Rightarrow \mathbb{Q} \quad \textit{quand } n \to +\infty \textit{ p.s.}$$

Démonstration Soit

$$\mathcal{D} = \left\{ \prod_{i=1}^{n}]a_i, b_i[, \text{ où } ai < b_i \text{ sont rationels} \right\}.$$

L'ensemble $\mathcal{D}$ est dénombrable. L'ensemble

$$\Omega_0 = \{\omega \in \Omega \text{ tels que } \forall B \in \mathcal{D},\ Q_n(B) \to \mathbb{Q}(B)\} = \bigcap_{B\in\mathcal{D}} \{\omega \text{ tels que } Q_n(B) \to \mathbb{Q}(B)\}\,.$$

est donc de probabilité 1 d'après l'équation (3).

Par ailleurs, pour tout ouvert $O \subset \mathbb{R}^k$, il existe une suite $(B_i)_{i\geq 1}$ d'ensembles disjoints dans $\mathcal{D}$ telle que $O = \cup_{i=1}^{+\infty} B_i$. On peut en effet obtenir cette suite par récurrence sur i, en choisissant B_{i+1} comme un élément maximal dans $\{B \in \mathcal{D} \text{ tels que } B \subset O \setminus (B_1 \cup \ldots \cup B_i)\}$. On laisse en exercice la preuve qu'une telle suite recouvre effectivement tout O (procéder par l'absurde).

En particulier, $\mathbb{Q}(O) = \sum_{i=1}^{+\infty} \mathbb{Q}(B_i)$, donc pour tout $\varepsilon > 0$, il existe $L \in \mathbb{N}^*$ tel que, pour tout $\ell \geq L$,

$$\mathbb{Q}(O) - \varepsilon \leq \sum_{i=1}^{\ell} \mathbb{Q}(B_i) = \sum_{i=1}^{\ell} \lim_{n\to+\infty} Q_n(B_i) = \lim_{n\to_\infty} \sum_{i=1}^{\ell} Q_n(B_i) \leq \liminf_{n\to+\infty} Q_n(O).$$

Puisque l'inégalité précédente est vraie pour tout $\varepsilon > 0$, $\mathbb{Q}(O) \leq \liminf Q_n(O)$. On conclut donc par le théorème de Portmanteau (théorème 2.8). □

3 Estimation paramétrique

On se place dans toute la suite sur un espace probabilisable $(\Omega, \mathcal{F})$. On considère le modèle statistique paramétrique

$$\left(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta}\right),$$

où $\mathcal{H} \subset \mathbb{R}^k$ et $\Theta \subset \mathbb{R}^d$. On supposera dans la suite que le **paramètre d'intérêt** (celui que l'on veut estimer) est $g(\theta)$ avec $g : \Theta \to \mathbb{R}^p$. Un exemple typique est le cas où g est la projection des p premières coordonnées de θ.

Par exemple, dans le modèle gaussien

$$\left(\mathbb{R}^n, \{\mathcal{N}(m, \sigma^2)^{\otimes n}\}_{m\in\mathbb{R},\sigma\in[0,+\infty)}\right),$$

le paramètre θ est le couple (m, σ), et l'ensemble Θ est $\mathbb{R}\times\mathbb{R}_+$. Le paramètre d'intérêt peut être par exemple la moyenne m de la loi inconnue, c'est-à-dire la première composante du paramètre.

La question statistique typique que se pose l'expérimentateur est la *détermination de la taille n de l'échantillon nécessaire à réaliser son objectif*, objectif qu'on supposera formulé en termes des paramètres du modèle. Dans l'exemple précédent, l'expérimentateur peut souhaiter *estimer* le paramètre m à une erreur $0,01$ près, ou bien il peut vouloir *tester* si le paramètre m appartient à un intervalle $[a, b]$ donné, qui pourrait avoir été obtenu indépendamment par une autre expérience, par exemple afin de valider le modèle gaussien.

3.1 Échantillons

La notion mathématique d'échantillon diffère légèrement de la définition statistique que nous avons donnée en section 2.2.1. Cette formalisation est nécessaire afin de pouvoir exploiter les propriétés probabilistes classiques (LGN, TCL) afin de répondre au problème statistique posé (par exemple la construction d'un intervalle de confiance).

Définition 3.1
— *Un* ***échantillon*** $(X_1, \ldots, X_n)$ *est une v.a. à valeurs dans $\mathcal{H}^n$ dont la loi est $\mathbb{P}_\theta$ pour un certain $\theta \in \Theta$ (à estimer).*
— ***L'échantillon de loi*** $\mathbb{P}_\theta$ *est une v.a.* $(X_1, \ldots, X_n)$ *sur* $(\Omega, \mathcal{F}, \mathbb{P})$ *de loi $\mathbb{P}_\theta$.*

Rappelons que le cas particulier où $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ pour tout $\theta \in \Theta$ signifie que tout échantillon est i.i.d. C'est une hypothèse faite le plus souvent lorsque le processus de sélection de l'échantillon (cf. section 2.2.1) est un tirage aléatoire avec remise dans la population, ou bien sans remise mais avec un population beaucoup plus grande que la taille n de l'échantillon.

Insistons sur la différence fondamentale entre échantillon et observations : un échantillon est une variable aléatoire $(X_1, \ldots, X_n)$ et les observations $(x_1, \ldots, x_n)$ sont une réalisation de l'échantillon :

$$\exists \omega \in \Omega, \quad (x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega)).$$

Afin de réaliser l'estimation des paramètres du modèle, on introduit la notion d'estimateur.

Définition 3.2 — *Une **statistique** S est une fonction mesurable de l'échantillon $(X_1, \ldots, X_n)$: $S = S(X_1, \ldots, X_n)$.*
— *Un **estimateur** est une statistique à valeurs dans $g(\Theta)$ (l'ensemble des valeurs possibles du paramètre d'intérêt).*

Les estimateurs sont classiquement notés avec des « chapeaux ». Par exemple, un estimateur du paramètre d'intérêt $g(\theta)$ (resp. du paramètre θ) sera noté $\hat{g}$ (resp. $\hat{\theta}$). Si on veut rendre compte de la dépendance de l'estimateur à la taille de l'échantillon, on notera $\hat{g}_n$ (resp. $\hat{\theta}_n$).

Remarquons qu'une statistique est un cas particulier de variable aléatoire. Remarquons également que la définition d'un estimateur ne comporte (pour le moment) aucune notion de qualité de l'approximation du paramètre réel du modèle. Il permet simplement de formaliser le fait qu'il faut chercher les approximations les plus performantes du paramètre d'intérêt réel dans la classe générale des estimateurs.

Pour revenir à l'exemple du jeu de pile-ou-face de la section 2.2.2, les v.a. X_1 et $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ sont tous les deux des estimateurs, mais l'un des deux semble intuitivement meilleur que l'autre pour estimer la probabilité d'obtenir pile. Comment préciser cette intuition ?

Rappelons que les données sont la réalisation $(x_1, \ldots, x_n)$ de l'échantillon $(X_1, \ldots, X_n)$ suivant une certaine loi $\mathbb{P}_{\theta_0}$ pour un paramètre $\theta_0 \in \Theta$ inconnu. Un estimateur $\hat{g}$ est une statistique, c'est-à-dire une fonction de l'échantillon : $\hat{g} = \hat{g}(X_1, \ldots, X_n)$. Si on remplace l'échantillon $(X_1, \ldots, X_n)$ par sa réalisation $(x_1, \ldots, x_n)$, on obtient la valeur $\hat{g}(x_1, \ldots, x_n)$, qui doit donner une bonne approximation de $g(\theta_0)$. Mais comme la valeur précise de θ_0 est inconnue, il faut que cette approximation soit bonne pour tout paramètre $\theta \in \Theta$. Autrement dit, on souhaite que pour tout $\theta \in \Theta$, si on considère une réalisation $(y_1, \ldots, y_n)$ de $(X_1, \ldots, X_n)$ sous $\mathbb{P}_\theta$, on ait de grandes chances que $\hat{g}(y_1, \ldots, y_n)$ soit une bonne approximation de $g(\theta)$. Finalement, la qualité d'un estimateur se mesure donc par des propriétés du type

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(|\hat{g} - g(\theta)| \leq \varepsilon) \geq 1 - \eta \tag{4}$$

pour un **seuil d'erreur** $\varepsilon > 0$ petit et un **niveau de confiance** $1-\eta$ proche de 1. Insistons sur le fait que, dans la formule précédente, $\hat{g} = \hat{g}(X_1, \ldots, X_n)$ et la probabilité $\mathbb{P}_\theta$ signifie que l'échantillon $(X_1, \ldots, X_n)$ a pour loi $\mathbb{P}_\theta$.

Puisque $(X_1, \ldots, X_n)$ a une loi différente pour chaque valeur de θ, le fait que la relation précédente soit vraie pour tout $\theta \in \Theta$ ne présente pas d'incompatibilité. Tout le problème est justement de trouver des estimateurs qui ont des valeurs typiques différentes sous différentes lois $\mathbb{P}_\theta$. Ce n'est pas possible qu'une telle propriété se produise avec probabilité 1 : par exemple, au jeu de pile-ou-face, il est toujours possible d'obtenir uniquement des piles à tous les lancers, quelle que soit la valeur de θ, mais c'est très improbable si θ est petit. C'est pourquoi on ne peut espérer que la relation (4) soit vraie presque sûrement (c'est-à-dire avec $\eta = 0$), mais seulement avec grande probabilité (avec $\eta > 0$ petit).

3.2 Estimation par insertion, méthode des moments

La première méthode d'estimation que nous allons étudier repose sur le principe suivant : considérons le modèle statistique i.i.d. $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\}_{\theta\in\Theta})$, où $\mathbb{Q}_\theta$ est une probabilité sur $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ et supposons que l'on peut écrire le paramètre d'intérêt $g(\theta)$ comme

$$g(\theta) = \varphi(\mathbb{Q}_\theta), \quad \forall \theta \in \Theta,$$

pour une certaine fonction φ définie sur les mesures de probabilités.

Dans ce cas, la méthode d'**estimation par insertion** consiste à remplacer la mesure $\mathbb{Q}_\theta$ par son approximation Q_n, la mesure empirique définie en section 2.2.4. On définit donc l'estimateur

$$\hat{g}_n = \varphi(Q_n).$$

Si la fonction φ est suffisamment régulière (par exemple si $\varphi(\mathbb{Q})$ est une fonction continue d'intégrales du type $\int_{\mathcal{H}} f \, d\mathbb{Q}$ pour des fonctions f continues bornées), on déduit du théorème de Varadarajan (théorème 2.27) que $\hat{g}_n$ converge presque sûrement quand $n \to +\infty$ vers la valeur réelle du paramètre d'intérêt $g(\theta_0)$ où $\theta_0 \in \Theta$ est telle que les données $(x_1, \ldots, x_n)$ sont une réalisation de l'échantillon de loi $\mathbb{Q}_{\theta_0}^{\otimes n}$.

La **méthode des moments** est un cas particulier de la méthode d'estimation par insertion, lorsque le paramètre d'intérêt $g(\theta)$ s'écrit comme une fonction d'un (vecteur de) moment(s) de la loi $\mathbb{Q}_\theta$.

Exemples de fonctions des moments Supposons que $\mathcal{H} \subset \mathbb{R}$, de sorte que $\mathbb{Q}_\theta$ est une mesure de probabilité sur $\mathbb{R}$.

— Si $g(\theta) = \int_{\mathbb{R}} x \, \mathbb{Q}_\theta(dx)$ (le premier moment), alors l'estimateur par la méthode des moments est la **moyenne empirique**, notée $\bar{X}_n$, de l'échantillon : en effet, dans ce cas

$$\hat{g}_n = \int_{\mathbb{R}} x \, Q_n(dx) = \frac{1}{n} \int_{\mathbb{R}} x \sum_{i=1}^n \delta_{X_i}(dx) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n.$$

— Si $g(\theta)$ est la variance de $\mathbb{Q}_\theta$, c'est-à-dire $g(\theta) = \int_{\mathbb{R}} x^2 \mathbb{Q}_\theta(dx) - \left(\int_{\mathbb{R}} x\, \mathbb{Q}_\theta(dx)\right)^2$, $g(\theta)$ est une fonction régulière des moments d'ordre 1 et 2. Dans ce cas, la méthode des moments donne pour estimateur la **variance empirique**, notée $\hat{\sigma}_n^2$, de l'échantillon :

$$\hat{g}_n = \int_{\mathbb{R}} x^2\, Q_n(dx) - \left(\int_{\mathbb{R}} x\, Q_n(dx)\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\bar{X}_n\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \hat{\sigma}_n^2.$$

— Si $g(\theta) = \int_{\mathbb{R}} x^k\, \mathbb{Q}_\theta(dx)$ (moment d'ordre k), alors $\hat{g}_n$ est le **moment empirique d'ordre** k, c'est-à-dire

$$\hat{g}_n = \int_{\mathbb{R}} x^k\, Q_n(dx) = \frac{1}{n}\sum_{i=1}^n X_i^k.$$

En particulier, l'estimateur construit dans l'exemple du jeu de pile-ou-face de la section 2.2.2 peut être obtenu par la méthode des moments. En effet, dans ce modèle, le paramètre θ est le premier moment de la loi $\mathbb{Q}_\theta = \mathrm{Bern}(\theta)$. Ci-dessous, nous donnons un autre exemple d'application de cette méthode.

Exemple Considérons le modèle statistique

$$\left(\mathbb{R}_+^n, \left\{\mathcal{U}([0,\theta])^{\otimes n}\right\}_{\theta>0}\right),$$

où $\mathcal{U}([a,b])$ désigne la loi uniforme sur l'intervalle $[a,b]$.

