Statistique et Apprentissage
Notes du cours MAP433

Gersende Fort, Matthieu Lerasle, Eric Moulines

16 juillet 2021

Notations

Ensembles

— $\mathbb{N}$: l'ensemble des entiers naturels, $\mathbb{N} = \{0, 1, 2, \ldots\}$.

— $\mathbb{N}^*$: l'ensemble des entiers naturels, privé de 0, $\mathbb{N}^* = \{1, 2, \ldots\}$.

— $\overline{\mathbb{N}}$: l'ensemble des entiers naturels étendu, $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$.

— $\mathbb{Z}$: l'ensemble des entiers relatifs, $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$.

— $\mathbb{R}$: l'ensemble des réels.

— $\mathbb{R}^d$: l'ensemble des vecteurs colonnes $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})'$.

— $\overline{\mathbb{R}}$: l'ensemble des réels complétés, $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

— $\lceil x \rceil$: le plus petit entier supérieur à x.

— $\lfloor x \rfloor$: le plus grand entier inférieur ou égal à x.

Espace métrique

— (X, d) : un espace métrique.

— Si $A \subset \mathsf{X}$, $d(x, A) = \inf\{d(x, y) : y \in A\}$.

— $\mathrm{B}(x, r)$: la boule ouverte de rayon $r > 0$ centrée en x,

$$\mathrm{B}(x, r) = \{y \in \mathsf{X} : d(x, y) < r\} .$$

— $\overline{U}$: la fermeture de $U \subset \mathsf{X}$ (l'intersection de l'ensemble des fermés contenant U).

— U^o : l'intérieur de U (l'union des ouverts inclus dans U).

— ∂U : la frontière de $U \subset \mathsf{X}$, $\partial U = \overline{U} \setminus U^o$.

Relations binaires

— $a \wedge b$: le minimum de a et b.

— $a \vee b$: le maximum de a et b

Soient $\{a_n,\ n \in \mathbb{N}\}$ et $\{b_n,\ n \in \mathbb{N}\}$ deux suites positives.

— $a_n \asymp b_n$: il existe deux constantes $0 < c$ et $C < \infty$ telles que $cb_n \leq a_n \leq CB_n$

— $a_n \sim b_n$: il existe une suite $\varepsilon_n \to 0$ telle que $(1 - \varepsilon_n)b_n \leq a_n \leq (1 + \varepsilon_n)b_n$.

Vecteurs, matrices

Par convention, les vecteurs sont des vecteurs colonne. Pour une matrice A, A^T désigne la matrice transposée de A.

— $\mathrm{Mat}_d(\mathbb{R})$: l'ensemble des matrices $d \times d$ à coefficients réels.
— Pour $x \in \mathbb{R}^d$, $\|x\|$ est la norme euclidienne de x.
— $\mathrm{Mat}_{n,p}(\mathbb{R})$: l'ensemble des matrices $n \times p$ à coefficients réels.
— Pour $M \in \mathrm{Mat}_d(\mathbb{C})$ et $\|\cdot\|$ une norme sur $\mathbb{R}^d$, $|||M|||$ est la norme opérateur

$$|||M||| = \sup\left\{ \frac{\|Mx\|}{\|x\|},\ x \in \mathbb{R}^d,\ x \neq 0 \right\} .$$

— I_d : matrice identité de taille $d \times d$.

Fonctions

— $\mathbb{1}_A$: fonction indicatrice de l'ensemble A, $\mathbb{1}_A(x) = 1$ si $x \in A$ et 0 autrement.
— f^+ : partie positive de la fonction f, i.e. $f^+(x) = f(x) \vee 0$,
— f^- : partie négative de f, i.e. $f^-(x) = -(f(x) \wedge 0)$.
— $f^{-1}(A)$: image réciproque de l'ensemble A par f.
— Si f est une fonction à valeurs réelles sur X, $\|f\|_\infty = \sup\{f(x) : x \in \mathsf{X}\}$ est la norme-sup.

Espaces de fonctions

Soit $(\mathsf{X}, \mathscr{X})$ un espace mesurable.

— $\mathbb{F}(X)$: l'espace mesurable des fonctions de $(\mathsf{X}, \mathscr{X})$ sur $]-\infty, \infty[$.
— $\mathbb{F}_+(X)$: le cône des fonctions mesurables de $(\mathsf{X}, \mathscr{X})$ sur $[0, \infty]$.
— $\mathbb{F}_b(X)$: le sous-ensemble de $\mathbb{F}(X)$ des fonctions bornées.
— Pour toute mesure ξ définie sur $(\mathsf{X}, \mathscr{X})$ et toute fonction $f \in \mathbb{F}_b(X)$, $\xi(f) = \int f \, \mathrm{d}\xi$.
— Si X est un espace topologique,
  — $\mathrm{C}_b(X)$ est l'espace de toutes les fonctions continues bornées sur X ;
  — $\mathrm{C}(\mathsf{X})$ est l'espace de toutes les fonctions continues sur X ;
— $\mathscr{L}^p(\mu)$: l'espace des fonctions mesurables f telles que $\int |f|^p \mathrm{d}\mu < \infty$.

Mesures

Soit $(\mathsf{X}, \mathscr{X})$ un espace mesurable.

— δ_x : mesure ponctuelle en x, i.e. $\delta_x(A) = 1$ si $x \in A$ et 0 autrement.
— $\lambda_{\mathrm{Leb}}^{\otimes d}$: mesure de Lebesgue sur $\mathbb{R}^d$; et λ_{Leb} : mesure de Lebesgue sur $\mathbb{R}$.
— $\mathbb{M}_{\mathrm{s}}(X)$: l'ensemble des mesures signées sur $(\mathsf{X}, \mathscr{X})$.
— $\mathbb{M}_+(X)$: l'ensemble des mesures positives sur $(\mathsf{X}, \mathscr{X})$.
— $\mathbb{M}_1(X)$: l'ensemble des probabilités sur $(\mathsf{X}, \mathscr{X})$.
— $\mu \ll \nu$: μ est absolument continue par rapport à ν.
— $\mu \sim \nu$: μ est équivalent à ν, i.e., $\mu \ll \nu$ et $\nu \ll \mu$.

Si X est un espace topologique (en particulier un espace métrique) alors $\mathscr{X} = \mathscr{B}(\mathsf{X})$ est la tribu borélienne, c'est-à-dire la tribu engendrée par la topologie de X.

Probabilité

Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité. Une variable aléatoire X est une fonction mesurable de $(\Omega, \mathscr{F})$ sur $(\mathsf{X}, \mathscr{X})$, où $(\mathsf{X}, \mathscr{X})$ est un espace mesurable.

— $\mathbb{E}(X)$ ou $\mathbb{E}[X]$: l'espérance (quand elle est définie) du vecteur aléatoire X par rapport à la probabilité $\mathbb{P}$,

$$\mathbb{E}[X] = \int X \mathrm{d}\mathbb{P} \, .$$

— $\mathrm{Cov}(X,Y)$: la matrice de covariance (ou matrice de variance-covariance) de taille $n \times m$, des vecteurs aléatoires $X \in \mathbb{R}^n$ et $Y \in \mathbb{R}^m$,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] \, .$$

— $\mathscr{L}_{\mathbb{P}}(X)$: distribution de la variable aléatoire X sur $(\mathsf{X}, \mathscr{X})$ sous $\mathbb{P}$, i.e. la probabilité image de $\mathbb{P}$ par X.

— $X_n \overset{\mathbb{P}}{\Longrightarrow} X$: la suite de variables aléatoires $\{X_n,\, n \in \mathbb{N}\}$ converge en loi vers X sous $\mathbb{P}$.

— $X_n \overset{\mathbb{P}-\mathrm{prob}}{\longrightarrow} X$: la suite de variables aléatoires $\{X_n,\, n \in \mathbb{N}\}$ converge vers X en probabilité sous $\mathbb{P}$.

— $X_n \overset{\mathbb{P}-\mathrm{p.s.}}{\longrightarrow} X$: la suite de variables aléatoires $\{X_n,\, n \in \mathbb{N}\}$ converge vers X $\mathbb{P}$-presque-sûrement.

Distributions usuelles

— $\mathrm{Ber}(p)$: distribution de Bernoulli de probabilité de succès p.

— $\mathrm{Bin}(n,p)$: distribution binômiale ; distribution du nombre de succès après n tirages indépendants de Bernoulli de probabilité de succès p.

— $\mathrm{N}(\mu, \sigma^2)$: distribution gaussienne (ou *normale*) de moyenne μ et variance σ^2.

— $\mathrm{Unif}(a,b)$: distribution uniforme sur $[a,b]$.

— χ^2_n : loi du chi-2 à n degrés de libertés

— $\mathrm{Poi}(\lambda)$: loi de Poisson d'intensité λ.

— $\mathrm{t}(n)$: loi de Student à n degrés de liberté.

Table des matières

Première partie

Inférence statistique

Chapitre I-1

Modèles statistiques

L'objectif de ce chapitre est d'introduire la notion de modèle statistique. Les notions principales sont celles d'observations, de modèle statistique, de modèle statistique induit, de modèle statistique dominé et de n-échantillon.
Des exemples de modèles classiques sont introduits : modèles de sondage et modèles de régression dont la régression logistique.

I-1.1 Exemples introductifs

Avant de procéder à la construction mathématique d'un modèle statistique, nous commençons par quelques exemples de problèmes concrets de statistiques.

Exemple I-1.1 (Sondage). Une élection entre deux candidats A et B a lieu : on effectue un sondage à la sortie des urnes. On interroge n votants, le nombre n étant considéré comme petit devant le nombre total de votants, et on obtient ainsi les nombres n_A et n_B de voix pour les candidats A et B respectivement ($n_A + n_B = n$, en ne tenant pas compte des votes blancs ou nuls pour simplifier). Les questions naturelles auxquelles nous tâcherons de répondre sont typiquement les suivantes.

— Quelle est la proportion d'électeurs ayant voté pour le candidat A ?

— Peut-on affirmer que A ou B a gagné au vu de n_A et n_B seulement ?

— Si l'on décide d'annoncer A (ou B) vainqueur, comment quantifier l'erreur de décision ? ♢

Exemple I-1.2 (Reconstruction d'un signal). On souhaite apprendre une fonction $t \mapsto f(t)$, et pour ce faire on dispose d'observations affectées par une erreur de mesure de la fonction inconnue f aux instants multiples de T_e, sur un intervalle de temps $[0,T]$: les relevés $y_k, k = 1,\dots,n$ dont on dispose sont donc des approximations des quantités inconnues $f(kT_e)$, $k = 1,\dots,n$, où $n := \lfloor T/T_e \rfloor$ - la partie entière inférieure de T/T_e.