Ici, le paramètre à estimer est θ, la borne supérieure des valeurs possibles des observations. Le maximum de l'échantillon n'est pas un moment de la loi $\mathbb{Q}_\theta = \mathcal{U}([0,\theta])$. En revanche, son premier moment s'exprime comme une fonction simple du paramètre θ :

$$\int_{\mathbb{R}} x\, \mathcal{U}([0,\theta])(dx) = \int_0^\theta x\, \frac{dx}{\theta} = \frac{\theta}{2}.$$

On déduit donc de la méthode des moments l'estimateur suivant pour le paramètre θ :

$$\hat{\theta}_n = 2\bar{X}_n.$$

Il est également possible ici d'obtenir un estimateur par la méthode d'inférence par insertion. On a en effet la relation suivante entre θ et la loi $\mathcal{U}([0,\theta])$:

$$\theta = \inf\{x \in \mathbb{R}_+ \text{ tels que } \mathbb{Q}_\theta(]x, +\infty[) = 0\}.$$

De cette relations, on déduit par insertion l'estimateur

$$\begin{aligned}\hat{\theta}'_n &= \inf\{x \in \mathbb{R}_+ \text{ tels que } Q_n(]x, +\infty[) = 0\} \\ &= \inf\{x \in \mathbb{R}_+ \text{ tels que } x \geq X_i,\ \forall i \in \{1, \ldots, n\}\},\end{aligned}$$

c'est-à-dire

$$\hat{\theta}'_n = \max_{1 \leq i \leq n} X_i.$$

Comment savoir lequel de ces deux estimateurs est le meilleur ?

3.3 Critères de performance en moyenne

Dans la suite, pour tout $\theta \in \Theta$, on notera $\mathbb{E}_\theta$ l'espérance par rapport à la loi $\mathbb{P}_\theta$, et $\mathbb{V}_\theta(Z)$ la matrice de variance-covariance de la v.a. Z sous la loi $\mathbb{P}_\theta$, c'est-à-dire, si $Z \in \mathbb{L}^2(\mathbb{P}_\theta)$,

$$\mathbb{V}_\theta(Z) = \mathbb{E}_\theta[\big[(Z - \mathbb{E}_\theta Z)(Z - \mathbb{E}_\theta Z)^T\big].$$

Nous commençons par quelques définitions.

Définition 3.3 *Une statistique S est d'ordre $p \in [1, +\infty)$ si $S \in \mathbb{L}^p(\mathbb{P}_\theta)$ pour tout $\theta \in \Theta$.*

Nous introduisons une première façon de mesurer la qualité d'un estimateur. Un estimateur est sans biais s'il donne le bon résultat en moyenne.

Définition 3.4 *Un estimateur $\hat{g}_n$ de $g(\theta)$ est dit :*

- ***sans biais*** *lorsque $\mathbb{E}_\theta \hat{g}_n = g(\theta)$ pour tout $\theta \in \Theta$; sinon, l'estimateur est dit **biaisé** ;*
- ***asymptotiquement sans biais*** *lorsque*

$$\lim_{n \to +\infty} \mathbb{E}_\theta \hat{g}_n = g(\theta), \quad \forall \theta \in \Theta.$$

On appelle $\mathbb{E}_\theta \hat{g} - g(\theta)$ le **biais** de l'estimateur $\hat{g}$.

Exemple 1 : Si $\mathbb{P}_\eta = \mathbb{Q}_\theta^{\otimes n}$ et $g(\theta)$ est la moyenne de la loi $\mathbb{Q}_\theta$, on a vu que la méthode de moments donne pour estimateur la moyenne empirique de l'échantillon

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

Cet estimateur est *sans biais*, puisque

$$\mathbb{E}_\theta \bar{X}_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta X_i = \mathbb{E}_\theta X_1 = \int_{\mathcal{H}} x \, \mathbb{Q}_\theta(dx).$$

Exemple 2 De même, si $g(\theta)$ est la variance de la loi $\mathbb{Q}_\theta$, la méthode des moments donne pour estimateur la variance empirique

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

Cet estimateur est *biaisé* mais *asymptotiquement sans biais*, puisque, en utilisant l'indépendance des X_i,

$$\begin{aligned}
\mathbb{E}_\theta \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta(X_i^2) - \frac{1}{n^2}\sum_{1\le i,j\le n} \mathbb{E}_\theta(X_i X_j) \\
&= \mathbb{E}_\theta(X_1^2) - \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_\theta(X_i^2) - \frac{1}{n^2}\sum_{i\neq j} \mathbb{E}_\theta X_i \mathbb{E}_\theta X_j \\
&= \left(1 - \frac{1}{n}\right)\mathbb{E}_\theta(X_1^2) - \frac{n(n-1)}{n^2}(\mathbb{E}_\theta X_1)^2 \\
&= \frac{n-1}{n}\mathbb{V}_\theta(X_1).
\end{aligned}$$

On préfère souvent à la variance empirique $\hat{\sigma}_n^2$ sa version *sans biais*

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

En effet,

$$\mathbb{E}_\theta \hat{\sigma}_n'^2 = \frac{n}{n-1}\mathbb{E}_\theta \hat{\sigma}_n^2 = \mathbb{V}_\theta(X_1).$$

Afin de mesurer la qualité d'un estimateur, on définit la notion de risque.

Définition 3.5 *Soit $\hat{g}$ un estimateur de $g(\theta)$ d'ordre de 2.*

— *Pour tout $\theta \in \Theta$, on définit le* ***risque quadratique de*** *$\hat{g}$* ***sous*** *$\mathbb{P}_\theta$ comme*

$$\mathcal{R}(\theta, \hat{g}) = \mathbb{E}_\theta\left(|\hat{g} - g(\theta)|^2\right).$$

— *L'estimateur $\hat{g}$ est* ***préférable*** *à un autre estimateur $\hat{g}'$ lorsque*

$$\mathcal{R}(\theta, \hat{g}) \le \mathcal{R}(\theta, \hat{g}'), \quad \forall \theta \in \Theta.$$

Afin de calculer le risque quadratique, le résultat suivant est souvent utile. Il fait apparaître la variance de l'estimateur et le carré de son biais.

Proposition 3.6 (décomposition biais-variance) *Soit $\hat{g}$ un estimateur de $g(\theta)$ d'ordre 2. Alors*

$$\mathcal{R}(\theta, \hat{g}) = |\mathbb{E}_\theta \hat{g} - g(\theta)|^2 + \mathbb{E}_\theta\,|\hat{g} - \mathbb{E}_\theta \hat{g}|^2 = |\mathbb{E}_\theta \hat{g} - g(\theta)|^2 + \mathbb{V}_\theta(\hat{g}).$$

Démonstration Rappelons que $|\cdot|$ désigne ici la norme euclidienne. En utilisant le fait que $\hat{g} - g(\theta) = (\hat{g} - \mathbb{E}_\theta \hat{g}) + (\mathbb{E}_\theta \hat{g} - g(\theta))$ et que pour tout vecteur $u, v \in \mathbb{R}^p$, $|u+v|^2 = |u|^2 + 2u^T v + |v|^2$, on obtient

$$\begin{aligned}\mathcal{R}(\theta, \hat{g}) &= \mathbb{E}_\theta |\hat{g} - \mathbb{E}_\theta \hat{g}|^2 + 2\mathbb{E}_\theta (\hat{g} - \mathbb{E}_\theta \hat{g})^T (\mathbb{E}_\theta \hat{g} - g(\theta)) + \mathbb{E}_\theta |\mathbb{E}_\theta \hat{g} - g(\theta)|^2 \\ &= \mathbb{E}_\theta |\hat{g} - \mathbb{E}_\theta \hat{g}|^2 + |\mathbb{E}_\theta \hat{g} - g(\theta)|^2\end{aligned}$$

puisque $\mathbb{E}_\theta(\hat{g} - \mathbb{E}_\theta \hat{g}) = 0$. □

3.4 Critères de performance asymptotique

Il est souvent utile de pouvoir relier la taille de l'échantillon et la qualité d'un estimateur, par exemple pour déterminer la taille d'un échantillon nécessaire à une estimation des paramètres avec une précision donnée.

Définition 3.7 *Un estimateur $\hat{g}_n$ est dit* ***consistant*** *lorsque*

$$\hat{g}_n \xrightarrow[n\to+\infty]{\mathbb{P}_\theta} g(\theta), \quad \forall \theta \in \Theta.$$

Il est important de comprendre que, dans la définition précédente, la loi $\mathbb{P}_\theta$ dépend également de n, puisque c'est la loi de l'échantillon $(X_1, \ldots, X_n)$. Par exemple, dans le cas d'un échantillon i.i.d., on a $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ et la consistance de l'estimateur $\hat{g}_n = \hat{g}_n(X_1, \ldots, X_n)$ signifie que

$$\hat{g}_n \xrightarrow[n\to+\infty]{\mathbb{Q}_\theta^{\otimes n}} g(\theta), \quad \forall \theta \in \Theta,$$

c'est-à-dire que, pour tout $\varepsilon > 0$ et tout $\theta \in \Theta$,

$$\lim_{n\to+\infty} \mathbb{Q}_\theta^{\otimes n} \left(|\hat{g}_n(X_1, \ldots, X_n) - g(\theta)| > \varepsilon \right) = 0.$$

Remarquons que cette propriété peut également être formulée avec une loi $\mathbb{P}'_\theta$ qui ne dépend pas de n : soit $\mathbb{P}'_\theta$ la loi d'une suite i.i.d. $(X_i)_{i\geq 1}$ où chaque X_i a la loi $\mathbb{Q}_\theta$. Alors la propriété précédente s'écrit

$$\lim_{n\to+\infty} \mathbb{P}'_\theta \left(|\hat{g}_n(X_1, \ldots, X_n) - g(\theta)| > \varepsilon \right) = 0.$$

Avec cette écriture, on peut définir une notion de consistance plus forte : on dit que $\hat{g}_n = \hat{g}_n(X_1, \ldots, X_n)$ est un estimateur **fortement consistant** de $g(\theta)$ si, pour tout $\theta \in \Theta$,

$$\lim_{n\to+\infty} \hat{g}_n = g(\theta), \quad \mathbb{P}'_\theta\text{-presque sûrement.}$$

Ceci implique la consistance au sens de la définition 3.7 grâce à la proposition 2.15.

Exemple : Dans le jeu de pile-ou-face, l'estimateur du paramètre θ par la méthode des moments est la moyenne empirique

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

C'est un estimateur *consistant*, et même *fortement consistant*. En effet, sous $\mathbb{P}_\theta$ (par abus de notation, on écrit ici $\mathbb{P}_\theta$ au lieu de $\mathbb{P}'_\theta$), pour $\theta \in [0,1]$, $(X_i)_{i\geq 1}$ est une suite i.i.d. de loi Bern(θ). Ainsi, la loi des grands nombres (théorème 2.16) assure que, sous $\mathbb{P}_\theta$ (définie plus haut), $\bar{X}_n$ converge presque sûrement vers $\mathbb{E}_\theta(X_1) = \theta$.

La définition suivante vise à quantifier la vitesse de convergence.

Définition 3.8 *Soit $(\nu_n)_{n\geq 1}$ une suite de réels positifs telle que $\nu_n \to +\infty$ quand $n \to +\infty$. On dit que $\hat{g}_n$ est un estimateur de $g(\theta)$* ***de vitesse*** *ν_n si, pour tout $\theta \in \Theta$, il existe une loi $\ell(\theta)$ sur $\mathbb{R}^p$ différente de δ_0, appelée* ***loi limite de*** *$\hat{g}_n$, telle que*

$$\nu_n(\hat{g}_n - g(\theta)) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \ell(\theta), \quad \forall \theta \in \Theta.$$

Si toutes les lois $\ell(\theta)$ sont gaussiennes, on dit que $\hat{g}_n$ est un estimateur ***asymptotiquement normal*** *(ou* ***asymptotiquement gaussien****).*

De nouveau, dans la définition précédente, la loi $\mathbb{P}_\theta$ dépend en fait de n, mais dans le cas d'échantillons i.i.d., on peut réécrire cette propriété à l'aide de la loi $\mathbb{P}'_\theta$ introduite plus haut, et qui ne dépend pas de n. A partir de la section 4, on se restreindra essentiellement au cas i.i.d., et on travaillera donc toujours sous la loi $\mathbb{P}'_\theta$, qu'on notera toujours $\mathbb{P}_\theta$ par abus de notation.

Exemple (suite) : Dans le jeu de pile-ou-face, la vitesse de l'estimateur $\bar{X}_n$ peut se déduire du théorème central limite (théorème 2.18). Soit $\theta \in [0,1]$ fixé. Sous $\mathbb{P}_\theta$, les v.a. X_i sont i.i.d. de loi Bern(θ), et donc de moyenne θ et variance $\theta(1-\theta)$. On obtient donc

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}(0, \theta(1-\theta)).$$

Ainsi, l'estimateur $\bar{X}_n$ est asymptotiquement normal, de vitesse $\sqrt{n}$.

Remarque 3.9 *Un estimateur $\hat{g}$ de vitesse $\nu_n \to +\infty$ est toujours consistant. En effet,*

$$\hat{g}_n - g(\theta) = \frac{1}{\nu_n} \times \nu_n(\hat{g}_n - g(\theta)).$$

Le premier facteur est déterministe et tend vers 0, donc en particulier il converge en probabilité vers 0, et le second facteur converge en loi vers $\ell(\theta)$. On déduit donc du lemme de Slutsky (lemme 2.17) que le couple

$$(1/\nu_n\,,\,\nu_n(\hat{g}_n - g(\theta)))$$

converge en loi vers $(0, \ell(\theta))$, et donc le produit des deux éléments du couple converge en loi vers le produit des limites. On en déduit que $\hat{g}_n - g(\theta)$ converge en loi vers 0. D'après la proposition 2.15, ceci implique la convergence en probabilité, d'où la consistance de l'estimateur $\hat{g}_n$.