Dans ce problème, on est typiquement intéressé par "estimer" la fonction f. Ceci signifie que nous allons chercher à construire une fonction $t \mapsto \widehat{f}\big(t;\, \{y_k\}_{k=1}^n\big)$ ne dépendant que de l'échantillon mesuré $\{y_k\}_{k=1}^n$, qui soit en un certain sens "proche" de la fonction inconnue f. La difficulté ici va dépendre de la période d'échantillonnage, du rapport entre l'amplitude du signal à reconstruire f et celle des erreurs de mesure $\{y_k - f(kT_e)\}_{k=1}^n$, et bien sûr de la "complexité" de la fonction f. Intuitivement, une fonction constante $f(t) = \mu$ ou ayant une forme prescrite, par exemple $f(t) = \beta_0 + \beta_1 t$ sera plus facile à reconstruire qu'une fonction très irrégulière. On peut s'intéresser au même problème en dimension supérieure, par exemple en dimension 2 où f peut représenter une image. La fonction f est alors définie sur le pavé $[0,1] \times [0,1]$. En pratique, cette image est discrétisée en pixels et la valeur de f en ce pixel quantifie un niveau de gris dans $\{0,\dots,M-1\}$. On collecte donc $n = N^2$ données $y_{k,\ell}$ pour $1 \le k,\ell \le N$, qui sont des mesures bruitées de la quantité inconnue $f(k/N, \ell/N)$. On cherche à reconstruire l'image f ou bien à décider si une certaine caractéristique est présente dans l'image. ♢

Exemple I-1.3 (Contrôle de qualité, données censurées). On cherche – en laboratoire – à tester la fiabilité d'un dispositif. On fait fonctionner en parallèle n appareils jusqu'à ce qu'ils tombent tous en panne. On note $x_1,\dots,x_n$ les instants de panne observés. On cherche alors à obtenir des informations sur la fiabilité du dispositif, par exemple

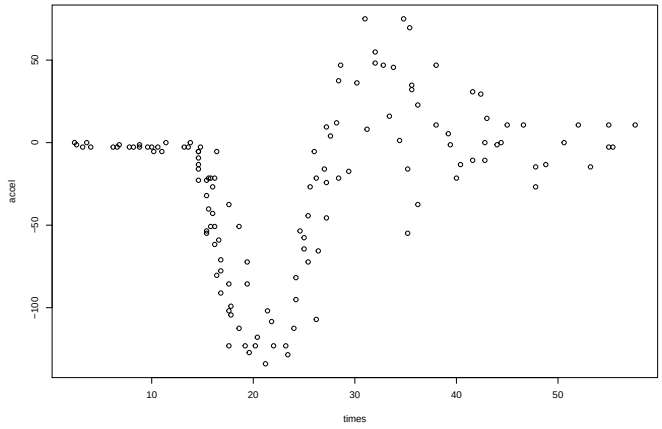

FIGURE I-1.1 – Accélération de la tête en fonction du temps suite à un impact

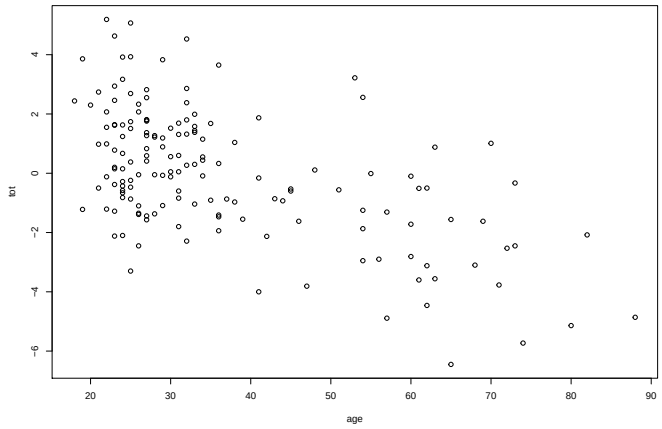

FIGURE I-1.2 – Qualité d'un greffon de rein en fonction de l'âge du donneur (données du laboratoire de néphrologie du Dr. B. Myers, Université de Stanford)

- garantir qu'avec une probabilité donnée, le système fonctionnera plus qu'un temps T donné.
- donner une valeur "représentative" de la durée de vie du dispositif.

Répondre à ces deux questions nécessite bien entendu de formaliser ce qu'on entend ici par probabilité ou "valeur représentative". Nous nous y attacherons dans la suite de l'exposé.

Si les appareils sont fiables et que le nombre n d'appareils testés est grand, attendre que tous les dispositifs soient tombés en panne peut s'avérer impossible. Une idée fréquemment utilisée en fiabilité est de fixer a priori un temps terminal τ et d'observer les temps de défaillance des systèmes apparaissant avant l'horizon τ, ce qui revient à observer $x_i^\star = \min\{x_i, \tau\}$, pour $i \in \{1,\dots,n\}$. Une question importante consiste à quantifier la perte d'information associée à l'horizon τ dans cette seconde expérience plus réaliste. ♢

Exemple I-1.4 (Influence d'une variable sur une autre). On s'intéresse à étudier le lien entre la qualité d'un greffon de rein (caractérisée par une note agrégeant un certain nombre de caractéristiques) et l'âge du donneur. On considère une population de n donneurs potentiels. Pour chaque donneur $i \in \{1,\dots,n\}$, on note x_i son âge et y_i la qualité du greffon. Il est bien entendu irréaliste de postuler l'existence d'une fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $y = f(x)$. Toutefois, il est raisonnable d'introduire un modèle statistique exprimant que y_i est en partie expliquée par x_i ; ce qui peut être fait en exprimant que chaque mesure y_i est une observation dans un bruit de mesure de $f(x_i)$ où f est une fonction inconnue. Le problème du statisticien sera ici de reconstruire la fonction de régression f et de caractériser l'erreur de

modélisation.

Le problème est proche de celui de l'exemple I-1.2, à ceci près que les points kT_e sont remplacés par les âges x_i. Les variables explicatives x_i ne sont pas nécessairement scalaires, on peut ainsi remplacer x_i par un vecteur $\mathbf{x}_i \in \mathbb{R}^k$ qui collecte un ensemble de variables explicatives possibles. Dans ce cas, f est une fonction $\mathbb{R}^k \to \mathbb{R}$.

On peut, si cette information est disponible, incorporer une information a priori sur la fonction f : par exemple, postuler que la fonction f est de la forme $f(\mathbf{x}) = t(\boldsymbol{\theta}^T \mathbf{x})$, où $t : \mathbb{R} \to \mathbb{R}$ est une fonction connue et $\theta = (\theta_1, \ldots, \theta_k)^T \in \mathbb{R}^k$ sont des paramètres inconnus. Dans les cas les plus élémentaires, $t(z) = z$, et on parle de *régression linéaire*. Il est bien entendu possible de considérer des modèles de régression plus généraux incorporant explicitement des non-linéarités et des interactions entre les différentes variables explicatives. La nature des informations exploitables pour construire un tel modèle dépend naturellement fortement des applications considérées. Nous reviendrons dans la suite sur l'importance centrale des modèles en statistique et des méthodes pour construire de tels modèles.

Il existe aussi des situations où y_i est une variable *qualitative*, c'est-à-dire ne prenant qu'un nombre fini de valeurs (instances). On peut penser que le risque d'être victime d'un infarctus est influencé par un ensemble de facteurs : consommation de tabac, d'alcool, taux de cholestérol, indice de masse corporelle, âge, terrain familial, pression artérielle, ces différents facteurs étant eux-mêmes souvent liés. ♢

I-1.2 Formulation mathématique

Construire un modèle statistique consiste à identifier trois éléments distincts :

Des données $z = (x_1, \ldots, x_n)$ où pour tout $i \in \{1, \ldots, n\}$, $x_i \in \mathbb{R}^d$ sont des vecteurs, mais on peut imaginer des situations plus complexes [1]. Ces données sont associées à la réalisation d'une expérience, et le point de départ du statisticien est donc le résultat de cette expérience.

Une famille de lois de probabilité associée à l'expérience qui a engendré les observations $Z = (X_1, \cdots, X_n)$ dont les mesures $z = (x_1, \cdots, x_n)$ sont une réalisation. Comme nous le verrons, spécifier un modèle statistique revient à postuler que la loi des observations est un élément d'une certaine famille de lois de probabilités. Cette loi traduit la connaissance a priori disponible sur la façon dont les observations sont produites. Par exemple, pour un sondage, l'observation dont nous disposons dépend de l'échantillon de la population qui a été interrogée : on peut avoir par exemple sélectionné des individus "au hasard" dans une population ; on peut avoir stratifié la population par sexes, classes d'âges, catégories socio-professionnelles, puis sélectionné "au hasard" un certain nombre d'individus dans chacune des strates. Chacune de ces procédures pour "construire" des échantillons conduit à différentes lois pour les observations – nous en discuterons dans la suite. En statistique, la loi des observations est le plus souvent connue de façon "partielle" – à la différence du calcul des probabilités, où la loi est toujours supposée comme une donnée initiale du problème. Les observations dont nous disposons vont nous permettre d'"affiner" notre compréhension du mécanisme de génération des données.

Une problématique associée au couple [données, modèle]. On s'intéressera principalement à estimer des paramètres inconnus, ou à prédire la valeur d'un signal $f(t_\star)$ en un point $t_\star$ non observé ; il faut alors contrôler la qualité de cette estimation et prédiction. On cherchera aussi à vérifier une "hypothèse" sur le mécanisme de génération des observations.

I-1.2.1 Modèles statistiques

Définition I-1.5 (Modèle statistique). *Un* modèle statistique *est la donnée de :*

— *un espace mesurable* $(\mathsf{Z}, \mathscr{Z})$*, dit l'*espace des observations*,*

— *une famille de probabilités* $\mathscr{C}$ *sur* $(\mathsf{Z}, \mathscr{Z})$.