3.5 Asymptotique de l'erreur d'estimation et intervalles de confiance

Définition 3.10 *Une v.a. G définie sur $(\Omega, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ est dite **pivotale** si sa loi sous $\mathbb{P}_\theta$ est indépendante du paramètre θ.*

Nous supposerons dans cette sous-section que l'estimateur $\hat{g}_n$ est de vitesse $\nu_n \to +\infty$ et que sa loi limite $\ell(\theta)$ satisfait que, si $Z \sim \ell(\theta)$, alors $Z = \sigma(\theta)G$ pour une certaine constant $\sigma(\theta) > 0$ et une certain v.a. G *pivotale*, c'est-à-dire dont la loi ne dépend pas de θ.

C'est par exemple le cas si $\ell(\theta) = \mathcal{N}(0, \sigma(\theta)^2)$, et dans ce cas la v.a. G a pour loi $\mathcal{N}(0, 1)$.

Supposons que l'on dispose d'un estimateur consistant $\hat{\sigma}_n$ de $\sigma(\theta)$. Alors, d'après le lemme de Slutsky (lemme 2.17),

$$(\nu_n(\hat{g}_n - g(\theta))\,,\,\hat{\sigma}_n) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} (\sigma(\theta)G\,,\,\sigma(\theta)).$$

Puisque $\sigma(\theta) > 0$, on en déduit la convergence du quotient des deux composantes ci-dessus, et donc

$$\frac{\nu_n}{\hat{\sigma}_n}(\hat{g}_n - g(\theta)) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} G.$$

Exemple (suite) : Dans le jeu de pile-ou-face, on a vu que l'estimateur $\bar{X}_n$ est asymptotiquement normal, de variance asymptotique $\theta(1-\theta)$. On peut donc appliquer le raisonnement précédent avec $\sigma(\theta) = \sqrt{\theta(1-\theta)}$ et $G \sim \mathcal{N}(0, 1)$. Puisque $\bar{X}_n$ est un estimateur consistant de θ, on en déduit que

$$\hat{\sigma}_n = \sqrt{\bar{X}_n(1 - \bar{X}_n)}$$

est un estimateur consistant de $\sigma(\theta)$. Finalement, en appliquant le lemme de Slutsky comme ci-dessous, on obtient la convergence

$$\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}\,(\bar{X}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}(0, 1). \tag{5}$$

Le résultat précédent est très utile pour construire des intervalles de confiance. Commençons par rappeler quelques définitions.

Définition 3.11 *Soit $\alpha \in]0,1[$ un niveau de confiance fixé. Un* ***intervalle de confiance pour*** $g(\theta) \in \mathbb{R}$ ***de niveau de confiance*** $1-\alpha$ *est une statistique I à valeur dans les intervalles de $\mathbb{R}$, c'est-à-dire de la forme $I = [a(X_1,\dots,X_n)\,,\, b(X_1,\dots,X_n)]$ pour des statistiques $a < b$, telle que*

$$\mathbb{P}_\theta(g(\theta) \in I) = 1-\alpha, \quad \forall \theta \in \Theta.$$

Un intervalle de confiance constitue un compromis entre l'erreur d'estimation (la longueur de l'intervalle) et la confiance dans le résultat (le paramètre $1-\alpha$). Si α est petit, l'intervalle de confiance sera grand ; si l'intervalle de confiance est petit, la confiance $1-\alpha$ sera faible.

Exemple : On considère le modèle statistique gaussien

$$\left(\mathbb{R}^n, \left\{\mathcal{N}(\theta,1)^{\otimes n}\right\}_{\theta\in\mathbb{R}}\right).$$

Soit $\alpha \in]0,1[$ et q_α le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi gaussienne $\mathcal{N}(0,1)$, c'est-à-dire l'unique nombre q tel que

$$\mathbb{P}(G \le q) = \int_{-\infty}^{q} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 1-\frac{\alpha}{2},$$

où $G \sim \mathcal{N}(0,1)$. Remarquons que $\sqrt{n}(\bar{X}_n - \theta)$ est une v.a. gaussienne de moyenne nulle et de variance $n\frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_\theta(X_i) = 1$. Donc

$$\mathbb{P}_\theta(\sqrt{n}|\bar{X}_n-\theta| > q_\alpha) = \mathbb{P}(G > q_\alpha \text{ ou } G < -q_\alpha) = \mathbb{P}(G > q_\alpha)+\mathbb{P}(-G > q_\alpha) = \alpha.$$

D'où

$$\mathbb{P}_\theta\left(\theta \in \left[\bar{X}_n - \frac{q_\alpha}{\sqrt{n}}\,,\, \bar{X}_n + \frac{q_\alpha}{\sqrt{n}}\right]\right) = \mathbb{P}_\theta(\sqrt{n}|\bar{X}_n-\theta| \le q_\alpha) = 1-\alpha.$$

Ainsi, $\left[\bar{X}_n - \frac{q_\alpha}{\sqrt{n}}\,,\, \bar{X}_n + \frac{q_\alpha}{\sqrt{n}}\right]$ est un intervalle de confiance pour θ de niveau de confiance $1-\alpha$.

Le plus souvent, on a seulement des informations partielle sur le niveau de confiance d'un intervalle de confiance. Ceci conduit aux deux définitions suivantes.

Définition 3.12 *Soit $\alpha \in]0,1[$ un niveau de confiance fixé. Une statistique I à valeur dans les intervalles de $\mathbb{R}$ est un* ***intervalle de confiance par excès*** *pour $g(\theta) \in \mathbb{R}$ de niveau de confiance $1-\alpha$ si*

$$\mathbb{P}_\theta(g(\theta) \in I) \ge 1-\alpha, \quad \forall \theta \in \Theta.$$

Définition 3.13 *Soit $\alpha \in]0,1[$ un niveau de confiance fixé. Une statistique I_n à valeur dans les intervalles de $\mathbb{R}$ est un* ***intervalle de confiance asymptotique*** *pour $g(\theta) \in \mathbb{R}$ de niveau de confiance $1-\alpha$ si*

$$\lim_{n\to+\infty} \mathbb{P}_\theta(g(\theta) \in I_n) = 1-\alpha, \quad \forall \theta \in \Theta.$$

Exemple du jeu de pile-ou-face (suite) : Dans le jeu de pile-ou-face, la convergence dans l'équation (5) permet d'obtenir un *intervalle de confiance asymptotique* : fixons un niveau de confiance $\alpha \in]0,1[$, et rappelons la notation q_α introduite plus haut pour les quantiles gaussiens. On déduit de la convergence (5) et du théorème de Portmanteau (théorème 2.8) que

$$\begin{aligned}\lim_{n\to+\infty} \mathbb{P}_\theta &\left(\theta \in \left[\bar{X}_n - q_\alpha\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\,,\, \bar{X}_n - q_\alpha\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]\right) \\ &= \lim_{n\to+\infty} \mathbb{P}_\theta\left(\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}\,(\bar{X}_n-\theta) \in [-q_\alpha, q_\alpha]\right) \\ &= \mathbb{P}(G \in [-q_\alpha, q_\alpha]) = 1-\alpha,\end{aligned}$$

où $G \sim \mathcal{N}(0,1)$.

Nous venons de voir comment déduire du TCL, ou de calculs exacts de la loi d'un estimateur, sa loi limite et un intervalle de confiance quand le paramètre à estimer est directement θ. Si le paramètre à estimer est $g(\theta)$ et que l'on connait seulement un estimateur $\hat{\theta}_n$ de θ et sa loi limite, le résultat suivant donne les propriétés de l'estimateur $\hat{g}_n = g(\hat{\theta}_n)$ de $g(\theta)$.

Proposition 3.14 (δ-méthode) *Soit Y_n une suite de v.a. à valeurs dans $\mathbb{R}^d$ et $y \in \mathbb{R}^d$. Supposons que $\nu_n(Y_n - y)$ converge en loi vers une v.a. Z, pour une certaine suite $\nu_n \to +\infty$. Soit $g : \mathbb{R}^d \to \mathbb{R}^p$ une fonction $\mathcal{C}^1$, et soit $J_g(y)$ sa matrice jacobienne au point $y \in \mathbb{R}^d$. Alors*

$$\nu_n\left(g(Y_n) - g(y)\right) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} J_g(y)Z.$$

Ce résultat permet de déduire, à partir de la vitesse et la loi limite d'un estimateur $\hat{\theta}_n$ de θ, la vitesse et la loi limite de l'estimateur $g(\hat{\theta}_n)$ de $g(\theta)$. En l'occurrence, la vitesse est la même, et la loi limite est multipliée par la jacobienne de g au point θ.

3.6 Exemple de la régression linéaire multiple

Cet exemple est détaillé dans l'exercice 8 de la feuille d'exercice. Nous nous contentons ici de donner le résultat principal.

On cherche à expliquer des **observations** $(x_1, \ldots, x_n)$, par exemple la consommation électrique de différents foyers, à l'aide d'un certain nombre

de facteurs (ou causes) quantitatifs connus, par exemple l'âge moyen des personnes du foyer, la catégorie socio-professionnelle ou le cours du pétrole. Ces facteurs sont appelés **régresseurs**, supposés de dimension k et on note par un vecteur $R_i \in \mathbb{R}^k$ les valeurs des régresseurs pour l'observation x_i. On note également R la matrice formée des vecteurs lignes $R_1, \dots, R_n$.

On suppose que la relation entre les observations et les régresseurs est linéaire (d'où le nom de **régression linéaire**, le terme **multiple** signifiant qu'il y a plusieurs régresseurs), et qu'un aléa extérieur ε supposé gaussien, représentant l'effet de tous les autres facteurs non pris en compte influençant les observation, vient bruiter cette relation linéaire de façon i.i.d. pour chaque observation. Autrement dit, si X est l'échantillon associé à cette expérience, on a

$$X = R\theta + \varepsilon, \quad \text{où } \varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathrm{Id}),$$

où les paramètres inconnus sont $\theta \in \mathbb{R}^k$, le vecteur des poids décrivant l'influence de chaque régresseur sur les observations, et σ^2 la variance du bruit.

On considère donc le modèle statistique gaussien

$$\left(\mathbb{R}^n, \left\{\mathcal{N}_n(R\theta, \sigma^2 \mathrm{Id})\right\}_{\theta \in \mathbb{R}^k,\ \sigma > 0}\right).$$

On peut toujours supposer (quitte à réduire le nombre de régresseurs si certains sont redondants) que la matrice R est de rang k (en particulier, $k \leq n$). Dans ce cas, on a le résultat suivant, où la notion d'*estimateur du maximum de vraisemblance* est définie dans la section suivante.

Théorème 3.15 *Sous les hypothèses précédentes, l'estimateur du maximum de vraisemblance de (θ, σ) est donné par $(\hat{\theta}_n, \hat{s}_n)$ où*

$$\hat{\theta}_n = (R^T R)^{-1} R^T X \quad \textit{et} \quad \hat{s}_n^2 = \frac{1}{n} |X - R\hat{\theta}_n|^2,$$

où $|\cdot|$ désigne la norme euclidienne. De plus, sous la loi $\mathbb{P}_{(\theta,\sigma)}$, ces estimateurs ont pour loi

$$\hat{\theta}_n \sim \mathcal{N}_k(\theta, \sigma^2 (R^T R)^{-1}) \quad \textit{et} \quad \frac{n}{\sigma^2} \hat{s}_n^2 \sim \chi^2(n-k),$$

où $\chi^2(n-k)$ est la loi du khi-deux à $n-k$ degrés de liberté.

En particulier, $\hat{\theta}_n$ est un estimateur sans biais de θ, et $\frac{n}{n-k}\hat{s}_n^2$ est un estimateur sans biais de σ^2.

4 Estimation par maximum de vraisemblance

Rappelons qu'on se place dans le cadre d'un modèle statistique paramétrique

$$\left(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta}\right),$$

où $\mathcal{H} \subset \mathbb{R}^k$ et $\Theta \subset \mathbb{R}^d$. On supposera dans cette partie que le paramètre d'intérêt (celui que l'on veut estimer) est θ. On a vu dans la proposition 3.14 comment en déduire des estimateurs et leurs propriétés (consistance, vitesse, loi limite, intervalles de confiance) pour un paramètre d'intérêt de la forme $g(\theta)$.

On se restreindra à deux cas dans la suite :

— Le **cas discret**, où $\mathcal{H}$ est un ensemble fini ou dénombrable (comme pour le jeu de pile-ou-face, par exemple). Dans ce cas, les lois $\mathbb{P}_\theta$ de l'échantillon $(X_1, \ldots, X_n)$ sont des lois discrètes, caractérisées par les **probabilités des événements élémentaires**, c'est-à-dire

$$\mathbb{P}_\theta\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\}, \quad \forall (x_1, \ldots, x_n) \in \mathcal{H}^n.$$

— Le **cas continu**, où $\mathcal{H}$ est un sous-ensemble non dénombrable de $\mathbb{R}^k$ (comme pour le modèle gaussiens considéré en section 3.6). Dans ce cas, on supposera toujours dans la suite que les lois $\mathbb{P}_\theta$ de l'échantillon $(X_1, \ldots, X_n)$ sont **absolument continues par rapport à la mesure de Lebesgue** sur $(\mathbb{R}^k)^n$, et on notera

$$f_\theta(x_1, \ldots, x_n), \quad \forall (x_1, \ldots, x_n) \in \mathcal{H}^n.$$

la **densité** de $\mathbb{P}_\theta$.

4.1 Vraisemblance

Commençons par deux exemples afin d'expliquer la terminologie de « vraisemblance » (ou « likelihood » en anglais).

Exemples

— Au jeu de pile-ou-face, considérons la suite de trois lancers PPF (P pour pile, F pour face). Avec les notations introduites en section 2.2.2, ceci correspond aux observations $(x_1, x_2, x_3) = (0, 0, 1)$. Pour la valeur $\theta_1 = 1/2$ du paramètre, la probabilité de ce tirage est

$$\mathbb{P}_{\theta_1}((X_1, X_2, X_3) = (0, 0, 1)) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$

Pour la valeur $\theta_2 = 1/4$, cette probabilité est

$$\mathbb{P}_{\theta_2}((X_1, X_2, X_3) = (0, 0, 1)) = \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{64}.$$

Puisque $3/64 < 1/8$, étant données nos trois observations, la valeur θ_1 du paramètre est **plus vraisemblable** que la valeur θ_2, puisqu'elle donne une plus grande probabilité d'observation.
Observons que la démarche décrite ici revient à étudier la fonction $\theta \mapsto \mathbb{P}_\theta((X_1, X_2, X_3) = (0, 0, 1))$, et constater que sa valeur est plus grande en θ_1 qu'en θ_2.

— Dans le modèle statistique gaussien $(\mathbb{R}, \{\mathcal{N}(\theta, 1)\}_{\theta\in\mathbb{R}})$, l'échantillon est de taille 1. Supposons que l'observation x est 0. La densité f_θ de l'échantillon étant ici gaussienne, on a [2]

$$\mathbb{P}_\theta(X \in [0, dx]) = f_\theta(0)\, dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}\, dx.$$

La « quantité » dx peut ici être interprétée comme une incertitude sur la mesure de l'observation [3]. Ainsi, la valeur $\theta_1 = 0$ du paramètre donne une probabilité $1/\sqrt{2\pi}\, dx$ aux observations, alors que la valeur $\theta_2 = 1$ donne une probabilité $e^{-1/2}/\sqrt{2\pi}\, dx$. Ainsi, la valeur θ_1 du paramètre est **plus vraisemblable** que la valeur θ_2.
Il est important d'observer que la démarche suivie dans cet exemple revient à **comparer les densités** des lois du modèle statistique, évaluées sur les observations, c'est-à-dire ici au point 0. Autrement dit, comparer la vraisemblance de différentes valeurs de θ revient à étudier la fonction $\theta \mapsto f_\theta(0)$.

Ceci conduit à la définition suivante :

Définition 4.1 *On considère le modèle statistique $(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$ et une observation $(x_1, \ldots, x_n) \in \mathcal{H}^n$.*

— *Dans le **cas discret**, la **vraisemblance** de l'observation $(x_1, \ldots, x_n)$ est l'application de Θ dans $[0, 1]$ définie par*

$$\theta \mapsto \mathbb{P}_\theta((X_1, \ldots, X_n) = (x_1, \ldots, x_n)).$$

— *Dans le **cas continu**, la **vraisemblance** de l'observation $(x_1, \ldots, x_n)$ est l'application de Θ dans $[0, 1]$ définie par*

$$\theta \mapsto f_\theta(x_1, \ldots, x_n),$$

où f_θ est la densité de la loi $\mathbb{P}_\theta$.

*Dans tous les cas, on note $\theta \mapsto L_n(x_1, \ldots, x_n; \theta)$ la **fonction de vraisemblance** définie ci-dessus.*

2. On utilise ici une notation « à la physicienne ». Si on voulait écrire ceci mathématiquement, il faudrait écrire un développement limité de cette probabilité quand $dx \to 0$.

3. Autrement dit, la seule chose que l'on sait à l'issue de l'expérience est que la valeur réelle des observations est dans l'intervalle $[0, dx]$ pour un nombre $dx > 0$ petit.

On définit également $\theta \mapsto \ell_n(x_1, \dots, x_n; \theta) = \log L_n(x_1, \dots, x_n; \theta)$ *la* ***fonction de log-vraisemblance***[4]*.*

Pour alléger les écritures, on notera avec des lettres grasses les vecteurs d'échantillons ou d'obervations $\mathbf{X} = (X_1, \dots, X_n)$ et $\mathbf{x} = (x_1, \dots, x_n)$, de sorte que la fonction de vraisemblance s'écrira $L_n(\mathbf{x}; \cdot)$.

Exemples :

— Dans le jeu de pile-ou-face avec n lancers,

$$L_n(\mathbf{x}; \theta) = \theta^{\text{nombre de pile}}(1-\theta)^{\text{nombre de face}} = \theta^{n\bar{x}_n}(1-\theta)^{n(1-\bar{x}_n)}, \quad (6)$$

où on rappelle que $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ est la moyenne empirique des observations.

— Dans le modèle statistique

$$\left(\mathbb{R}^n, \left\{\mathcal{N}(m, \sigma^2)^{\otimes n}\right\}_{m\in\mathbb{R},\ \sigma>0}\right),$$

la vraisemblance de $x_1, \dots, x_n$ s'écrit

$$L_n(\mathbf{x}; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-m)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i-m)^2}{2\sigma^2}\right).$$

Le résultat suivant indique que, presque sûrement, les observations ont une vraisemblance strictement positive.

Proposition 4.2 *Pour tout* $\theta \in \Theta$*,*

$$L_n(\mathbf{X}; \theta) = L_n(X_1, \dots, X_n; \theta) > 0, \quad \mathbb{P}_\theta\text{-}p.s.$$

Démonstration Dans le cas discret, le résultat est évident. En effet, pour tout $(x_1, \dots, x_n) \in \mathcal{H}^n$, $L_n(\mathbf{x}; \theta) > 0$ ssi $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) > 0$, puisque les deux quantités sont égales.

Dans le cas continu,

$$\mathbb{P}_\theta(L_n(\mathbf{X}; \theta) = 0) = \int_{\{L_n(\cdot;\theta)=0\}} f_\theta(\mathbf{x})\, d\mathbf{x} = \int_{\{L_n(\cdot;\theta)=0\}} L_n(\mathbf{x}; \theta)\, d\mathbf{x} = 0. \quad \square$$

Terminons par une propriété évidente mais fondamentale dans le cas des échantillons i.i.d.

Proposition 4.3 *Si pour tout* $\theta \in \Theta$*,* $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$*, alors*

$$L_n(x_1, \dots, x_n; \theta) = \prod_{i=1}^n L(x_i; \theta) \quad \textit{et} \quad \ell_n(x_1, \dots, x_n; \theta) = \sum_{i=1}^n \ell(x_i; \theta),$$

où $L(x; \theta) = L_1(x; \theta)$ *et* $\ell(x; \theta) = \ell_1(x; \theta)$*.*

4. Ici et dans toute la suite du cours, log désigne le logarithme en base naturelle, ou logarithme népérien, fonction réciproque de la fonction $x \mapsto e^x$.

4.2 Estimation par maximum de vraisemblance : définition

Dans la suite, l'abréviations **EMV** signifie « estimation par maximum de vraisemblance » ou « estimateur du maximum de vraisemblance ». En anglais, on utilise l'abréviation **MLE** pour « maximum likelihood estimation » ou « maximum likelihood estimator ».

Le principe de cette méthode repose sur l'intuition que la probabilité sous $\mathbb{P}_\theta$ des observations $(x_1, \ldots, x_n)$ doit être élevée pour θ proche de θ_0, la valeur réelle du paramètre.

Définition 4.4 *Un* ***estimateur du maximum de vraisemblance*** *(EMV) de θ est un estimateur $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ tel que*

$$L_n(X_1, \ldots, X_n; \hat{\theta}) = \sup_{\theta \in \Theta} L_n(X_1, \ldots, X_n; \theta).$$

De plus, $\hat{\theta}$ est un EMV, alors $\hat{\theta}(x_1, \ldots, x_n)$ est appellé **estimateur ponctuel du maximum de vraisemblance**.

Remarquons qu'il n'y a pas nécessairement unicité d'un EMV, car plusieurs valeurs de θ pourraient maximiser la vraisemblance.

Remarquons que la construction d'un EMV nécessite de trouver un argmax de la vraisemblance, c'est-à-dire la solution d'un problème d'optimisation. Cette solution n'est généralement pas explicite, et nécessite dans les cas pratiques de réaliser une optimisation numérique. On renvoie aux algorithmes bien connus de Newton-Raphson et de Gauss-Newton pour des exemples simples de méthodes d'optimisation. Notons également que la vraisemblance n'est pas toujours explicite dans les modèles compliqués, où la denstié des observations est difficile à évaluer. D'un point de vue pratique, il est alors nécessaire d'utiliser des méthodes d'approximation de densité, comme par exemple l'algorithme de Metropolis-Hastings ou les méthodes MCMC (Markov Chain Monte Carlo). Nous n'aborderons pas dans ce cours ces difficultés, et les exercices ne porteront que sur des cas explicites.

Remarque 4.5 *Dans le cas d'échantillons indépendants et identiquement distribués ($\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ pour tout $\theta \in \Theta$), on préfère généralement chercher un EMV $\hat{\theta}$ en maximisant la log-vraisemblance :*

$$\ell_n(X_1, \ldots, X_n; \hat{\theta}) = \sup_{\theta \in \Theta} \ell_n(X_1, \ldots, X_n; \theta).$$

En effet, les techniques de calcul différentiel sont souvent plus simples à mettre en œuvre pour des sommes de fonctions que pour des produits (cf. proposition 4.3).

Exemple : Dans le jeu de pile-ou-face, d'après l'équation (6), la log-vraisemblance est donnée par

$$\ell_n(x_1, \ldots, x_n; \theta) = n\bar{x}_n \log \theta + n(1 - \bar{x}_n) \log(1 - \theta).$$

Ainsi,

$$\frac{\partial}{\partial \theta} \ell_n(x_1, \ldots, x_n; \theta) = \frac{n\bar{x_n}}{\theta} - \frac{n(1 - \bar{x}_n)}{1 - \theta}$$

s'annule ssi $\theta = \bar{x}_n$, est positive avant et négative après. La log-vraisemblance est donc maximale en ce point uniquement. Il existe donc un unique EMV, donné par $\hat{\theta}_n = \bar{X}_n$. On retrouve dans cet exemple le même estimateur que par la méthode des moments (cf. section 3.2).

La suite de ce chapitre est consacrée à l'étude des propriétés théoriques de l'EMV (consistance, normalité asymptotique, optimalité). Nous allons donc nous placer sous des hypothèses générales afin de faciliter notre étude. Ces hypothèses ne sont pas toujours vérifiées dans les cas pratiques, mais les arguments présentés ici peuvent souvent être adaptés aux cas particuliers, comme nous le verrons dans la feuille d'exercice.

4.3 Information de Kullback-Leibler

Notre premier objectif est d'étudier la consistance de l'EMV. Pour cela, nous introduisons la définition suivante.

Définition 4.6 *Pour tout $\alpha, \theta \in \Theta$, l'**information de Kullback-Leibler** (ou* divergence de Kullback-Leibler*, ou* entropie relative*) entre $\mathbb{P}_\alpha$ et $\mathbb{P}_\theta$ est*

$$K_n(\alpha, \theta) = -\mathbb{E}_\theta \log \frac{L_n(\mathbf{X}; \alpha)}{L_n(\mathbf{X}; \theta)} = \mathbb{E}_\theta \left[\ell_n(\mathbf{X}; \theta) - \ell_n(\mathbf{X}; \alpha)\right]$$

si $\ell_n(\mathbf{X}; \alpha)$ et $\ell_n(\mathbf{X}; \alpha)$ appartiennent à $\mathbb{L}^1(\mathbb{P}_\theta)$ et $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$, et $K_n(\alpha, \theta) = +\infty$ sinon.

Remarquons que, lorsque $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$, alors la densité de $\mathbb{P}_\alpha$ par rapport à $\mathbb{P}_\theta$ est donnée par (cf. théorème 2.6)

$$\frac{L_n(\mathbf{x}; \alpha)}{L_n(\mathbf{x}; \theta)}, \quad \forall \mathbf{x} \in \mathcal{H}^n.$$

Exemple : Dans le jeu de pile-ou-face, d'après la formule (6),

$$\begin{aligned} K_n(\alpha, \theta) &= \mathbb{E}_\theta \left[n\bar{X}_n \log \frac{\theta}{\alpha} + n(1 - \bar{X}_n) \log \frac{1 - \theta}{1 - \alpha} \right] \\ &= n\theta \log \frac{\theta}{\alpha} + n(1 - \theta) \log \frac{1 - \theta}{1 - \alpha}. \end{aligned}$$

L'information de Kullback-Leibler est une mesure de dissimilarité entre les lois $\mathbb{P}_\alpha$ et $\mathbb{P}_\theta$:

Proposition 4.7 *Pour tout $\alpha, \theta \in \Theta$, $K_n(\alpha, \theta) \geq 0$.*
Si de plus le modèle statistique est identifiable, alors $K_n(\alpha, \theta) = 0$ si et seulement si $\alpha = \theta$.