Nous notons $(\mathsf{Z}, \mathscr{Z}, \mathscr{C})$ *le modèle statistique.*

Le modèle est dit paramétrique lorsque $\mathscr{C}$ *correspond à une famille* $\{\mathbb{P}_\theta\,,\ \theta \in \Theta\}$*, où* Θ *est un sous-ensemble de* $\mathbb{R}^d$*, avec* $d \geq 1$ *; i.e., il existe une fonction associant à chaque* $\theta \in \Theta$ *un élément* $\mathbb{P}_\theta$ *de* $\mathscr{C}$*.*

Le modèle est identifiable si la fonction $\theta \mapsto \mathbb{P}_\theta$ *est injective, i.e. si* $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ *implique* $\theta = \theta'$*.*

1. On peut considérer des données qualitatives, que l'on pourra coder par des entiers. On peut aussi considérer des données de type surface ou des trajectoires d'un processus stochastique.

Exemple I-1.6 (Modèle de sondage). Reprenons l'exemple I-1.1. On dispose d'une population de N individus (N est typiquement très grand). Considérons que l'expérience consiste à un tirage uniforme de n individus, avec replacement dans la population. On appelle x_i la réponse du i-ème individu sondé : $x_i = 1$ si le i-ème individu sondé vote A et $x_i = 0$ sinon. Le modèle statistique associé est alors donné par

$$\Big(\{0,1\}^n, \mathscr{P}(\{0,1\}^n), \big\{ \mathrm{Ber}_\theta^{\otimes n}, \theta \in \Theta \big\}\Big),$$

où $\mathscr{P}(\{0,1\}^n)$ est l'ensemble des parties de $\{0,1\}^n$ et $\mathrm{Ber}_\theta^{\otimes n}$ est le produit de n lois de Bernoulli de paramètre θ (voir Appendice A.1.5 et Appendice A.3.5), i.e.

$$\mathrm{Ber}_\theta^{\otimes n}(\{x_1,\dots,x_n\}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{n-\Sigma_{i=1}^n x_i} .$$

Par ce choix de la famille de lois, on exprime que les tirages sont indépendants et le résultat de chaque tirage est une variable de Bernoulli de paramètre $\theta \in \Theta$.

Une autre manière de mener l'expérience est de considérer que l'on n'observe que le nombre de votants n_A pour le candidat A. Dans ce cas, l'espace des observations est $\mathsf{Z} := \{0,\dots,n\}$ qui est muni de la tribu $\mathscr{P}(\mathsf{Z})$ - l'ensemble des parties de Z. La famille $\mathscr{C}$ est paramétrique : pour chaque $\theta \in \Theta := [0,1]$, nous prendrons pour $\mathbb{P}_\theta$ la loi binomiale de paramètres (n,θ)

$$\mathbb{P}_\theta(\{k\}) = \binom{n}{k}\theta^k(1-\theta)^{n-k} , \quad k \in \{0,\dots,n\} .$$

Le modèle statistique s'écrit

$$(\{0,\dots,n\}, \mathscr{P}(\{0,\dots,n\}), \{\mathrm{Bin}(n,\theta),\ \theta \in \Theta\}) .$$

Supposons maintenant que les tirages sont sans replacement. On veut introduire un modèle qui exprime que dans la population, $N\theta$ individus votent pour le candidat A, où $\theta \in \Theta := \{k/N,\ k \in \{0,\dots,N\}\}$. L'espace des observations est encore $\mathsf{Z} := \{0,\dots,n\}$ et $\mathscr{Z} := \mathscr{P}(\mathsf{Z})$. Par contre, la loi des observations est cette fois donnée par

$$\mathbb{P}_\theta(\{k\}) = \begin{cases} \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}}, & \text{si } k \in \{\max(n-N(1-\theta),0),\dots,\min(N\theta,n)\}, \\ 0, & \text{si } k \notin \{\max(n-N(1-\theta),0),\dots,\min(N\theta,n)\}. \end{cases}$$

C'est le nombre de façons de choisir k individus dans une population de taille $N\theta$, puis indépendamment $(n-k)$ individus dans une population de taille $N(1-\theta)$ divisé par le nombre total de façons de choisir n individus parmi N. La loi $\mathbb{P}_\theta$ définie ci-dessus est appelée *hypergéométrique*, notée $\mathrm{Hyper}(N\theta,N,n)$. ♢

Exemple I-1.7. Dans le problème de reconstruction d'un signal (Exemple I-1.2), les observations sont clairement à valeurs dans $\mathsf{Z} := \mathbb{R}^n$ avec $n := \lfloor T/T_e \rfloor$. Il est naturel de munir cet ensemble de sa tribu borélienne (voir Définition A.3) : $\mathscr{Z} := \mathscr{B}(\mathbb{R}^n)$. Pour construire un modèle statistique, il faut se donner un ensemble $\mathscr{C}$ de lois possibles pour les observations. Supposons que :

— la fonction f est modélisée par une combinaison linéaire de fonctions de base connues $f(t) = \sum_{k=1}^p \beta_k \phi_k(t) \in \mathbb{R}$,

— les erreurs de mesure sont modélisées comme la réalisation de variables aléatoires gaussiennes centrées et de même variance σ^2.

L'ensemble des paramètres inconnus est donc ici $\theta = (\beta_1,\dots,\beta_p,\sigma^2)$ où $(\beta_1,\dots,\beta_p) \in \mathbb{R}^p$ et $\sigma^2 \in \mathbb{R}_+$. Nous poserons donc $\Theta := \mathbb{R}^p \times \mathbb{R}_+^*$. Par suite, pour tout $\theta \in \Theta$, dans le modèle $\mathbb{P}_\theta$, la loi du vecteur d'observations est la loi Gaussienne sur $\mathbb{R}^n$, d'espérance $(\mu_1(\theta),\mu_2(\theta),\dots,\mu_n(\theta))$ avec $\mu_i = \sum_{\ell=1}^p \beta_\ell \phi_\ell(iT_e)$ et de matrice de covariance $\sigma^2 \mathrm{I}_{n\times n}$. Nous écrivons le modèle statistique suivant :

$$\big(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \big\{ p_\theta \cdot \lambda_{\mathrm{Leb}}^{\otimes n},\ \theta \in \Theta := \mathbb{R}^p \times \mathbb{R}_+^* \big\}\big) ,$$

où

$$p_\theta(x_1,\dots,x_n) := (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \left(x_i - \sum_{k=1}^p \beta_k \phi_k(iT_e) \right)^2 \right) . \qquad ♢$$

Définition I-1.8 (Statistique). *Soient $(\mathsf{Z},\mathscr{Z},\mathscr{C})$ un modèle statistique et $(\mathsf{T},\mathscr{T})$ un espace mesurable. On appelle* statistique *sur le modèle statistique $(\mathsf{Z},\mathscr{Z},\mathscr{C})$ une application mesurable T de $(\mathsf{Z},\mathscr{Z})$ à valeurs dans $(\mathsf{T},\mathscr{T})$.*

Remarquons, et cela est très important, que T *ne dépend pas* de la loi $\mathbb{P} \in \mathscr{C}$ et donc ne dépend pas du paramètre si le modèle est paramétrique.

Pour toute loi $\mathbb{P} \in \mathscr{C}$, la statistique T comme application de $(\mathsf{Z}, \mathscr{Z})$ dans $(\mathsf{T}, \mathscr{T})$ est un élément aléatoire dont on note $\mathbb{P}^T$ la loi de probabilité (qui est la loi image de $\mathbb{P}$ par T). En écrivant

$$\mathscr{C}^T = \left\{ \mathbb{P}^T \,,\, \mathbb{P} \in \mathscr{C} \right\} \,,$$

on obtient ainsi le modèle statistique $(\mathsf{T}, \mathscr{T}, \mathscr{C}^T)$ induit par la statistique T.

Définition I-1.9 (Statistiques indépendantes). *Nous dirons que les statistiques X et Y sur* $(\mathsf{Z}, \mathscr{Z}, \mathscr{C})$ *sont indépendantes si pour toute loi* $\mathbb{P} \in \mathscr{C}$*, les éléments aléatoires X et Y sont indépendants sous* $\mathbb{P}$ *:*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \,.$$

Pour servir de support à l'intuition, il est souvent pratique d'introduire des statistiques associées aux observations individuelles. On parlera *d'observations individuelles* $X_1, \cdots, X_n$; on dit aussi que X_i est la *statistique associée* à la i-ème donnée collectée.

Exemple I-1.10 (Modèle de sondage (suite)). Considérons tout d'abord le modèle statistique donné par

$$(\mathsf{Z}, \mathscr{Z}, \mathscr{C}) = \left(\{0,1\}^n, \mathscr{P}(\{0,1\}^n), \left\{ \mathrm{Ber}_\theta^{\otimes n} \,,\, \theta \in \Theta := [0,1] \right\} \right) .$$

Notons X_i la statistique correspondant au résultat du i-ème sondage, $i \in \{1, \ldots, n\}$: pour tout $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$, $X_i(\mathbf{x}) = x_i$. En notant $\mathbb{P}_\theta = \mathrm{Ber}_\theta^{\otimes n}$, la loi image $\mathbb{P}_\theta^{X_i}$ est alors une loi de Bernoulli de paramètre $\theta \in \Theta$. On vérifie aussi que les statistiques $X_1, \ldots, X_n$ sont indépendantes : pour tout $(x_1, \ldots, x_n) \in \{0,1\}^n$, nous avons puisque $\mathbb{P}_\theta$ est $\mathrm{Ber}_\theta^{\otimes n}$

$$\mathbb{P}_\theta(\{X_1 = x_1, \ldots, X_n = x_n\}) = \mathbb{P}_\theta^Z(\{x_1, \ldots, x_n\}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \prod_{i=1}^n \mathbb{P}_\theta(\{X_i = x_i\}) \,,$$

où $Z = (X_1, \cdots, X_n)$. Pour le modèle *modelecanon*, les statistiques $(X_1, \ldots, X_n)$ sont indépendantes et identiquement distribuées (i.i.d.) : pour tout $\theta \in \Theta$ et $i \in \{1, \ldots, n\}$, l'observation X_i est distribuée suivant une loi de Bernoulli de paramètre θ. ♢

Exemple I-1.11. Considérons le modèle

$$(\mathsf{Z}, \mathscr{Z}, \mathscr{C}) = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{ \mathbb{P}_\theta \,,\, \theta \in \Theta := \mathbb{R} \times \mathbb{R}_+^* \right\})$$

où pour tout $\theta \in \Theta$, $\mathbb{P}_\theta := p_\theta \cdot \lambda_{\mathrm{Leb}}^{\otimes n}$ où p_θ est une densité gaussienne sur $\mathbb{R}^n$ de moyenne $\mu \mathbf{1}_n$ ($\mathbf{1}_n$ désigne le vecteur $(1, \cdots, 1)$ de longueur n) et de covariance $\sigma^2 \mathrm{I}_{n \times n}$:

$$p_\theta(x_1, \ldots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \,.$$

Notons X_i la i-ème observation, $i \in \{1, \ldots, n\}$. Cette statistique est définie, pour $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ par

$$X_i(\mathbf{x}) = x_i \,.$$

Pour tout $\theta \in \Theta$, la loi image de $\mathbb{P}_\theta$ par la statistique X_i est une loi gaussienne sur $\mathbb{R}$ de moyenne μ et de variance σ^2, dont la densité est donnée par

$$p_\theta^{X_i}(x_i) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) \,.$$

Sous $\mathbb{P}_\theta$, les statistiques $Z = (X_1, \ldots, X_n)$ sont i.i.d. de densité gaussienne de moyenne $\mu \in \mathbb{R}$ et variance $\sigma^2 \in \mathbb{R}_+^*$, puisque nous avons :

$$p_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n p_\theta^{X_i}(x_i) \,.$$

♢

Définition I-1.12 (Famille dominée et modèle statistique dominé). *Soient* $(\mathsf{Z},\mathscr{Z})$ *un espace mesurable et* μ *une mesure* σ*-finie sur* $(\mathsf{Z},\mathscr{Z})$*. Une* famille *de loi* $\mathscr{C}$ *est* dominée *par la mesure* μ *si toute loi de* $\mathbb{P}\in\mathscr{C}$ *admet une densité de probabilité* p *par rapport à* μ*, i.e.* $\mathbb{P}=p\cdot\mu$*.*

Le modèle statistique $(\mathsf{Z},\mathscr{Z},\mathscr{C})$ *est dit* dominé *lorsque la famille de lois* $\mathscr{C}$ *est dominée.*

Dans une section qui peut être omise en première lecture (voir Section I-1.3), nous discutons de l'unicité de la mesure de domination ; et donnons des exemples de modèles statistiques non dominés.

I-1.2.2 Modèle statistique produit et n-échantillon

Soient $(\mathsf{X},\mathscr{X},\mathscr{C})$ et $(\mathsf{X}',\mathscr{X}',\mathscr{C}')$ deux modèles statistiques. Nous allons construire le produit de ces deux modèles. Un élément z de l'espace des observations $\mathsf{Z}:=\mathsf{X}\times\mathsf{X}'$ est donc un couple $z:=(x,x')$ avec $x\in\mathsf{X}$ et $x'\in\mathsf{X}'$. Nous munissons cet espace de la tribu produit $\mathscr{Z}:=\mathscr{X}\otimes\mathscr{X}'$ qui est par définition la plus petite tribu qui contient les pavés mesurables $A\times A'$ où $A\in\mathscr{X}$ et $A'\in\mathscr{X}'$ (voir Appendice A.1.5). Pour $\mathbb{Q}\in\mathscr{C}$ et $\mathbb{Q}'\in\mathscr{C}'$, $\mathbb{Q}\otimes\mathbb{Q}'$ est la mesure produit (voir Appendice A.3.5) définie, pour tout $(A,A')\in\mathscr{X}\times\mathscr{X}'$ par :

$$\mathbb{Q}\otimes\mathbb{Q}'(A\times A'):=\mathbb{Q}(A)\mathbb{Q}'(A')\,.$$

Considérons le modèle statistique

$$(\mathsf{Z},\mathscr{Z},\mathscr{C}\otimes\mathscr{C}')\quad\text{où}\quad\mathscr{C}\otimes\mathscr{C}':=\left\{\mathbb{Q}\otimes\mathbb{Q}',\,\mathbb{Q}\in\mathscr{C},\mathbb{Q}'\in\mathscr{C}'\right\}\,. \tag{I-1.1}$$

Soient $X:\mathsf{Z}\to\mathsf{X}$ et $X':\mathsf{Z}\to\mathsf{X}'$ les statistiques définies pour tout $z=(x,x')\in\mathsf{Z}=\mathsf{X}\times\mathsf{X}'$ par

$$X(x,x')=x\quad\text{et}\quad X'(x,x')=x'\,.$$

Pour tout $\mathbb{P}=\mathbb{Q}\otimes\mathbb{Q}'\in\mathscr{C}\otimes\mathscr{C}'$ et $(A,A')\in\mathscr{X}\times\mathscr{X}'$, nous avons

$$\begin{aligned}\mathbb{P}^{X,X'}(A\times A')&=\mathbb{P}(X\in A,X'\in A')\\&=\mathbb{Q}\otimes\mathbb{Q}'(A\times A')=\mathbb{Q}(A)\mathbb{Q}'(A')=\mathbb{P}(X\in A)\mathbb{P}(X'\in A')\,.\end{aligned}$$

Les statistiques X et X' sont donc indépendantes. Les modèles statistiques induits par X et X' sont respectivement $(\mathsf{X},\mathscr{X},\mathscr{C})$ et $(\mathsf{X}',\mathscr{X}',\mathscr{C}')$. Considérer des produits de modèles statistiques correspond, dans la pratique, à étudier des systèmes d'observations indépendantes.

Définition I-1.13 (Produit de modèles statistiques). *Soient* $(\mathsf{X},\mathscr{X},\mathscr{C})$ *et* $(\mathsf{X}',\mathscr{X}',\mathscr{C}')$ *deux modèles statistiques. On appelle* produit de ces deux modèles *et on note* $(\mathsf{X},\mathscr{X},\mathscr{C})\otimes(\mathsf{X}',\mathscr{X}',\mathscr{C}')$ *le modèle statistique* $(\mathsf{X}\times\mathsf{X}',\ \mathscr{X}\otimes\mathscr{X}',\ \mathscr{C}\otimes\mathscr{C}')$ *où :*

$$\mathscr{C}\otimes\mathscr{C}':=\left\{\mathbb{P}=\mathbb{Q}\otimes\mathbb{Q}',\,\mathbb{Q}\in\mathscr{C},\,\mathbb{Q}'\in\mathscr{C}'\right\}\,.$$

Nous allons maintenant définir le modèle statistique associé à un *n-échantillon*. La construction est tout à fait similaire à celle d'un produit. Soient $(\mathsf{X},\mathscr{X},\mathscr{C})$ un modèle statistique et $n\in\mathbb{N}^*$. Posons

$$\mathsf{Z}:=\mathsf{X}^n,\qquad\mathscr{Z}:=\mathscr{X}^{\otimes n}\,.$$

Un élément z de l'espace des observations Z est donc un n-uplet $z=(x_1,\ldots,x_n)$ où, pour tout $i\in\{1,\ldots,n\}$, $x_i\in\mathsf{X}$. Soit $\mathscr{X}^{\otimes n}:=\mathscr{X}\otimes\cdots\otimes\mathscr{X}$ la tribu produit (voir Appendice A.1.5) i.e.la plus petite tribu contenant

les pavés mesurables $A_1 \times \cdots \times A_n$, où $(A_1, \ldots, A_n) \in \mathscr{X}^n$. Pour tout $\mathbb{Q} \in \mathscr{C}$, notons $\mathbb{Q}^{\otimes n}$ la probabilité produit sur $\mathscr{X}^{\otimes n}$ (voir Appendice A.3.5) définie, pour tout $(A_1, \ldots, A_n) \in \mathscr{X}^n$ par

$$\mathbb{Q}^{\otimes n}(A_1 \times \cdots \times A_n) := \prod_{i=1}^{n} \mathbb{Q}(A_i) \, .$$

Considérons le modèle statistique

$$(\mathsf{X}, \mathscr{X}, \mathscr{C})^n := (\mathsf{X}^n, \mathscr{X}^{\otimes n}, \{\mathbb{Q}^{\otimes n}, \, \mathbb{Q} \in \mathscr{C}\}) \, . \tag{I-1.2}$$

Remarquons que $(\mathsf{X}, \mathscr{X}, \mathscr{C})^n$ n'est pas (en général) égal au produit des expériences $(\mathsf{X}, \mathscr{X}, \mathscr{C})$. Pour $n = 2$ par exemple,

$$(\mathsf{X}, \mathscr{X}, \mathscr{C}) \otimes (\mathsf{X}, \mathscr{X}, \mathscr{C}) = (\mathsf{X}^2, \mathscr{X}^{\otimes 2}, \{\mathbb{Q} \otimes \mathbb{Q}', \, \mathbb{Q} \in \mathscr{C}, \mathbb{Q}' \in \mathscr{C}\})$$

alors que

$$\mathscr{C}^2 = (\mathsf{X}^2, \mathscr{X}^{\otimes 2}, \{\mathbb{Q}^{\otimes 2}, \, \mathbb{Q} \in \mathscr{C}\}) \, .$$

Définition I-1.14 (n-échantillon). *Soient $(\mathsf{X}, \mathscr{X}, \mathscr{C})$ un modèle statistique et $n \in \mathbb{N}^*$. On appelle n-*échantillon *de $(\mathsf{X}, \mathscr{X}, \mathscr{C})$ le modèle statistique*

$$(\mathsf{X}, \mathscr{X}, \mathscr{C})^n := (\mathsf{X}^n, \ \mathscr{X}^{\otimes n}, \{\mathbb{Q}^{\otimes n}, \, \mathbb{Q} \in \mathscr{C}\}) \, .$$

On appelle i-ème observation (canonique) la statistique X_i définie pour $z = (x_1, \ldots, x_n) \in \mathsf{X}^n$ par $X_i(z) = x_i$.

Lemme I-1.15. *Soit $(\mathsf{X}, \mathscr{X}, \mathscr{C})^n$ un n-échantillon du modèle $(\mathsf{X}, \mathscr{X}, \mathscr{C})$.*

(i) Les observations $Z = (X_1, \ldots, X_n)$ sont indépendantes.

(ii) Pour tout $i \in \{1, \ldots, n\}$, le modèle induit par la statistique X_i est $(\mathsf{X}, \mathscr{X}, \mathscr{C})$.