Démonstration Supposons que $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$ (sinon $K_n(\alpha, \theta) = +\infty$ et il n'y a rien à démontrer). Puisque la fonction $-\log$ est convexe, l'inégalité de Jensen (cf section 2.1.3) implique que

$$K_n(\alpha, \theta) \geq -\log \mathbb{E}_\theta \frac{L_n(\mathbf{X}; \alpha)}{L_n(\mathbf{X}; \theta)}. \tag{7}$$

Dans le cas discret,

$$\mathbb{E}_\theta \frac{L_n(\mathbf{X}; \alpha)}{L_n(\mathbf{X}; \theta)} = \sum_{\mathbf{x} \in \mathcal{H}^n} \frac{L_n(\mathbf{x}; \alpha)}{L_n(\mathbf{x}; \theta)} \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x} \in \mathcal{H}^n} L_n(\mathbf{x}; \alpha) = \sum_{\mathbf{x} \in \mathcal{H}^n} \mathbb{P}_\alpha(\mathbf{X} = \mathbf{x}) = 1.$$

De même, dans le cas continu,

$$\mathbb{E}_\theta \frac{L_n(\mathbf{X}; \alpha)}{L_n(\mathbf{X}; \theta)} = \int_{\mathcal{H}^n} \frac{L_n(\mathbf{x}; \alpha)}{L_n(\mathbf{x}; \theta)} f_\theta(d\mathbf{x}) = \int_{\mathcal{H}^n} L_n(\mathbf{x}; \alpha)\, d\mathbf{x} = \int_{\mathcal{H}^n} f_\alpha(\mathbf{x})\, d\mathbf{x} = 1,$$

où f_θ est la densité de $\mathbb{P}_\theta$. Dans les deux cas, on déduit de (7) que $K_n(\alpha, \theta) \geq 0$. De plus, le cas d'égalité de l'inégalité de Jensen nous assure que $K_n(\alpha, \theta) = 0$ si et seulement si $\frac{L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)}$ est constant $\mathbb{P}_\theta$-presque sûrement, c'est-à-dire si $L_n(\mathbf{x}; \alpha) = C L_n(\mathbf{x}; \theta)$ pour une certaine constante C pour $\mathbb{P}_\theta$-presque tout $\mathbf{x} \in \mathcal{H}^n$. Puisque $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$, on en déduit que la densité de $\mathbb{P}_\alpha$ par rapport à $\mathbb{P}_\theta$ est la constante C. Puisque $\mathbb{P}_\alpha$ et $\mathbb{P}_\theta$ sont tous deux de masse 1, ceci implique que $C = 1$ et donc que $\mathbb{P}_\alpha = \mathbb{P}_\theta$. Par identifiabilité, on en déduit que $\alpha = \theta$. □

4.4 EMV : consistance

Le résultat suivant montre que l'EMV est consistant dans le cas d'échantillons i.i.d.

Théorème 4.8 (consistance de l'EMV) *Supposons que le modèle est identifiable, que $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ pour tout $\theta \in \Theta$ et que $\Theta \subset \mathbb{R}^d$ est compact.*

- *Sous l'hypothèse que $\log L(x; \cdot)$ est continue sur Θ pour tout $x \in \mathcal{H}$, il existe un EMV.*
- *Sous l'hypothèse supplémentaire que, pour tout $\theta \in \Theta$, il existe un voisinage V de θ dans Θ et il existe une v.a. $H \in \mathbb{L}^1(\mathbb{Q}_\theta)$ tel que $\sup_{\alpha \in V} |\log L(X_1; \alpha)| \leq H$, alors l'EMV est* consistant.

Remarquons que les hypothèses de ce théorème sont très restrictives, et que ses conclusions sont valides dans des cas beaucoup plus généraux. En effet, la condition de continuité de $\log L(x; \cdot)$ impose que $L_n(x; \theta) > 0$ pour tout $x \in \mathcal{H}$ et $\theta \in \Theta$. On verra dans la feuille d'exercice plusieurs exemples où cette condition n'est pas satisfaite.

Exemple 1 : On considère le modèle statistique exponentiel

$$\left(\mathbb{R}_+^n, \left\{\mathrm{Exp}(\theta)^{\otimes n}\right\}_{\theta>0}\right).$$

Dans ce cas, la vraisemblance est donnée par

$$L_n(x_1,\dots,x_n;\theta) = \prod_{i=1}^n L(x_i;\theta) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-n\theta \bar{x}_n},$$

où on rappelle que $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. La log-vraisemblance est donc donnée par

$$\ell_n(x_1,\dots,x_n;\theta) = n\log\theta - n\theta\bar{x}_n.$$

La dérivée de cette quantité par rapport à θ est $n/\theta - n\bar{x}_n$. En tant que fonction de $\theta > 0$, elle est donc maximale pour $1/\bar{x}_n$. Ainsi, $\hat{\theta} = 1/\bar{X}_n$ est l'unique EMV de ce modèle. La consistance de cet estimateur découle de la loi forte des grands nombres (théorème 2.16), puisque $\bar{X}_n \to \mathbb{E}_\theta X_1 = 1/\theta$ en probabilité quand $n \to +\infty$. Si on voulait obtenir ce résultat à l'aide du théorème 4.8, la plupart des hypothèses seraient vérifiées, puisque $\ell_n(\mathbf{x};\theta)$ est une fonction continue de $\theta > 0$, et $|\ell(X_1;\theta)| \le |\log\theta| + \theta X_1$, qui est localement $\mathbb{L}^1$. Cependant, l'espace des paramètres $\Theta = \mathbb{R}_+^*$ n'est pas compact. Ainsi, le théorème 4.8 ne s'applique pas, même si sa conclusion reste valide.

Exemple 2 : Dans le jeu de pile-ou-face, on a

$$\log L(X_1;\theta) = \ell(X_1;\theta) = X_1\log\theta + (1-X_1)\log(1-\theta),$$

où $X_1 \in \{0,1\}$. Donc $\theta \mapsto \ell(0;\theta)$ et $\theta \mapsto \ell(1;\theta)$ sont continus sur $]0,1[$, et

$$|\ell(X_1;\theta)| \le \log\theta + \log(1-\theta).$$

On peut donc appliquer le théorème 4.8, mais seulement si on réduit l'ensemble des paramètres à un intervalle du type $[\varepsilon, 1-\varepsilon]$ pour $\varepsilon > 0$. Pourtant, nous avons vu en section 2.2.2 que l'EMV $\bar{X}_n$ est consistant sur l'ensemble des paramètres complet $\Theta = [0,1]$.

Démonstration L'existence d'un EMV $\hat{\theta}_n$ est une conséquence de la continuité de la fonction de vraisemblance et de la compacité de Θ.

Démontrons maintenant le second point du théorème 4.8. Soit $\theta \in \Theta$ fixé. Pour tout $\alpha \in \Theta$, on pose

$$U_n(\alpha) = \frac{1}{n}\log L_n(X_1,\dots,X_n;\alpha) = \frac{1}{n}\sum_{i=1}^n \ell(X_i;\alpha)$$

et

$$U(\alpha) = \mathbb{E}_\theta \log L_n(X_1,\dots,X_n;\alpha).$$

D'après la loi forte des grands nombres (théorème 2.16), $U_n \to U$ en $\mathbb{P}_\theta$-probabilité. De plus, par définition d'un EMV, $U_n(\hat{\theta}_n) = \sup_{\alpha\in\Theta} U_n(\alpha)$. Le résultat repose sur le lemme suivant, prouvé à la fin de cette preuve.

Lemme 4.9 *La suite U_n converge uniformément en $\mathbb{P}_\theta$-probabilité vers U, c'est-à-dire que pour tout $\varepsilon > 0$,*

$$\lim_{n\to+\infty} \mathbb{P}_\theta \left(\sup_{\alpha\in\Theta} |U_n(\alpha) - U(\alpha)| > \varepsilon \right) = 0.$$

Puisque

$$\left| \sup_{\alpha\in\Theta} U_n(\alpha) - \sup_{\alpha\in\Theta} U(\alpha) \right| \leq \sup_{\alpha\in\Theta} |U_n(\alpha) - U(\alpha)|,$$

on déduit du lemme que $U_n(\hat{\theta}_n)$ converge vers $\sup_{\alpha\in\Theta} U(\alpha)$ en $\mathbb{P}_\theta$-probabilité. De plus, les hypothèses du théorème impliquent que

$$\sup_{\alpha\in V} |\log L_n(X_1, \ldots, X_n; \alpha)| = \sup_{\alpha\in V} \left| \sum_{i=1}^{n} \log L(X_i; \alpha) \right| \leq n\, H,$$

et on peut donc appliquer le théorème de continuité sous le signe somme (théorème 2.3) pour en déduire que $\alpha \mapsto U(\alpha)$ est continue. Par compacité de Θ, il existe $\tau \in \Theta$ tel que $U(\tau) = \sup_{\alpha\in\Theta} U(\alpha)$. Or l'information de Kullback-Leibler vérifie

$$K_n(\tau, \theta) = U(\theta) - U(\tau) = U(\theta) - \sup_{\alpha\in\Theta} U(\alpha) \leq 0.$$

On déduit donc de la proposition 4.7 et du fait que le modèle est supposé identifiable que $K_n(\tau, \theta) = 0$ et donc que $\tau = \theta$.

Ainsi, $U_n(\hat{\theta}_n) \to U(\theta)$ et $K_n(\hat{\theta}_n, \theta) \to 0$ en $\mathbb{P}_\theta$-probabilité, puisque

$$0 \leq K_n(\hat{\theta}_n, \theta) = U(\theta) - U(\hat{\theta}_n) \leq |U(\theta) - U_n(\hat{\theta}_n)| + \sup_{\alpha\in\Theta} |U_n(\alpha) - U(\alpha)| \xrightarrow[n\to+\infty]{\mathbb{P}_\theta} 0$$

d'après le lemme 4.9.

Puisque $K_n(\alpha, \theta) = 0$ ssi $\alpha = \theta$ et $\alpha \mapsto K_n(\alpha, \theta)$ est continue, on a la propriété suivante : pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que, pour tout $\alpha \in \Theta$ tel que $|\alpha - \theta| > \eta$, on a $K_n(\alpha, \theta) > \varepsilon$ (pour le démontrer, procéder par l'absurde afin de montrer que sinon, il existerait une suite $(\alpha_n)_{n\in\mathbb{N}}$ qui convergerait vers θ telle que $K_n(\alpha_n, \theta) > \varepsilon$ pour tout $n \in \mathbb{N}$, ce qui contredirait la continuité de K_n). Ceci implique que

$$\limsup_{n\to+\infty} \mathbb{P}_\theta \left(|\hat{\theta}_n - \theta| > \eta \right) \leq \limsup_{n\to+\infty} \mathbb{P}_\theta \left(K_n(\hat{\theta}_n, \theta) > \varepsilon \right) = 0.$$

Puisque cette propriété est vraie pour tout $\varepsilon > 0$, on a démontré la consistance de l'estimateur $\hat{\theta}_n$. □

Démonstration du lemme 4.9 Rappelons que $\theta \in \Theta$ est fixé dans toute la preuve. Pour tout $x \in \mathcal{H}$ et $\eta > 0$, posons

$$h(x, \eta) = \sup_{\alpha,\beta\in\Theta,\ \text{t.q. } |\alpha-\beta|\leq\eta} |\log L(x; \alpha) - \log L(x; \beta)|\,.$$

La compactité de Θ et l'hypothèse de domination du théorème 4.8 permettent d'appliquer le théorème de convergence dominée (théorème 2.2) pour en déduire que, pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que $\mathbb{E}_\theta h(X_1; \eta) < \varepsilon/3$.

Par compacité de Θ, il existe $N \in \mathbb{N}$ et $\theta_1, \ldots, \theta_N > 0$ tels que

$$\Theta = \bigcup_{j=1}^{N} B(\theta_j, \eta),$$

où $B(\theta, r)$ est la boule ouverte de Θ centrée en $\theta \in \Theta$ et de rayon $r > 0$. On a alors

$$\begin{aligned}
\sup_{\alpha\in\Theta} |U_n(\alpha) - U(\alpha)| &= \max_{1\leq j\leq N} \sup_{\alpha\in B(\theta_j,\eta)} |U_n(\alpha) - U(\alpha)| \\
&\leq \max_{1\leq j\leq N} \sup_{\alpha\in B(\theta_j,\eta)} |U_n(\alpha) - U_n(\theta_j)| + \max_{1\leq j\leq N} |U_n(\theta_j) - U(\theta_j)| \\
&\quad + \max_{1\leq j\leq N} \sup_{\alpha\in B(\theta_j,\eta)} |U(\theta_j) - U(\alpha)| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} h(X_i;\eta) + \max_{1\leq j\leq N} |U_n(\theta_j) - U(\theta_j)| + \mathbb{E}_\theta h(X_1,\eta).
\end{aligned}$$

D'après la loi des grands nombres (théorème 2.16), le premier terme du membre de droite converge en $\mathbb{P}_\theta$-probabilité vers $\mathbb{E}_\theta h(X_1, \eta)$, qui est inférieur à $\varepsilon/3$, et le second terme du membre de droite converge en $\mathbb{P}_\theta$-probabilité vers 0. On en déduit que

$$\begin{aligned}
\mathbb{P}_\theta\left(\sup_{\alpha\in\Theta} |U_n(\alpha) - U(\alpha)| > \varepsilon\right) &\leq \mathbb{P}_\theta\left(\left|\frac{1}{n}\sum_{i=1}^{n} h(X_i;\bar\eta) - \mathbb{E}_\theta h(X_1,\bar\eta)\right| > \frac{\varepsilon}{6}\right) \\
&\quad + \mathbb{P}_\theta\left(\max_{1\leq j\leq N} |U_n(\theta_j) - U(\theta_j)| > \frac{\varepsilon}{6}\right),
\end{aligned}$$

qui converge vers 0 quand $n \to +\infty$. □

4.5 Information de Fisher

On s'intéresse maintenant à la normalité asymptotique de l'EMV. Pour cela, nous avons besoin d'introduire la notion d'information de Fisher.

Dans la suite, on note ∇ le gradient par rapport à la variable $\theta \in \Theta$ et ∇^2 la matrice hessienne, $\mathbb{V}_\theta(Z)$ la matrice de variance-covariance du vecteur aléatoire Z par rapport à la loi $\mathbb{P}_\theta$, et $\mathrm{Cov}_\theta(Z_1, Z_2)$ la covariance entre les v.a. réelles Z_1 et Z_2 par rapport à $\mathbb{P}_\theta$.