Démonstration. En effet, pour tout $\mathbb{P} = \mathbb{Q}^{\otimes n} \in \mathscr{C}^{\otimes n}$ et $(A_1, \ldots, A_n) \in \mathscr{X}^n$, nous avons

$$\begin{aligned} \mathbb{P}^Z(A_1 \times \cdots \times A_n) &= \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) \\ &= \mathbb{Q}^{\otimes n}(A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} \mathbb{Q}(A_i) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) \, . \end{aligned}$$

□

Soit $(\mathsf{X}, \mathscr{X}, \{\mathbb{Q}_\theta, \, \theta \in \Theta\})$ un modèle paramétrique. Supposons qu'il existe une mesure σ-finie μ sur $(\mathsf{X}, \mathscr{X})$ telle que la famille paramétrique $\{\mathbb{Q}_\theta, \, \theta \in \Theta\}$ soit dominée par rapport à μ. Notons, pour tout $\theta \in \Theta$, $q_\theta(\cdot)$ la densité de la loi $\mathbb{Q}_\theta$ par rapport à la mesure de domination μ :

$$\mathbb{Q}_\theta = q_\theta(\cdot) \cdot \mu \, .$$

Pour tout $\theta \in \Theta$, la loi produit $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ est dominée par $\mu^{\otimes n}$ et admet pour densité

$$(x_1, \ldots, x_n) \mapsto p_\theta(x_1, \ldots, x_n) := \prod_{i=1}^{n} q_\theta(x_i) \, . \tag{I-1.3}$$

Exemple I-1.16. Pour l'exemple I-1.3 du contrôle de qualité, un modèle classique de durée de vie est fourni par la famille de lois exponentielles de paramètre $\theta \in \mathbb{R}_+^*$, de densité par rapport à la mesure de Lebesgue λ_{Leb} donnée par

$$q_\theta(x) := \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x) \, . \tag{I-1.4}$$

Le n-échantillon du modèle statistique $\left(\mathbb{R}_+, \mathscr{B}(\mathbb{R}_+), \{q_\theta \cdot \lambda_{\text{Leb}}, \, \theta \in \Theta := \mathbb{R}_+^*\}\right)$ est

$$\left(\mathbb{R}_+^n, \mathscr{B}(\mathbb{R}_+^n), \{q_\theta^{\otimes n} \cdot \lambda_{\text{Leb}}^{\otimes n}, \, \theta \in \Theta\}\right) \, ,$$

où $\lambda_{\mathrm{Leb}}^{\otimes n}$ est la mesure de Lebesgue sur $\mathbb{R}^n$ et

$$q_\theta^{\otimes n}(x_1,\dots,x_n)=\prod_{i=1}^n q_\theta(x_i).$$

Pour tout $\theta\in\Theta$, sous $\mathbb{P}_\theta$, les observations $(X_1,\dots,X_n)$, définies par $X_i(x_1,\dots,x_n)=x_i$ pour tout $i\in\{1,\dots,n\}$ et $(x_1,\dots,x_n)\in\mathbb{R}_+^n$ sont indépendantes. Pour tout i, l'expérience statistique induite par X_i est

$$\left(\mathbb{R}_+,\mathscr{B}(\mathbb{R}_+),\left\{q_\theta\cdot\lambda_{\mathrm{Leb}}\,,\ \theta\in\Theta:=\mathbb{R}_+^*\right\}\right)\ .$$

Supposons que les observations $(X_1,\dots,X_n)$ soient censurées à un instant terminal τ connu. Pour $i\in\{1,\dots,n\}$, nous considérons les statistiques

$$X_i^*:=\min(X_i,\tau)\ ,\quad i\in\{1,\dots,n\}\ .$$

Pour $\theta\in\Theta$, notons $\mathbb{Q}_\theta^*$ la loi image $\mathbb{Q}_\theta^{X_i^*}$. Le modèle statistique induit par la variable X_i^* est donc

$$([0,\tau],\mathscr{B}([0,\tau]),\{\mathbb{Q}_\theta^*\,,\ \theta\in\Theta\})\ .$$

Notons que la loi $\mathbb{Q}_\theta^*$ n'est pas absolument continue par rapport à la mesure de Lebesgue. En revanche, la famille $\left\{\mathbb{Q}_\theta^*\,,\ \theta\in\Theta\right\}$ est dominée par $\mu=\lambda_{\mathrm{Leb}}+\delta_\tau$, où δ_τ désigne la mesure de Dirac en τ. Pour tout $\theta\in\Theta$, la densité de $\mathbb{Q}_\theta^*$ par rapport à μ est donnée par

$$q_\theta^*(x):=\theta\mathrm{e}^{-\theta x}\mathbb{1}_{\{x<\tau\}}+c(\theta)\mathbb{1}_{\{x=\tau\}}\ ,\quad \text{avec } c(\theta):=\int_\tau^{+\infty}\theta\mathrm{e}^{-\theta t}\mathrm{d}t=\mathrm{e}^{-\theta\tau}\ .$$

Le modèle statistique induit par $(X_1^*,\dots,X_n^*)$ est donc donné par

$$\left([0,\tau]^n,\mathscr{B}([0,\tau]^n),\left\{(q_\theta^*)^{\otimes n}\cdot\mu^{\otimes n}\,,\ \theta\in\Theta\right\}\right)\ .$$

C'est un n-échantillon du modèle statistique

$$([0,\tau],\mathscr{B}([0,\tau]),\{q_\theta^*\cdot\mu\,,\ \theta\in\Theta\})\ .$$ ♢

Exemple I-1.17 (Regression logistique). Une nouvelle molécule anti-cancéreuse est en cours de développement [2]. Pour déterminer la dose à administrer aux futurs patients, un "bio-essai" est mené. Celui-ci consiste à injecter des doses croissantes $x_1<\dots<x_q$ de cette molécule à des populations de $n_1,n_2,\dots,n_q$ souris et à mesurer dans chaque population le nombre de souris y_i qui décèdent. On dispose donc de q-couples de données $\{(x_i,y_i)\}_{i=1}^q$. Rappelons tout d'abord que la loi binomiale $\mathrm{Bin}_{n,\pi}$, $\pi\in]0,1[$ est la loi sur $\{0,\dots,n\}$ de densité par rapport à la mesure de comptage donnée par

$$\mathrm{Bin}_{n,\pi}(k)=\binom{n}{k}\pi^k(1-\pi)^{n-k}\ .$$

L'application

$$\pi\mapsto\mathrm{logit}(\pi)=\log\left(\frac{\pi}{1-\pi}\right)$$

est appelée "application logit". C'est une fonction croissante, $\lim_{\pi\downarrow 0^+}\mathrm{logit}(\pi)=-\infty$ et $\lim_{\pi\uparrow 1^-}\mathrm{logit}(\pi)=\infty$. Les chercheurs postulent un lien linéaire entre la dose x et le taux de mortalité. Pour chaque population individuelle $i\in\{1,\dots,q\}$, nous postulons donc le modèle statistique binomial

$$\mathscr{E}_i:=\left(\{0,\dots,n_i\},\mathscr{P}(\{0,\dots,n_i\}),\left\{\mathrm{Bin}_{n_i,\pi(\theta,x_i)}\,,\ \theta=(\theta_0,\theta_1)\in\Theta=\mathbb{R}^2\right\}\right)$$

où

$$\pi(\theta,x):=\frac{\exp(\theta_0+\theta_1x)}{1+\exp(\theta_0+\theta_1x)}.$$

On vérifie aisément que

$$\mathrm{logit}(\pi(\theta,x))=\theta_0+\theta_1x\,.$$

Considérons le modèle statistique

$$\left(\prod_{i=1}^q\{0,\dots,n_i\},\mathscr{P}\left(\prod_{i=1}^q\{0,\dots,n_i\}\right),\left\{\bigotimes_{i=1}^q\mathrm{Bin}_{n_i,\pi(\theta,x_i)}\,,\ \theta\in\Theta\right\}\right)\ .$$

Pour $i\in\{1,\dots,n\}$ on note Y_i la i-ème observation qui est la statistique définie, pour tout $z=(y_1,\dots,y_q)\in\mathsf{Z}$ par $Y_i(z)=y_i$. Pour tout $\theta\in\Theta$, sous $\mathbb{P}_\theta$, les observations $(Y_1,\dots,Y_n)$ sont indépendantes et le modèle induit par chaque observation est le modèle binomial $\mathscr{E}_i$. Par abus de langage, nous dirons que "les observations $(Y_1,\dots,Y_n)$ sont indépendantes et pour $i\in\{1,\dots,n\}$, Y_i est distribué suivant le modèle binomial de paramètre $\pi(\theta,x_i)$". ♢

2. Cet exemple est emprunté à [**?**, Chapitre 8].

I-1.3 Pour aller plus loin

Le théorème de Radon-Nykodim (Théorème A.47) montre que le modèle $(\mathsf{Z}, \mathscr{Z}, \mathscr{C})$ est dominé si et seulement si toute loi de $\mathbb{P} \in \mathscr{C}$ est absolument continue par rapport à μ (voir Définition A.46).

La mesure de domination peut être choisie de multiples façons. Soient μ et ν sont deux mesures σ-finies distinctes sur $(\mathsf{Z}, \mathscr{Z})$. Supposons que le modèle statistique canonique paramétrique $(\mathsf{Z}, \mathscr{Z}, \mathscr{C}) = \{(\mathsf{Z}, \mathscr{Z}),\ \{\mathbb{P}_\theta,\ \theta \in \Theta\}\}$ est dominé par rapport à μ et à ν. Cette hypothèse implique que, pour tout $\theta \in \Theta$, nous pouvons associer deux fonctions mesurables positives, p_θ et q_θ telles que

$$\mathbb{P}_\theta = p_\theta \cdot \mu = q_\theta \cdot \nu\ .$$

La fonction p_θ est la densité de $\mathbb{P}_\theta$ par rapport à μ et q_θ est la densité de $\mathbb{P}_\theta$ par rapport à ν. Il est facile de montrer que ces deux densités diffèrent d'un facteur multiplicatif. Le raisonnement élémentaire est le suivant. Notons tout d'abord que les mesures σ-finies μ et ν sont absolument continues par rapport à $\lambda := \mu + \nu$. En effet, pour tout $B \in \mathscr{Z}$, $\lambda(B) = \mu(B) + \nu(B) = 0$ implique que $\mu(B) = 0$ et $\nu(B) = 0$. En appliquant le théorème A.47, il existe donc des fonctions mesurables positives m et n, uniques à une λ-équivalence près, telles que $\mu = m \cdot \lambda$ et $\nu = n \cdot \lambda$. Nous avons donc, pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta = p_\theta \cdot \mu = p_\theta m \cdot \lambda = q_\theta \cdot \nu = q_\theta n \cdot \lambda\ ,$$

et donc pour tout $\theta \in \Theta$, $p_\theta m = q_\theta n \quad \lambda$-p.p. Posons $\mathsf{Z}_0 = \{z \in \mathsf{Z},\ m(z) > 0\}$. Notons que

$$\mu(\mathsf{Z}_0^c) = \int_{\mathsf{Z}_0^c} m(z)\lambda(\mathrm{d}z) = 0\ ,$$

et donc, pour tout $\theta \in \Theta$, $\mathbb{P}_\theta(\mathsf{Z}_0^c) = 0$. Pour tout $z \in \mathsf{Z}_0$ et $\theta \in \Theta$, nous avons

$$p_\theta(z) = q_\theta(z)\frac{n(z)}{m(z)}\ , \quad \lambda - \text{p.p.}$$

Les deux exemples qui suivent sont plus avancés et peuvent être omis en première lecture.

Exemple I-1.18. Un exemple où il n'existe pas de mesure dominante est la famille paramétrique $\{\mathbb{P}_\theta = \delta_\theta, \theta \in \Theta = \mathbb{R}\}$, où δ_θ est la mesure de Dirac au point θ. Cet exemple correspond à l'"expérience parfaite" où une seule réalisation de la loi permet de déterminer sans erreur la valeur du paramètre. Montrons que ce modèle n'est pas dominé. Nous procédons par contradiction et supposons que ce modèle est dominé par une mesure σ-finie μ. Alors, nous avons $\mu(\{\theta\}) > 0$ pour tout $\theta \in \Theta$ (car $\mathbb{P}_\theta \ll \mu$ et $\mathbb{P}_\theta(\{\theta\}) > 0$), ce qui est impossible car une mesure σ-finie a au plus un nombre dénombrable d'atomes.
Nous allons prouver cette dernière propriété par contradiction. Comme μ est σ-finie, il existe une suite $\{F_k,\ k \in \mathbb{N}\}$ telle que $\mu(F_k) < \infty$ pour tout $k \in \mathbb{N}$, et $\mathbb{R} = \bigcup_{k=1}^\infty F_k$. On note $I = \{\theta \in \mathbb{R} :\ \mu(\{\theta\}) > 0\}$ et on rappelle que, par hypothèse, cet ensemble n'est pas dénombrable. Comme $I = \bigcup_{k=1}^\infty F_k \cap I$, il existe $n \in \mathbb{N}$ tel que $I \cap F_n$ est non dénombrable. Pour tout $p \in \mathbb{N}$, notons alors $I_p = \{\theta \in I \cap F_n :\ \mu(\{\theta\}) > 1/p\}$. Comme $I \cap F_n = \bigcup_{p=1}^\infty I_p$, il existe $q \in \mathbb{N}$ tel que I_q est non dénombrable. En particulier, I_q est donc infini. Pour tout sous-ensemble J de cardinal N de I_q, nous avons

$$\mu(F_n) \geq \mu(I_q) \geq \mu(J) = \sum_{\theta \in J} \mu(\{\theta\}) \geq \frac{N}{p}\ .$$

Comme p est fixé et que N peut être choisi arbitrairement grand, nous avons nécessairement $\mu(F_n) = \infty$, ce qui conduit à une contradiction. ♢

Exemple I-1.19. Un exemple plus subtil est donné par le modèle statistique canonique dont la famille de loi est donnée par

$$P_\theta = \sum_{k=1}^\infty \mathrm{e}^{-k}\delta_{\theta k}\ , \quad \text{où } \theta \in \Theta = \mathbb{R}_+ \setminus \{0\} \text{ est le paramètre.}$$

Dans ce cas, la loi de l'observation est non dégénérée (réduite à un point). Le modèle n'est pas dominé car là aussi, toute mesure de domination aurait alors nécessairement une infinité non dénombrable d'atomes. ♢

Chapitre I-2

Estimation ponctuelle

Ce chapitre est consacré à une introduction à l'estimation ponctuelle, centrée sur quelques méthodes élémentaires classiques ; ceci afin de comprendre les enjeux.
Les notions principales sont celles d'estimation ponctuelle ; d'estimation par la méthode des moments et par maximum de vraisemblance ; et de M et Z-estimateurs.

Définition I-2.1 (Estimateur ponctuel). *Soit* $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ *un modèle statistique paramétrique et* $g : \Theta \mapsto \mathbb{R}^q$ *une fonction. Un* estimateur ponctuel T *de* $g(\theta)$ *est une statistique à valeur dans* $\mathbb{R}^q$.

I-2.1 Méthode des moments

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'un modèle statistique paramétrique $(\mathsf{X}, \mathscr{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta\})$, où $\Theta \subseteq \mathbb{R}^d$. Pour tout $\theta \in \Theta$, nous notons $\mathbb{P}_\theta := \mathbb{Q}_\theta^{\otimes n}$. Le paramètre $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})$ est inconnu et le but est d'utiliser les observations pour approcher cette quantité.

Pour cela, la *méthode des moments* consiste à choisir d fonctions $T_j : \mathsf{X} \to \mathbb{R}$, $j \in \{1, \ldots, d\}$ telles que, pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[|T_j(X_1)|] < \infty$. On pose alors, pour tout $\theta \in \Theta$ et tout $j = 1, \ldots, d$,

$$e_j(\theta) := \mathbb{E}_\theta[T_j(X_1)]\ .$$

Les "moments" $\mathbb{E}_\theta[T_j(X_1)]$ ne sont pas connus, mais on peut les estimer par les moments empiriques $n^{-1}\sum_{i=1}^n T_j(X_i)$. En supposant que $\mathbb{E}_\theta[|T_j^2(X_1)] < \infty$, l'inégalité de Bienayme-Tchebychev (rappelée au lemme IV-1.2) montre que, pour tout $\varepsilon > 0$ et pour tout j,

$$\mathbb{P}_\theta\left(\left|n^{-1}\sum_{i=1}^n T_j(X_i) - \mathbb{E}_\theta[T_j(X_1)]\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}_\theta(T_j(X_1))}{n\varepsilon^2}\ .$$

En particulier, si $\{\varepsilon_n,\ n \in \mathbb{N}\}$ est une suite de nombres positifs tels que $\lim_{n\to\infty} n\varepsilon_n^2 = +\infty$, nous avons

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\left|n^{-1}\sum_{i=1}^n T_j(X_i) - \mathbb{E}_\theta[T_j(X_1)]\right| \geq \varepsilon_n\right) = 0\ .$$

Cette propriété garantit la *consistance* de l'estimateur $n^{-1}\sum_{i=1}^n T_j(X_i)$, notion sur laquelle nous reviendrons au Section II-1.4.2 : elle exprime que plus le nombre d'observations n est grand et meilleure est l'approximation de la quantité inconnue $\mathbb{E}_\theta[T_j(X_1)]$ par l'estimateur.

Pour estimer le paramètre inconnu $\theta \in \Theta$, on considère alors le système de d équations à d inconnues

$$\theta \in \Theta \quad \text{tel que pour tout } j \in \{1, \cdots, d\} \qquad n^{-1} \sum_{i=1}^{n} T_j(X_i) = e_j(\theta) \, .$$

En supposant que ce système admette une solution unique notée $\hat{\theta}_n$, on appelle $\hat{\theta}_n$ l'*estimateur des moments* de θ associé aux fonctions T_j, $j = 1 \dots d$. L'estimateur des moments est donc égal à la valeur du paramètre θ pour laquelle les moments exacts et les moments empiriques sont égaux.

Exemple I-2.2. Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle exponentiel

$$\left(\mathbb{R}^+, \mathscr{B}(\mathbb{R}^+), \left\{\text{Expo}(\theta) \, , \; \theta \in \Theta := \mathbb{R}_+^* \right\}\right) \, .$$

Rappelons que la loi exponentielle de paramètre $\theta > 0$ admet une densité $q_\theta(x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par

$$q_\theta(x) := \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x) \, . \tag{I-2.1}$$

Considérons deux stratégies d'estimation de θ, données par la méthode des moments associée aux deux fonctions $T(x) = x$ et $\tilde{T}(x) = x^2$. Déterminons tout d'abord les fonctions e et $\tilde{e}$ associées. Un calcul élémentaire montre que

$$\begin{aligned} e(\theta) &= \mathbb{E}_\theta \left[T(X_1)\right] = \int_0^{+\infty} x\theta \exp(-\theta x) \mathrm{d}x = \frac{1}{\theta} \\ \tilde{e}(\theta) &= \mathbb{E}_\theta \left[\tilde{T}(X_1)\right] = \int_0^{+\infty} x^2 \theta \exp(-\theta x) \mathrm{d}x = \frac{2}{\theta^2} \, . \end{aligned}$$

Les estimateurs des moments associés sont les solutions des équations

$$\frac{1}{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{et} \quad \frac{2}{\theta^2} = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \, .$$

Ces équations ont des solutions uniques (nous avons la contrainte $\theta > 0$), qui définissent les estimateurs des moments suivants :

$$\hat{\theta}_{n,1} := \frac{1}{n^{-1} \sum_{i=1}^{n} X_i} \, , \quad \text{et} \quad \hat{\theta}_{n,2} := \left(\frac{2}{n^{-1} \sum_{i=1}^{n} X_i^2} \right)^{1/2} \, . \tag{I-2.2}$$

♢

Exemple I-2.3. Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle de Cauchy

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\text{Cauchy}(\theta) \, , \; \theta \in \Theta := \mathbb{R}\}\right) \, .$$

Rappelons que la loi de Cauchy de paramètre de translation θ et d'échelle 1 admet une densité par rapport à la mesure de Lebesgue donnée par

$$q_\theta(x) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)} , \; x \in \mathbb{R}.$$

La densité $q_\theta(\cdot)$ n'a pas de moment d'ordre k pour $k \geq 1$, et le choix $T(x) = x^k$ avec k entier ne s'applique pas ici. Prenons $T(x) = \text{signe}(x)$ où la fonction signe est définie par

$$\text{signe}(x) := \begin{cases} -1 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0. \end{cases} \tag{I-2.3}$$

On a

$$\mathbb{E}_\theta \left[T(X_1)\right] = \int \text{signe}(x) q_\theta(x) \mathrm{d}x = 1 - 2F(-\theta) \, ,$$

où F est la fonction de répartition de la loi de Cauchy de paramètre de translation 0 et d'échelle 1 :

$$F(t) = \frac{1}{\pi} \int_{-\infty}^{t} \frac{\mathrm{d}x}{1+x^2} = \frac{1}{\pi} \arctan(t) + \frac{1}{2}.$$

Par suite, $e(\theta) = 2\arctan(\theta)/\pi$ et on obtient l'estimateur des moments en résolvant

$$\frac{2}{\pi} \arctan(\theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i \, , \quad \text{où } Y_i := \text{signe}(X_i);$$

d'où l'estimateur

$$\widehat{\theta}_n := \tan\left(\frac{\pi}{2n} \sum_{i=1}^{n} Y_i \right).$$

♢

Exemple I-2.4. Soit q une densité sur $\mathbb{R}$ vérifiant

$$\int xq(x)\mathrm{d}x = 0\,, \qquad m_2 := \int x^2 q(x)\mathrm{d}x > 0\,, \qquad m_4 := \int x^4 q(x)\mathrm{d}x < \infty\,.$$

Par exemple, on peut prendre pour q la densité d'une gaussienne centrée réduite, $q(x) = (2\pi)^{-1/2}\exp(-x^2/2)$: dans ce cas, $m_2 = 1$ et $m_4 = 3$; on peut aussi considérer la densité de la loi de Laplace $q(x) = (1/2)\exp(-|x|)$ auquel cas $m_2 = 2$ et $m_4 = 4! = 24$.