On suppose ici que Θ est un ouvert de $\mathbb{R}^d$ et que $\nabla \log L_n(\mathbf{X};\theta) \in \mathbb{L}^2(\mathbb{P}_\theta)$ pour tout $\theta \in \Theta$, et on note K l'application de Θ dans $\mathbb{R}_+$ définie par

$$K : \alpha \mapsto K_n(\alpha, \theta),$$

pour $\theta \in \Theta$ fixé.

Le calcul qui suit est formel, au sens où on s'autorise à appliquer le théorème de dérivation sous le signe somme (théorème 2.4) sans en vérifier les hypothèses. On verra dans la proposition 4.12 ci-dessous comment justifier ce calcul. On a

$$\mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta)) = \mathbb{E}_\theta\left(\frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)}\right)^2 - \left(\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)}\right)^2.$$

Dans le cas discret,

$$\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} = \sum_{\mathbf{x}\in\mathcal{H}^n} \nabla L_n(\mathbf{x};\theta) = \nabla \sum_{\mathbf{x}\in\mathcal{H}^n} L_n(\mathbf{x};\theta) = \nabla 1 = 0. \tag{8}$$

De même, dans le cas continu,

$$\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} = \mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\theta)}{f_\theta(\mathbf{X})} = \int_{\mathcal{H}^n} \nabla L_n(\mathbf{x};\theta)\, d\mathbf{x} = \nabla \int_{\mathcal{H}^n} f_\theta(\mathbf{x})\, d\mathbf{x} = 0. \tag{9}$$

Ainsi,

$$\mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta)) = \mathbb{E}_\theta \left(\frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} \right)^2.$$

Par ailleurs,

$$\nabla K(\alpha) = -\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\alpha)}$$

et donc

$$\nabla^2 K(\theta) = \mathbb{E}_\theta \left(\frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} \right)^2 - \mathbb{E}_\theta \frac{\nabla^2 L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)}.$$

On démontre de la même façon que dans (8) et (9) que le dernier terme du membre de droite est nul, de sorte que

$$\nabla^2 K(\theta) = \mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta)).$$

Ceci nous conduit à la définition suivante.

Définition 4.10 *On suppose que Θ est un ouvert $\mathbb{R}^d$ et que $\nabla \log L_n(\mathbf{X};\theta) \in \mathbb{L}^2(\mathbb{P}_\theta)$ pour tout $\theta \in \Theta$. L'**information de Fisher** est définie pour tout $\theta \in \Theta$ par $I_n(\theta) = \mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta))$, c'est-à-dire*

$$I_n(\theta) = \left(Cov_\theta \left(\frac{\partial}{\partial \theta_i} \log L_n(\mathbf{X};\theta)\, ;\, \frac{\partial}{\partial \theta_j} \log L_n(\mathbf{X};\theta) \right) \right)_{1\le i,j\le d}.$$

Le calcul précédent montre que $I_n(\theta)$ décrit la courbure de $\alpha \mapsto K_n(\alpha,\theta)$ en son minimum $\alpha = \theta$: au voisinage de θ,

$$K_n(\alpha,\theta) = \frac{1}{2}(\alpha-\theta)^T\, \mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta))\, (\alpha-\theta) + o(|\alpha-\theta|^2).$$

L'information de Fisher permet donc de quantifier, dans un modèle statistique donné, le pouvoir de discrimination de l'information de Kullback-Leibler entre deux valeurs proches du paramètre.

Exemple : Dans le jeu de pile-ou-face, d'après (6),

$$\begin{aligned} I_n(\theta) &= \mathbb{V}_\theta \left(\nabla \left(n\bar{X}_n \log\theta + n(1-\bar{X}_n)\log(1-\theta) \right) \right) \\ &= \left(\frac{1}{\theta} + \frac{1}{1-\theta} \right)^2 \mathbb{V}_\theta(n\bar{X}_n) \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

Ainsi, l'EMV a une faible incertitude pour θ proche de 0 ou 1 (voir le théorème 4.14 ci-dessous).

Proposition 4.11 *Dans le cas d'échantillons i.i.d., c'est-à-dire si* $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ *pour tout* $\theta \in \Theta$*, on a* $I_n(\theta) = nI(\theta)$*, où*

$$I(\theta) = \mathbb{V}_\theta(\nabla \log L(X_1;\theta)) = \mathbb{E}_{\mathbb{Q}_\theta} \left[\nabla \log L(X_1;\theta) \left(\nabla \log L(X_1;\theta) \right)^T \right]$$

*est l'**information de Fisher** du modèle* $(\mathcal{H}, \{\mathbb{Q}_\theta\}_{\theta\in\Theta})$*.*

Démonstration Puisque

$$\nabla \log L_n(X_1,\ldots,X_n;\theta) = \sum_{i=1}^n \nabla \log L(X_i;\theta)$$

et que les X_i sont i.i.d., on a

$$I_n(\theta) = \sum_{i=1}^n \mathbb{V}_\theta(\nabla \log L(X_i;\theta)) = n\mathbb{V}_\theta(\nabla \log L(X_1;\theta)). \qquad \square$$

Récapitulons pour un usage futur les résultats obtenus dans (8) et (9).

Proposition 4.12 *Supposons que, pour tout* $\theta \in \Theta$*, il existe un voisinage* V *de* θ *dans* Θ *tel que* $\sup_{\alpha\in V} \left| \frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} \right| \in \mathbb{L}^1(\mathbb{P}_\theta)$*. Alors*

$$\mathbb{E}_\theta \nabla \log L_n(\mathbf{X};\theta) = 0. \tag{10}$$

Si de plus $\sup_{\alpha\in V} \left| \frac{\nabla^2 L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} \right| \in \mathbb{L}^1(\mathbb{P}_\theta)$*, alors* $I_n(\theta)$ *existe et*

$$I_n(\theta) = -\mathbb{E}_\theta \nabla^2 \log L_n(\mathbf{X};\theta).$$

Démonstration Nous allons seulement prouver en détail la propriété (10). Le reste se démontre de la même façon. Il s'agit de justifier l'interversion entre le gradient et la somme ou l'intégrale dans les calculs (8) et (9) à l'aide du théorème de dérivation sous le signe somme (théorème 2.4). Pour $\theta \in \Theta$ fixé, d'après nos hypothèses, il existe un voisinage V de θ tel que

$$\sup_{\alpha\in V} \left| \frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} \right| \in \mathbb{L}^1(\mathbb{P}_\theta).$$

Dans le cas discret, cela signifie que l'application qui à $\mathbf{x} \in \mathcal{H}^n$ associe

$$\sup_{\alpha \in V} |\nabla L_n(\mathbf{x}; \alpha)|$$

est dans $\mathbb{L}^1(\mu)$, où $\mu = \sum_{\mathbf{x} \in \mathcal{H}^n} \delta_{\mathbf{x}}$ est la mesure de comptage sur $\mathcal{H}^n$, et dans le cas continue, que l'application qui à $\mathbf{x} \in \mathcal{H}^n$ associe

$$\sup_{\alpha \in V} |\nabla L_n(\mathbf{x}; \alpha)|$$

est dans $\mathbb{L}^1(\text{Leb})$, où Leb désigne la mesure de Lebesgue sur $\mathcal{H}^n$. On peut donc appliquer le théorème 2.4 pour en déduire, dans le cas discret, que

$$\sum_{\mathbf{x} \in \mathcal{H}^n} \nabla L_n(\mathbf{x}; \theta) = \int_{\mathcal{H}^n} \nabla L_n(\mathbf{x}; \theta)\, \mu(d\mathbf{x}) = \nabla \int_{\mathcal{H}^n} L_n(\mathbf{x}; \theta)\, \mu(d\mathbf{x}) = \nabla \sum_{\mathbf{x} \in \mathcal{H}^n} L_n(\mathbf{x}; \theta),$$

et dans le cas continu que

$$\int_{\mathcal{H}^n} \nabla L_n(\mathbf{x}; \theta)\, d\mathbf{x} = \nabla \int_{\mathcal{H}^n} L_n(\mathbf{x}; \theta)\, d\mathbf{x}.$$

Ainsi, les calculs (8) et (9) sont justifiés. □

4.6 EMV : normalité asymptotique

Commençons par donner l'hypothèse principale du résultat de normalité asymptotique de l'EMV.

Définition 4.13 *Le modèle statistique $(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta \in \Theta}$ est dit* ***régulier*** *si*

- *pour tout $\theta \in \Theta$, $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$;*
- *pour tout $x \in \mathcal{H}$, l'application $\theta \mapsto \log L(x; \theta)$ est continue sur Θ ;*
- *pout tout $\theta \in \theta$, il existe un voisinage V de θ dans Θ tel que*

$$\sup_{\alpha \in V} \left| \frac{\nabla L_n(\mathbf{X}; \alpha)}{L_n(\mathbf{X}; \theta)} \right| + \left| \frac{\nabla^2 L_n(\mathbf{X}; \alpha)}{L_n(\mathbf{X}; \theta)} \right| \in \mathbb{L}^1(\mathbb{P}_\theta),$$

  de sorte que $\mathbb{E}_\theta \nabla \log L_n(\mathbf{X}; \theta) = 0$ et $I_n(\theta) = -\mathbb{E}_\theta \nabla^2 \log L_n(\mathbf{X}; \theta)$, d'après la proposition 4.12 ;
- *pour tout $\theta \in \Theta$, $I_n(\theta)$ est une matrice inversible ;*

Exemple : Il est facile de vérifier que le modèle statistique du jeu de pile-ou-face avec $\Theta =]0, 1[$ est régulier.

Théorème 4.14 (nomalité asymptotique de l'EMV) *On considère $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\})_{\theta \in \Theta})$ un modèle statistique régulier. Si un EMV $\hat{\theta}_n$ existe pour tout n suffisamment grand et est consistant, alors*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \to +\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}_d(0, I(\theta)^{-1}), \quad \forall \theta \in \Theta.$$

Avec le vocabulaire de la section 3.4, ce résultat signifie que l'EMV a pour vitesse $\sqrt{n}$ et pour loi limite $\mathcal{N}_d(0, I(\theta)^{-1})$.

Les hypothèses de ce résultat sont nécessaires, car il existe des exemples de modèles statistiques i.i.d. pour lesquels l'EMV est asymptotiquement non-normal.

Exemple : Le modèle statistique

$$\left(\mathbb{R}_n^n, \{\mathcal{U}([0,\theta])^{\otimes n}\}_{\theta>0}\right),$$

où $\mathcal{U}([a,b])$ désigne la loi uniforme sur l'intervalle $[a,b]$, a pour vraisemblance pour tout $(x_1,\dots,x_n) \in \mathbb{R}_+^n$

$$L_n(x_1,\dots,x_n;\theta) = \theta^{-n}\mathbb{1}_{0\le x_1,\dots,x_n\le\theta}.$$

Il en résulte que l'EMV est la plus petite valeur de θ (afin de maximiser θ^{-n}) telle que l'indicatrice est non nulle dans l'expression précédente. Donc l'EMV est donné ici par

$$\hat{\theta}_n = \max_{1\le i\le n} X_i.$$

Remarquons que $0 \le \hat{\theta}_n \le \theta$ $\mathbb{P}_\theta$-presque sûrement et que, pour tout $t \in [-n\theta, 0]$,

$$\mathbb{P}_\theta\left(n(\hat{\theta}_n - \theta) \le t\right) = \mathbb{P}_\theta\left(\max_{1\le i\le n} X_i \le \theta + \frac{t}{n}\right) = \left(1 + \frac{t}{n\theta}\right)^n$$

et $\mathbb{P}_\theta\left(n(\hat{\theta}_n - \theta) \le t\right) = 0$ si $t < -n\theta$. Ainsi,

$$\mathbb{P}_\theta(n(\theta - \hat{\theta}_n) \le t) \xrightarrow[n\to+\infty]{} 1 - e^{-t/\theta}, \quad \forall t \ge 0.$$

On reconnait la fonction de répartition de la loi exponentielle de paramètre $1/\theta$. D'où

$$n(\hat{\theta} - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} -\text{Exp}(1/\theta).$$

Ainsi, l'EMV est ici de vitesse n et asymptotiquement exponentiel.

Démonstration Fixons $\theta \in \Theta$ dans toute la preuve. On définit

$$U_n(\alpha) = \log L_n(\mathbf{X};\alpha) = \sum_{i=1}^n \log L(X_i;\alpha), \quad \forall \alpha \in \Theta.$$

Rappelons que $\nabla U_n(\hat{\theta}_n) = 0$ car $\hat{\theta}_n$ est un EMV. La formule de Taylor avec reste intégral assure que

$$0 = \nabla U_n(\hat{\theta}_n) = \nabla U(\theta) + (\hat{\theta}_n - \theta)\int_0^1 \nabla^2 U_n(\theta + t(\hat{\theta}_n - \theta))\, dt,$$

de sorte que

$$-\frac{1}{\sqrt{n}}\nabla U_n(\theta) = \sqrt{n}(\hat{\theta}_n - \theta)\bar{U}_n,$$

où

$$\bar{U}_n = \frac{1}{n}\int_0^1 \nabla^2 U_n(\theta + t(\hat{\theta}_n - \theta))\, dt.$$

Or

$$\mathbb{V}_\theta(\nabla \log L(X_1;\theta)) = I(\theta) \quad \text{et} \quad \mathbb{E}_\theta \nabla \log L(X_1;\theta) = 0,$$

donc le théorème central limite (théorème 2.18) assure que

$$\frac{1}{\sqrt{n}}\nabla U_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla \log L(X_i;\theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}_d(0, I(\theta)).$$

Pour terminer la preuve du théorème 4.14, il nous suffit donc de démontrer que

$$\bar{U}_n \xrightarrow[n\to+\infty]{\mathbb{P}_\theta} -I(\theta), \tag{11}$$

puisqu'on en déduirait par le lemme de Slutsky (lemme 2.17) que

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} I(\theta)^{-1}\mathcal{N}_d(0, I(\theta)) = \mathcal{N}_d(0, I(\theta)^{-1}).$$

Montrons donc (11). Pour tout $x \in \mathcal{H}$ et tout $r > 0$, on définit

$$\sigma(x, r) = \sup_{\alpha,\theta\in\Theta \text{ t.q. } |\alpha-\theta|\le r} \left|\nabla^2 \log L(x;\alpha) - \nabla^2 \log L(x;\theta)\right|.$$

Le fait que le modèle est régulier assure que $\sigma(X_1, r) \in \mathbb{L}^1(\mathbb{P}_\theta)$ pour $r > 0$ suffisamment petit et, d'après le théorème de convergence dominée (théorème 2.2), pour $\varepsilon > 0$ fixé, il existe $r > 0$ suffisamment petit tel que $\mathbb{E}_\theta\sigma(X_1, r) < \varepsilon/2$. Or

$$\bar{U}_n = \frac{1}{n}\sum_{i=1}^n \int_0^1 \nabla^2 \log L(X_i;\theta + t(\hat{\theta}_n - \theta))\, dt,$$

donc

$$\begin{aligned}
\mathbb{P}_\theta\left(|I(\theta) + \bar{U}_n| \ge \varepsilon\right) &\le \mathbb{P}_\theta\left(\left|I(\theta) + \frac{1}{n}\sum_{i=1}^n \int_0^1 \nabla^2 \log L(X_i;\theta)\, dt\right| \ge \frac{\varepsilon}{2}\right) \\
&\quad + \mathbb{P}_\theta\left(\frac{1}{n}\sum_{i=1}^n \int_0^1 \left|\nabla^2 \log L(X_i;\theta) - \nabla^2 \log L(X_i;\theta + t(\hat{\theta}_n - \theta))\right| dt \ge \frac{\varepsilon}{2}\right) \\
&\le \mathbb{P}_\theta\left(\left|I(\theta) + \frac{1}{n}\sum_{i=1}^n \nabla^2 \log L(X_i;\theta)\right| \ge \frac{\varepsilon}{2}\right) \\
&\quad + \mathbb{P}_\theta\left(|\hat{\theta}_n - \theta| \ge r\right) + \mathbb{P}_\theta\left(\frac{1}{n}\sum_{i=1}^n \sigma(X_i, r) \ge \frac{\varepsilon}{2}\right).
\end{aligned}$$

Le premier terme du membre de droite converge vers 0 en $\mathbb{P}_\theta$-probabilité grâce à la loi des grands nombres (théorème 2.16) puisque $I(\theta) = -\mathbb{E}_\theta\nabla^2 \log L(X_i;\theta)$ d'après la proposition 4.12. Le second terme du membre de droite tend vers 0 car l'EMV

$\hat{\theta}_n$ est supposé consistant. Enfin, le troisième terme du membre de droite converge également vers 0 puisque, par la loi des grands nombres,

$$\frac{1}{n}\sum_{i=1}^{n}\sigma(X_i,r) \xrightarrow[n\to+\infty]{\mathbb{P}_\theta} \mathbb{E}_\theta\sigma(X_1,r) < \frac{\varepsilon}{2}.$$

Ceci termine la preuve du théorème 4.14. □

4.7 Propriétés théoriques de l'EMV

Dans le cas des échantillons i.i.d. (et sous certaines hypothèses techniques vues précédemment), l'EMV satisfait plusieurs « bonnes » propriétés qui montrent qu'il est optimal (en un certain sens). En particulier,

— il est **consistant** ;
— il est **asymptotiquement normal, de vitesse** $\sqrt{n}$;
— il est **asymptotiquement efficace**.

Nous avons déjà prouvé les deux premiers points, il nous reste à justifier le dernier. Afin de définir la notion d'efficacité asymptotique, nous énonçons et démontrons une popriété générale des estimateurs sans biais d'ordre 2.

Théorème 4.15 (Cramer-Rao) *Soit un modèle statistique* $(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$ *avec* $\mathcal{H} \subset \mathbb{R}^k$ *et* $\Theta \subset \mathbb{R}^d$*, et paramètre d'intérêt* $g(\theta)$ *pour une fonction* $g : \Theta \to \mathbb{R}$ $\mathcal{C}^1$*. On suppose que, pour tout* $\theta \in \Theta$*, il existe un voisinage* V *de* θ *dans* Θ *tel que la v.a.*

$$H = \sup_{\alpha\in V}\left|\frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)}\right|$$

est telle que $H \in \mathbb{L}^2(\mathbb{P}_\theta)$*. Sous ces hypothèses, si* $\hat{g}$ *est un estimateur d'ordre 2 de* $g(\theta)$ *sans biais, alors*

$$\mathcal{R}(\theta,\hat{g}) \geq \nabla g(\theta)^T I_n(\theta)^{-1}\nabla g(\theta), \tag{12}$$

où $\mathcal{R}(\theta,\hat{g})$ *est le risque quadratique de l'estimateur* $\hat{g}$ *de* $g(\theta)$ *défini en section 3.3. L'équation* (12) *s'appelle* ***borne de Cramer-Rao****.*

Exemple : Dans le jeu de pile-ou-face, puisque l'EMV $\bar{X}_n$ est sans biais, d'après la poposition 3.6,

$$\mathcal{R}(\theta,\bar{X}_n) = \mathbb{V}_\theta(\bar{X}_n) = \frac{1}{n}\mathbb{V}_\theta(X_1) = \frac{\theta(1-\theta)}{n}.$$

Or, on a vu en section 4.5 que $I_n(\theta) = \frac{n}{\theta(1-\theta)}$ dans ce modèle, de sorte que l'EMV du jeu de pile-ou-face réalise l'égalité dans la borne de Cramer-Rao (12). On déduit donc du théorème 4.15 que, dans le jeu de pile-ou-face, il n'existe pas de meilleur estimateur d'ordre 2 et sans biais que l'EMV, en terme de risque quadratique.

Le preuve du théorème 4.15 repose sur le lemme suivant, prouvé après le théorème.

Lemme 4.16 *Sous les hypothèses du théorème 4.15, pour tout $\theta \in \Theta$,*

$$\mathbb{E}_\theta \nabla \log L_n(\mathbf{X};\theta) = 0$$

et

$$\mathbb{E}_\theta \left[\hat{g} \nabla \log L_n(\mathbf{X};\theta)\right] = \nabla \mathbb{E}_\theta(\hat{g}).$$

Démonstration du théorème 4.15 Le lemme 4.16 implique que

$$\nabla g(\theta) = \mathbb{E}_\theta \left[\hat{g} \nabla \log L_n(\mathbf{X};\theta)\right] = \mathbb{E}_\theta \left[(\hat{g} - g(\theta)) \nabla \log L_n(\mathbf{X};\theta)\right].$$

Donc, pour tout $u \in \mathbb{R}^d$, en notant $\langle \cdot, \cdot \rangle$ le produit scalaire usuel,

$$\begin{aligned}\langle u, \nabla g(\theta)\rangle^2 &= \Big(\mathbb{E}_\theta \left[(\hat{g} - g(\theta)) \langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle\right]\Big)^2 \\ &\leq \mathcal{R}(\theta, \hat{g}) \mathbb{E}_\theta \langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle^2,\end{aligned}$$

où la dernière ligne découle de l'inégalité de Cauchy-Schwarz. Or, par définition de l'information de Fisher,

$$\begin{aligned}\mathbb{E}_\theta \langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle^2 &= u^T \mathbb{E}_\theta \left[\nabla \log L_n(\mathbf{X};\theta) \nabla \log L_n(\mathbf{X};\theta)^T\right] u \\ &= u^T \mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta)) u = u^T I_n(\theta) u.\end{aligned}$$

Donc, en choisissant $u = I_n(\theta)^{-1} \nabla g(\theta)$,

$$\mathbb{E}_\theta \langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle^2 = \nabla g(\theta)^T I_n(\theta)^{-1} \nabla g(\theta)$$

et

$$\langle u, \nabla g(\theta)\rangle = \nabla g(\theta)^T I_n(\theta)^{-1} \nabla g(\theta),$$

d'où

$$\mathcal{R}(\theta, \hat{g}) \geq \nabla g(\theta)^T I_n(\theta)^{-1} \nabla g(\theta). \qquad \square$$

Démonstration du lemme 4.16 La première équation découle de la proposition 4.12. Pour la seconde équation, la preuve suit la même démarche que celle de la proposition 4.12 : nous distinguons suivant le cas discret et le cas continu. Dans le cas discret,

$$\mathbb{E}_\theta \left[\hat{g} \nabla \log L_n(\mathbf{X};\theta)\right] = \sum_{\mathbf{x} \in \mathcal{H}^n} \hat{g}(\mathbf{x}) \frac{\nabla L_n(\mathbf{x};\theta)}{L_n(\mathbf{x};\theta)} \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x} \in \mathcal{H}^n} \hat{g}(\mathbf{x}) \nabla L_n(\mathbf{x};\theta).$$

Or, pour tout $\alpha \in V$,

$$|\hat{g}(\mathbf{x}) \nabla L_n(\mathbf{x};\alpha)| \leq \frac{1}{2} |\hat{g}(\mathbf{x})|^2 L_n(\mathbf{x};\theta) + \frac{1}{2} H^2 L_n(\mathbf{x};\theta) \tag{13}$$

est dans $\mathbb{L}^1(\mu)$, où $\mu = \sum_{\mathbf{x}\in\mathcal{H}^n} \delta_{\mathbf{x}}$. Ainsi, on peut appliquer le théorème de dérivation sous le signe somme (théorème 2.4) pour en déduire que

$$\mathbb{E}_\theta \left[\hat{g}\nabla \log L_n(\mathbf{X};\theta)\right] = \nabla \sum_{\mathbf{x}\in\mathcal{H}^n} \hat{g}(\mathbf{x})\mathbb{P}_\theta(\mathbf{X}=\mathbf{x}) = \nabla \mathbb{E}_\theta \hat{g}.$$

Dans le cas continu,

$$\mathbb{E}_\theta \left[\hat{g}\nabla \log L_n(\mathbf{X};\theta)\right] = \int_{\mathcal{H}^n} \hat{g}(\mathbf{x}) \frac{\nabla L_n(\mathbf{x};\theta)}{L_n(\mathbf{x};\theta)} f_\theta(\mathbf{x})\, dx = \int_{\mathcal{H}^n} \hat{g}(\mathbf{x}) \nabla L_n(\mathbf{x};\theta)\, dx.$$

Puisque le membre de droite de (13) est dans $\mathbb{L}^1(\text{Leb})$, où Leb est la mesure de Lebesgue sur $\mathcal{H}^n$, on peut de nouveau appliquer le théorème 2.4 pour en déduire que

$$\mathbb{E}_\theta \left[\hat{g}\nabla \log L_n(\mathbf{X};\theta)\right] = \nabla \int_{\mathcal{H}^n} \hat{g}(\mathbf{x}) f_\theta(\mathbf{x})\, dx = \nabla \mathbb{E}_\theta \hat{g}. \qquad \square$$

Nous pouvons maintenant définir la notion d'estimateur efficace.

Définition 4.17 — *Un estimateur $\hat{g}$ sans biais et d'ordre 2 est* ***efficace*** *si son risque quadratique atteint (c'est-à-dire réalise l'égalité dans) la borne de Cramer-Rao* (12).

— *Dans la suite de modèle statistiques $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\}_{\theta\in\Theta})$, une suite $\hat{g}_n$ d'estimateurs sans biais et d'ordre 2 est* ***asymptotiquement efficace*** *si*

$$\lim_{n\to+\infty} n\mathcal{R}(\theta, \hat{g}_n) = \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta).$$

Au vu de la normalité asymptotique de l'EMV (théorème 4.14), on a le résultat suivant.

Théorème 4.18 (efficacité asymptotique de l'EMV) *Si $\Theta \subset \mathbb{R}$ et si les hypothèses du théorème 4.14 sont vérifiées, l'EMV $\hat{\theta}_n$ est asymptotiquement efficace, c'est-à-dire*

$$\lim_{n\to+\infty} n\mathcal{R}(\theta; \hat{\theta}_n) = \frac{1}{I(\theta)}.$$

4.8 Intervalles de confiance et test de Wald

On peut se servir de la normalité asymptotique de l'EMV pour construire des intervalles de confiance et des tests d'hypothèses sur les paramètres d'un modèle statistique paramétrique i.i.d., de la forme $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\}_{\theta\in\Theta})$.

Dans le cas de paramètres de dimension 1, la construction d'un intervalle de confiance est classique et suit la démarche présentée en section 3.5. Rappelons que, pour tout $\alpha \in]0,1[$, q_α est le quantile d'ordre $1-\alpha/2$ de la loi gaussienne centrée réduite.

Théorème 4.19 (intervalle de confiance asymptotique) *Si $\Theta \subset \mathbb{R}$, les hypothèses du théorème 4.14 sont satisfaites, et l'application qui à $x \in \mathcal{H}$ associe*

$$\sup_{\alpha \in \Theta} \frac{(\nabla L(x;\alpha))^2}{L(x;\alpha)}$$

est $\mathbb{L}^1$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^k$ dans le cas continu (respectivement par rapport à la mesure $\mu = \sum_{x \in \mathcal{H}^n} \delta_x$ dans le cas discret), alors pour tout $\alpha \in]0,1[$,

$$\left[\hat{\theta}_n - \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}} \, , \, \hat{\theta}_n + \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}} \right] \tag{14}$$

est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour le paramètre θ.