Considérons un n-échantillon $(X_1,\ldots,X_n)$ du modèle $(\mathsf{X},\mathscr{X},\{q_\theta\cdot\mu\,,\ \theta\in\Theta\})$ où

$$q_\theta(x) := \frac{1}{\sigma}q\left(\frac{x-m}{\sigma}\right)\,, \quad \theta := (m,\sigma^2)\in\Theta := \mathbb{R}\times\mathbb{R}_+^*\,, \tag{I-2.4}$$

Il est facile de vérifier que pour tout $\theta\in\Theta$

$$\mathbb{E}_\theta[X_1] = m\,, \qquad \mathrm{Var}_\theta(X_1) = \sigma^2 m_2.$$

On peut donc choisir $T_1(x) = x$ et $T_2(x) = x^2$ et identifier l'estimateur des moments à la solution du système d'équations

$$m = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i\,, \tag{I-2.5}$$

$$\sigma^2 m_2 + m^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i^2\,. \tag{I-2.6}$$

Pour tout $\theta\in\Theta$, la matrice jacobienne de cette transformation est donnée par

$$\mathrm{J}_{\mathbf{e}}(\mu,\sigma^2) = \begin{bmatrix} 1 & 0 \\ 2m & m_2 \end{bmatrix}\,;$$

elle est donc inversible et son inverse est donnée par

$$\left\{\mathrm{J}_{\mathbf{e}}(\mu,\sigma^2)\right\}^{-1} = \frac{1}{m_2}\begin{bmatrix} m_2 & 0 \\ -2m & 1 \end{bmatrix}\,.$$

Cette transformation est aussi injective, car

$$\begin{bmatrix} m \\ m^2+m_2\sigma_1^2 \end{bmatrix} = \begin{bmatrix} \bar{m} \\ \bar{m}^2+m_2\sigma_2^2 \end{bmatrix} \quad \Leftrightarrow \begin{bmatrix} m \\ \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \bar{m} \\ \sigma_2^2 \end{bmatrix}\,.$$

Dans ce cas particulier d'ailleurs, l'inverse a une expression explicite et (I-2.5) admet donc une unique solution $\hat{\theta}_n := (\hat{m}_n,\hat{\sigma}_n^2)$ donnée par

$$\begin{cases} \hat{m}_n &= \frac{1}{n}\sum_{i=1}^n X_i\,, \\ \hat{\sigma}_n^2 &= \frac{1}{nm_2}\sum_{i=1}^n (X_i-\hat{m}_n)^2\,. \end{cases} \tag{I-2.7}$$

♢

I-2.2 Z-estimateurs

Soit $(X_1,\ldots,X_n)$ un n-échantillon d'un modèle statistique paramétrique $(\mathsf{X},\mathscr{X},\{\mathbb{Q}_\theta\,,\ \theta\in\Theta\})$. Pour tout $\theta\in\Theta$, nous notons $\mathbb{P}_\theta := \mathbb{Q}_\theta^{\otimes n}$.

L'estimateur des moments basé sur les statistiques $\mathbf{T} = (T_1,\ldots,T_d)$ consiste à identifier les moments empiriques et les moments exacts. Il est donc la solution d'un système de d équations à d inconnues

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(\theta,X_i) = 0\,, \tag{I-2.8}$$

où nous avons défini la fonction

$$\boldsymbol{\psi}(\theta,x) = [\psi_1(\theta,x),\ldots,\psi_d(\theta,x)]^T := \mathbf{T}(x) - \mathbb{E}_\theta[\mathbf{T}(X_1)]\,. \tag{I-2.9}$$

Nous verrons dans la suite qu'il est intéressant de considérer des estimateurs définis comme solution du système d'équations (I-2.8), mais pour des fonctions plus générales que (I-2.9). Supposons que nous disposions de fonctions

$$\psi_j\ : (\theta,x)\mapsto \psi_i(\theta,x), \qquad j = 1,\ldots,d\,,$$

vérifiant, pour tout $\theta_0 \in \Theta$ et $j \in \{1,\dots,d\}$,

$$\mathbb{E}_{\theta_0}[|\psi_j(\theta_0, X_1)|] < \infty\,. \tag{I-2.10}$$

Notons alors $\boldsymbol{\psi}(\theta,x) = [\psi_1(\theta,x),\dots,\psi_d(\theta,x)]^T$ et supposons que, pour tout $\theta_0 \in \Theta$,

$$\mathbb{E}_{\theta_0}[\boldsymbol{\psi}(\theta_0, X_1)] = \mathbf{0}_{d\times 1}\,. \tag{I-2.11}$$

Cette condition est vérifiée par les estimateurs de moments (I-2.9).

Définition I-2.5 (*Z*-estimateur, cas multidimensionnel). *On appelle* Z-estimateur *associé à* $\boldsymbol{\psi}$ *tout estimateur* $\hat\theta_n$ *satisfaisant* $\Psi_n(\hat\theta_n) = 0$*, où*

$$\Psi_n(\theta) := \frac{1}{n}\sum_{i=1}^n \psi(\theta, X_i)\,. \tag{I-2.12}$$

Exemple I-2.6 (Estimation du paramètre de translation). Soit F une fonction de répartition d'une loi symétrique sur $\mathbb{R}$, i.e. $F(x) = 1 - F(-x)$ pour tout $x \in \mathbb{R}$.

Soit $(X_1,\dots,X_n)$ un n-échantillon du modèle $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathbb{Q}_\theta\,,\ \theta \in \Theta = \mathbb{R}\})$ où, pour tout $\theta \in \Theta$ et $x \in \mathbb{R}$,

$$\mathbb{Q}_\theta(]-\infty, x]) = F_\theta(x) := F(x-\theta)\,.$$

Ici, θ est un paramètre de translation (ou de position). Supposons tout d'abord que F admette un moment d'ordre 1 : $\int |x| F(\mathrm{d}x) < \infty$. Comme la loi F est symétrique, $\int x F(x)\mathrm{d}x = 0$. On a donc

$$\mathbb{E}_\theta[X] = \int x F_\theta(\mathrm{d}x) = \int (x+\theta) F(\mathrm{d}x) = \theta,$$

ce qui suggère de construire l'estimateur des moments $\hat\theta_n$ qui est solution de l'équation

$$\frac{1}{n}\sum_{i=1}^n \psi(\theta, X_i) = 0\,, \quad \text{en ayant posé} \quad \psi(\theta,x) := x - \theta\,.$$

La solution de ce problème est ici explicite et élémentaire : il s'agit de la moyenne empirique

$$\hat\theta_n = n^{-1}\sum_{i=1}^n X_i.$$

On peut bien entendu considérer d'autres fonctions ψ. Supposons par exemple que la fonction de répartition F soit continue en 0. Comme F est symétrique, nous avons $\int \psi(x) F(\mathrm{d}x) = 0$ où $\psi(x) = \mathrm{signe}(x)$. Un Z-estimateur du paramètre de translation associé à cette fonction ψ est une solution de l'équation

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

où la fonction signe$(\cdot)$ est définie par (I-2.3). La solution est donnée par la *médiane empirique* de l'échantillon, définie par

$$\hat\theta_n = \begin{cases} X_{m:n} & \text{si } n = 2m-1 \\ (1/2)(X_{m:n} + X_{m+1:n}) & \text{si } m = 2n \end{cases} \tag{I-2.13}$$

où $X_{m:n}$ est la m-ème statistique d'ordre (voir Définition IV-2.14). Ainsi, la médiane empirique est un Z-estimateur ; ce n'est par contre pas un estimateur de moment. ♢

I-2.3 Maximum de vraisemblance

L'estimateur du maximum de vraisemblance joue un rôle important en statistique inférentielle. Avant de procéder à une construction générale, considérons tout d'abord un exemple élémentaire.

Exemple I-2.7 (Sondage). Soit $(X_1,\dots,X_n)$ un n-échantillon d'un modèle de Bernoulli

$$(\{0,1\},\mathscr{P}(\{0,1\}),\{\mathrm{Ber}(\theta)\,,\ \theta\in\Theta:=[0,1]\})\,.$$

Pour tout $\theta\in\Theta$, nous notons $\mathbb{P}_\theta:=\mathrm{Ber}^{\otimes n}(\theta)$. Ce modèle est dominé par la mesure de comptage sur $\{0,1\}^n$ et sa densité est donnée, pour tout $(x_1,\dots,x_n)\in\{0,1\}^n$ par

$$\begin{aligned}p_\theta(x_1,\dots,x_n)&=\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\\&=\prod_{i=1}^n q(\theta,x_i)\quad\text{en posant } q(\theta,x_i):=\theta^{x_i}(1-\theta)^{1-x_i}\,.\end{aligned}$$

Pour une réalisation donnée $(X_1,\dots,X_n)$, telle que $\bar X_n=0$ où on a posé

$$\bar X_n:=\frac{1}{n}\sum_{i=1}^n X_i\,,$$

le maximum de la fonction de vraisemblance

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

est atteint en $\hat\theta_n=0$. Lorsque $\bar X_n=1$, le maximum de la fonction de vraisemblance est atteint en $\hat\theta_n=1$. Dans toute la suite, considérons le cas où $\bar X_n\in\,]0,1[$: pour maximiser la fonction de vraisemblance, nous allons considérer son logarithme et se restreindre au cas où $\theta\in\,]0,1[$ (on pourra vérifier a posteriori que dans le cas $\bar X_n\notin\{0,1\}$, la fonction de vraisemblance n'est pas maximale sur les bords de Θ). Soit la fonction définie sur $]0,1[$ par

$$\begin{aligned}\theta\mapsto\ell_n(\theta):=n^{-1}\sum_{i=1}^n\{X_i\log(\theta)+(1-X_i)\log(1-\theta)\}=\bar X_n\log(\theta)+(1-\bar X_n)\log(1-\theta)\\=n^{-1}\sum_{i=1}^n\log q(\theta,X_i)\,.\end{aligned}$$