Ce résultat peut être également utilisé pour déterminer le nombre de données à collecter pour garantir une précision d'estimation $\varepsilon > 0$ donnée du paramètre θ avec un niveau de confiance $1 - \alpha > 0$ donné. On commence par réaliser une première estimation grossière de θ avec un nombre limité n_0 de données. Ceci nous donne un premier intervalle de confiance I_0 de niveau α. Ensuite, on calcule en fonction de n la taille maximale de l'intervalle de confiance (14) sur I_0 :

$$\sup_{\theta \in I_0} \frac{q_\alpha}{\sqrt{nI(\theta)}} = \frac{1}{\sqrt{n}} \sup_{\theta \in I_0} \frac{q_\alpha}{\sqrt{I(\theta)}}$$

et on choisit pour n le premier entier tel que cette largeur soit inférieure au seuil de précision ε. C'est le nombre de mesures à réaliser afin d'obtenir une estimation avec précision inférieure à ε et niveau de confiance $1 - \alpha$.

Démonstration Fixons $\theta \in \Theta$. Le théorème 4.14 implique que

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \to +\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}(0, I(\theta)^{-1}). \tag{15}$$

La difficulté est que $I(\theta)$ est inconnu (puisque θ l'est), mais on peut utiliser la consistance de l'estimateur $\hat{\theta}_n$ pour démontrer que $I(\hat{\theta}_n)$ est un estimateur consistant de $I(\theta)$. Pour cela, nous avons besoin de montrer que la fonction I est continue. Or, dans le cas continu,

$$I(\theta) = -\mathbb{E}_\theta(\nabla \log L(X_1;\theta))^2 = -\int_{\mathcal{H}} \frac{(\nabla L(x;\theta))^2}{L(x;\theta)} \, dx.$$

Notre hypothèse de domination permet d'appliquer le théorème de continuité sous le signe somme (théorème 2.3) afin de déduire la continuité de I. Ainsi,

$I(\hat{\theta}_n)$ converge en probabilité vers $I(\theta)$. En combinant ce résultat avec (15), le lemme de Slutsky permet de déduire que

$$\sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}(0,1).$$

La construction d'un intervalle de confiance est ensuite classique : si $G \sim \mathcal{N}(0,1)$, $\mathbb{P}(|G| \leq q_\alpha) = 1-\alpha$, et donc, d'après le théorème de Portmanteau (théorème 2.8),

$$\lim_{n\to+\infty} \mathbb{P}\left(\theta \in \left[\hat{\theta}_n - \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}}\,,\, \hat{\theta}_n + \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}}\right]\right)$$
$$\lim_{n\to+\infty} \mathbb{P}\left(\sqrt{nI(\hat{\theta}_n)}\,|\hat{\theta}_n - \theta| \leq q_\alpha\right) = \mathbb{P}(|G| \geq q_\alpha) = 1-\alpha. \quad \square$$

Pour des paramètres en dimension 2 ou plus, on peut de la même manière construire des régions de confiance de forme ellipsoïdale. La méthode est la suivante : puisque $I(\theta)$ est une matrice de variance covariance supposée inversible, elle est symétrique définie positive. Il est alors classique de construire [5] une matrice $A(\theta)$ symétrique définie positive telle que $A^2(\theta) = I(\theta)$, appelée *racine carrée matricielle*. On déduit alors de la proposition 2.21 que

$$\sqrt{n}A(\theta)(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}_d(0, \mathrm{Id}).$$

Il est facile de construire une région de confiance pour un vecteur aléatoire $G \sim \mathcal{N}_d(0,\mathrm{Id})$, puisque $|G|^2$ suit la loi du $\chi^2(d)$ à d degrés de liberté. Pour tout $d \geq 2$ et $\alpha \in]0,1[$, on définit le quantile $q_{d,\alpha}$ de niveau $1-\alpha$ pour la loi $\chi^2(d)$, c'est-à-dire l'unique solution q de

$$\int_0^q \frac{(1/2)^{k/2}}{\Gamma(k/2)}\, r^{k/2-1} e^{-r/2}\, dr = 1-\alpha,$$

où $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\,dt$, de sorte que $\mathbb{P}(|G| \leq \sqrt{q_{d,\alpha}}) = 1-\alpha$.

Afin d'appliquer la méthode du théorème 4.19, il ne reste plus qu'à démontrer la consistance de l'estimateur $A(\hat{\theta}_n)$ de $A(\theta)$. Ceci découle de la continuité de l'application $\theta \mapsto A(\theta)$, qui est elle-même une conséquence

5. La méthode consiste à diagonaliser la matrice $I(\theta)$ avec un changement de base orthonormée de matrice P, de sorte que $P^{-1}I(\theta)P = P^T I(\theta)P = D = \mathrm{diag}(\lambda_1,\ldots,\lambda_d)$. La matrice A s'obtient alors en prenant les racines carrées des éléments diagonaux puis en appliquant le changement de base inverse, c'est-à-dire

$$A = P\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_d})P^{-1}.$$

de la continuité de la racine carrée matricielle (une propriété vraie sur l'ensemble des matrices définies positives, laissée en exercice [6]) et de la continuité de $\theta \mapsto I(\theta)$, qui se démontre comme dans le théorème 4.19. On obtient finalement le résultat suivant.

Théorème 4.20 (région de confiance asymptotique) *On suppose que $\Theta \subset \mathbb{R}^d$ avec $d \geq 2$, que les hypothèses du théorème 4.14 sont satisfaites, et que l'application qui à $x \in \mathcal{H}$ associe la matrice*

$$\sup_{\alpha \in \Theta} \left| \frac{\nabla L(x;\alpha)(\nabla L(x;\alpha))^T}{L(x;\alpha)} \right|$$

est $\mathbb{L}^1$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^k$ dans le cas continu (respectivement par rapport à la mesure $\mu = \sum_{x \in \mathcal{H}^n} \delta_x$ dans le cas discret), alors pour tout $\alpha \in]0,1[$,

$$\hat{\theta}_n + \sqrt{\frac{q_{d,\alpha}}{n}}\, A(\hat{\theta}^n)^{-1} B(0,1) = \left\{ \hat{\theta}_n + \sqrt{\frac{q_{d,\alpha}}{n}}\, A(\hat{\theta}^n)^{-1} u, \ \textit{t.q. } u \in B(0,1) \right\}$$

est une région de confiance de niveau asymptotique $1-\alpha$ pour le paramètre θ.

Remarquons que, de façon générale, pour toute matrice A de taille $d \times d$, l'ensemble $AB(0,1)$ est un ellipsoïde de $\mathbb{R}^d$ centré en 0. Ainsi, la région de confiance du résultat précédent est un ellipsoïde centré en $\hat{\theta}_n$ et dont le diamètre est d'ordre $1/\sqrt{n}$.

Nous terminons ce chapitre avec le test de Wald, qui est un test statistique portant sur le paramètre θ. Rappelons d'abord le contexte général des tests statistiques.

Un test d'hypothèse statistique vise à répondre par oui ou non à une question formulée en terme d'une hypothèse. On distingue les **tests non-paramétriques**, qui sont souvent non-asymptotiques et ne font pas intervenir de distributions connues (comme les lois gaussiennes ou du χ^2), voire pas d'hypothèse de modèle paramétrique du tout. C'est le cas par exemple des tests d'adéquation entre deux échantillons, où le but est de déterminer si les deux échantillons sont issus d'une même loi, non paramétrique. La famille de tests la plus répandue est cette des **tests paramétriques**, où les observations sont supposées suivre un modèle paramétrique, dont les hypothèses sont directement formulées en terme des paramètres du modèle et où on s'intéresser le plus souvent à des propriétés asymptotiques en la taille de l'échantillon.

Un test statistique consiste à définir deux hypothèses :

6. On peut par exemple utiliser le développement en série entière de la fonction $\sqrt{1+x}$.

— l'**hypothèse nulle**, notée souvent H_0, qui est l'hypothèse communément admise pour laquelle on souhaite savoir si les observations permettent de la réfuter ; sinon, l'hypothèse est conservée (pensez par exemple à la présomption d'innocence en justice) ;
— l'**hypothèse alternative**, notée souvent H_1, et qui n'est pas toujours spécifiée dans les applications, mais qui est généralement la négation de l'hypothèse nulle.

On se fixe un seuil de confiance $\alpha \in]0,1[$ (valeur typique $\alpha = 5\%$). Un test repose sur une fonction des observations, c'est-à-dire une statistique T, appelée **statistique de test**, à partir de laquelle un **critère de rejet** est défini, prenant la forme d'une **région de rejet** telle que, si la valeur observée de T est dans cette région, on dit que l'hypothèse H_0 est **rejetée** ; sinon, elle est dite **acceptée**. Le plus souvent, $T \in \mathbb{R}$ et la région de rejet a pour forme $\{T \geq t\}$ pour un **seuil de rejet** $t \in \mathbb{R}$ à déterminer.

On appelle **erreur de première espèce** la probabilité de rejeter H_0 lorsque H_0 est vraie (risque de **faux positif**). On dit que le test est **de niveau** α si l'erreur de première espèce est α, et est **de niveau asymptotique** α si la limite de l'erreur de première espèce quand la taille de l'échantillon tend vers l'infini tend vers α (on suppose ici que la statistique de test $T = T_n$ peut dépendre de n).

Dans le cas d'une statistique uni-dimensionnelle et d'une région de rejet de la forme $\{T \geq t\}$, on appelle **p-valeur** la probabilité sous l'hypothèse H_0 que T soit supérieure ou égale à t_0, la valeur observée de la statistique T. C'est une façon de mesurer la certitude avec laquelle on rejette l'hypothèse H_0 : plus la p-valeur est petite, plus la valeur observée t_0 est irréaliste sous l'hypothèse H_0.

Remarquons que le risque de première espèce et la p-valeur ne dépendent que de H_0 et des observations, et pas du choix particulier de l'hypothèse alternative H_1. Le **risque de seconde espèce**, souvent noté β, est la probabilité d'accepter H_0 alors que H_1 est vraie, et la **puissance du test** est la probabilité de rejeter H_0 lorsque H_1 est vraie. Une fois le niveau du test fixé, la qualité d'un test se mesure à la petitesse de son risque de seconde espèce.

Soit un modèle statistique paramétrique i.i.d. de la forme $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\}_{\theta\in\Theta})$. Le test de Wald est un test paramétrique visant à tester une valeur précise du paramètre θ, en exploitant la normalité asymptotique de l'estimateur du maximum de vraisemblance. Dans ce cas, l'hypothèse nulle a la forme

$$(H_0) \quad \theta = \theta_0$$

pour un θ_0 fixé. L'hypothèse alternative peut avoir la forme

$$(H_1) \quad \theta \neq \theta_0$$

ou

$$(H_1') \quad \theta = \theta_1$$

pour un certain $\theta_1 \neq \theta_0$.

Soit $\alpha \in]0,1[$. Puisque, comme nous l'avons vu plus haut,

$$\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} G,$$

où $G \sim \mathcal{N}_d(0, \mathrm{Id})$, on a

$$\left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta)\right|^2 \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \chi^2(d).$$

On définit donc la statistique de test

$$T_n = \left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)\right|^2,$$

le seuil de rejet $t = q_{d,\alpha}$ et la région de rejet $\{T_n \geq q_{d,\alpha}\}$. Le test ainsi construit est asymptotiquement de niveau α, puisque l'erreur de première espèce est $\mathbb{P}_{\theta_0}(T_n \geq q_{d,\alpha})$ et, d'après le théorème de Portmanteau (théorème 2.8),

$$\lim_{n\to+\infty} \mathbb{P}_{\theta_0}(T_n \geq q_{d,\alpha}) = \mathbb{P}(|G|^2 \geq q_{d,\alpha}) = \alpha.$$

On peut calculer l'erreur de seconde espèce sous l'hypothèse alternative (H_1') définie plus haut :

$$\beta_n = \mathbb{P}_{\theta_1}(T_n \leq q_{d,\alpha}).$$

Or, sous $\mathbb{P}_{\theta_1}$,

$$\begin{aligned}\sqrt{T_n} &= \left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1) + \sqrt{n}A(\hat{\theta}_n)(\theta_1 - \theta_0)\right| \\ &\geq \left|\sqrt{n}A(\hat{\theta}_n)(\theta_1 - \theta_0)\right| - \left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1)\right|,\end{aligned}$$

où le second terme du membre de droite converge en loi sous $\mathbb{P}_{\theta_1}$ vers $|G|$, et le premier terme du membre de droite est une constante de la forme $a\sqrt{n}$ pour un certain $a > 0$. On en déduit donc que

$$\beta_n = \mathbb{P}_{\theta_1}(\sqrt{T_n} \leq \sqrt{q_{d,\alpha}}) \leq \mathbb{P}_{\theta_1}\left(\left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1)\right| \geq a\sqrt{n} - \sqrt{q_{d,\alpha}}\right).$$

En particulier, pour toute constant $C > 0$, à partir d'un certain rang n,

$$\beta_n \leq \mathbb{P}_{\theta_1}\left(\left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1)\right| \geq C\right) \xrightarrow[n\to+\infty]{} \mathbb{P}(|G| \geq C).$$

Puisque ceci est vrai pour toute valeur de C et que $\mathbb{P}(|G| \geq C) \to 0$ quand $C \to +\infty$, on en déduit que

$$\lim_{n\to+\infty} \beta_n = 0.$$

Ainsi, la puissance du test $1 - \beta_n$ tend vers 1 lorsque $n \to +\infty$.

En utilisant la théorie des grandes déviations, il est même possible de démontrer que $\beta_n \leq Ce^{-bn}$ pour des constantes C et $b > 0$ explicites, de sorte que l'on peut déterminer une valeur de n, et donc le nombre de données à collecter, à partir de laquelle la puissance du test est aussi bonne qu'un seuil fixé au préalable. Pourvu que l'on puisse réaliser autant de mesures que l'on veut, il est donc possible d'accepter ou rejeter H_0 contre H_1' avec n'importe quel niveau de confiance fixé à l'avance.