Cette fonction est appelée *log-vraisemblance* (normalisée). Fixons $\theta_0\in\,]0,1[$. Nous avons $\mathbb{E}_{\theta_0}[X_i]=\theta_0$ pour tout $i\in\{1,\dots,n\}$ et donc, pour tout $\theta\in\,]0,1[$,

$$\mathbb{E}_{\theta_0}[\log q(\theta,X_1)]=\theta_0\log(\theta)+(1-\theta_0)\log(1-\theta)\,.$$

La dérivée seconde de la fonction $\theta\mapsto\mathbb{E}_{\theta_0}[\log q(\theta,X_1)]$ est donnée, pour tout $\theta\in\,]0,1[$, par

$$-\frac{\theta_0}{\theta^2}-\frac{1-\theta_0}{(1-\theta)^2}<0\,.$$

Cette fonction est de dérivée nulle en θ_0 et strictement concave sur $]0,1[$, elle admet donc un maximum unique au point θ_0. En utilisant l'inégalité de Bienayme-Tchebychev (Lemme IV-1.2), pour tout $\theta\in\,]0,1[$ et $\delta>0$,

$$\mathbb{P}_{\theta_0}\left(\left|n^{-1}\sum_{i=1}^n\log q(\theta,X_i)-\mathbb{E}_{\theta_0}[\log q(\theta,X_i)]\right|\geq\delta\right)\leq\frac{\mathrm{Var}_{\theta_0}\left(\log q(\theta,X_1)\right)}{n\delta^2}\quad.$$

Comme de plus

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[\left\{\log q(\theta,X_i)-\mathbb{E}_{\theta_0}[\log q(\theta,X_i)]\right\}^2\right]&=\mathbb{E}_{\theta_0}[(X_i-\theta_0)^2]\left\{\log\left(\frac{\theta}{1-\theta}\right)\right\}^2\\&=\theta_0(1-\theta_0)\left\{\log\left(\frac{\theta}{1-\theta}\right)\right\}^2,\end{aligned}$$

pour tout $\theta\in\,]0,1[$ il vient

$$\lim_{n\to\infty}\mathbb{P}_{\theta_0}\left(\left|\ell_n(\theta)-\mathbb{E}_{\theta_0}[\log q(\theta,X_i)]\right|\geq\delta\right)=\lim_{n\to\infty}\mathbb{P}_{\theta_0}\left(\left|n^{-1}\sum_{i=1}^n\log q(\theta,X_i)-\mathbb{E}_{\theta_0}[\log q(\theta,X_i)]\right|\geq\delta\right)=0\,.$$

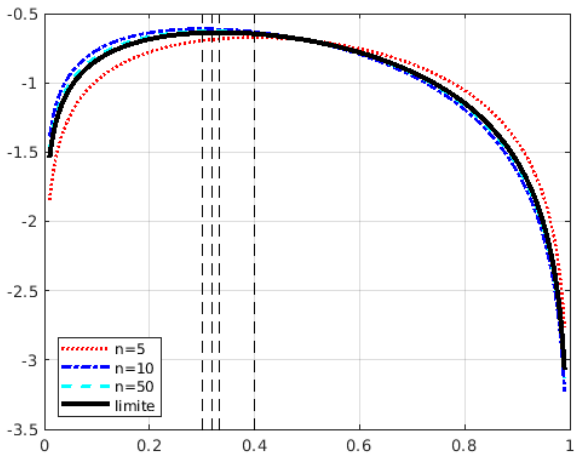

FIGURE I-2.1 – On dispose de 50 mesures binaires $(x_1, \cdots, x_{50})$ telles que $\bar{x}_n := n^{-1}\sum_{i=1}^n x_i$ vaut 0.4 lorsque $n = 5$, 0.3 lorsque $n = 10$ et 0.32 lorsque $n = 50$. Ces mesures ont été obtenues comme la réalisation de v.a. de Bernoulli indépendantes de paramètres $\theta_0 = 1/3$. On trace les fonctions $\theta \mapsto \bar{x}_n \log(\theta) + (1-\bar{x}_n)\log(1-\theta)$ sur $]0,1[$ dans le cas $n = 5$, $n = 10$ et $n = 50$. On trace aussi la fonction $\theta \mapsto \theta_0 \log(\theta) + (1-\theta_0)\log(1-\theta)$ sur $]0,1[$ en noir. Noter que lorsque $n \to \infty$, sous $\mathbb{P}_{\theta_0}$, $\bar{X}_n$ converge presque-sûrement vers θ_0. Les traits verticaux sont placé en θ_0, 0.3, 0.32 et 0.4.

Par conséquent, la fonction $\theta \mapsto \ell_n(\theta)$ approche la fonction $\theta \mapsto \mathbb{E}_{\theta_0}[\ell_n(\theta)]$. Nous verrons plus tard que cette approximation est en fait uniforme si θ appartient à un sous-ensemble $[a,b]$ avec $0 < a < b < 1$. Ceci est illustré dans la fig. I-2.1 La fonction $\theta \mapsto \ell_n(\theta)$ est strictement concave sur $]0,1[$ et $\lim_{\theta\to 0} \ell_n(\theta) = -\infty$ et $\lim_{\theta\to 1} \ell_n(\theta) = -\infty$. Par conséquent la fonction $\theta \mapsto \ell_n(\theta)$ admet un maximum unique qui est caractérisé par l'équation

$$0 = \ell_n'(\theta) = n^{-1}\sum_{i=1}^n \left\{ \frac{X_i}{\theta} - \frac{1-X_i}{1-\theta} \right\} = \frac{\bar{X}_n}{\theta} - \frac{1-\bar{X}_n}{1-\theta} .$$

Cette équation a une unique solution dans $[0,1]$, donnée par

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

Ce maximum unique est appelé *estimateur du maximum de vraisemblance*. C'est la valeur du paramètre qui maximise la vraisemblance des observations.

Dans cet exemple, $\hat{\theta}_n$ est aussi un estimateur des moments associé à la statistique $T(x) = x$; nous avons $\mathbb{E}_\theta[T(X_1)] = \theta$ pour tout $\theta \in \Theta$. Cette propriété sera vérifiée dans les modèles statistique de type exponentiel (voir Chapitre IV-4).♢

Nous allons maintenant donner une définition plus formelle de la vraisemblance.

Définition I-2.8 (Vraisemblance). *Soit $(\mathsf{X}, \mathscr{X})$ un espace mesurable et $\mu \in \mathbb{M}_+(\mathscr{X})$ une mesure σ-finie.*

On appelle fonction de vraisemblance (ou vraisemblance) associée au n-échantillon $(X_1, \ldots, X_n)$ du modèle statistique $(\mathsf{X}, \mathscr{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$ l'application

$$\theta \in \Theta \mapsto \mathrm{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^n q_\theta(X_i) .$$

Quand il n'y a pas de risque de confusion, nous utiliserons la notation abrégée

$$\mathrm{L}_n(\theta) = \mathrm{L}_n(\theta, X_1, \ldots, X_n) .$$

Exemple I-2.9 (Loi de Poisson). Dans ce cas $\mathsf{X} = \mathbb{N}$, que nous équipons de la mesure de comptage μ sur $\mathbb{N}$. Pour $\theta \in \Theta := \mathbb{R}_+^*$, nous appelons loi de Poisson de paramètre θ, la loi de densité par rapport à μ,

$$q_\theta(x) = \mathrm{e}^{-\theta}\,\frac{\theta^x}{x!}\,, \quad x \in \mathbb{N}, \theta \in \Theta\,.$$

Soit $(X_1,\ldots,X_n)$ un n-échantillon du modèle $(\mathsf{X}, \mathscr{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$. La vraisemblance s'écrit, pour tout $\theta > 0$,

$$\mathrm{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^n \mathrm{e}^{-\theta}\,\frac{\theta^{X_i}}{X_i!} = \frac{1}{\prod_{i=1}^n X_i!}\exp\left(-n\theta + n\log(\theta)\bar{X}_n\right)\,.$$

La vraisemblance admet un maximum unique en

$$n^{-1}\sum_{i=1}^n X_i\,.$$

♢

Exemple I-2.10 (Censure). Nous reprenons l'exemple I-1.3. Les observations sont des survies censurées à droite

$$X_i := \min\{Y_i, \tau\},\ \ i = 1,\ldots,n$$

où $\tau > 0$ est un instant de censure et $(Y_1,\ldots,Y_n)$ est un n-échantillon du modèle exponentiel

$$(\mathbb{R}_+, \mathscr{B}(\mathbb{R}_+), \left\{\mathrm{Expo}(\theta)\,,\ \theta \in \Theta := \mathbb{R}_+^*\right\})\,.$$

Attention : les mesures dont on dispose sont modélisées par le modèle statistique induit par les v.a. $(X_1,\ldots,X_n)$ et non pas par les v.a. $(Y_1,\cdots,Y_n)$. Posons $\mu := \lambda_{\mathrm{Leb}} + \delta_\tau$, où λ_{Leb} est la mesure de Lebesgue sur $\mathbb{R}$ et δ_τ est la mesure de Dirac au point τ. Pour $\theta \in \mathbb{R}_+^*$, considérons la densité par rapport à la mesure μ donnée par

$$q_\theta(x) := \theta\mathrm{e}^{-\theta x}\mathbb{1}_{\{x<\tau\}} + c(\theta)\mathbb{1}_{\{x=\tau\}}, \qquad c(\theta) := \int_\tau^{+\infty} \theta\mathrm{e}^{-\theta t}\mathrm{d}t = \mathrm{e}^{-\theta\tau}\,.$$

Nous avons établi (voir exemple I-1.3) que $(X_1,\ldots,X_n)$ est un n-échantillon du modèle

$$\left([0,\tau], \mathscr{B}([0,\tau]), \left\{q_\theta\,,\ \theta \in \Theta = \mathbb{R}_+^*\right\}\right)\,.$$

La vraisemblance s'écrit

$$\theta \mapsto \mathrm{L}_n(\theta, X_1, \ldots, X_n) = \theta^{\mathrm{card}\,N_n^-}\exp\left(-\theta\sum_{i\in N_n^-} X_i\right) c(\theta)^{\mathrm{card}\,N_n^+}\,,$$

où $N_n^- := \{i \le n :\ X_i < \tau\}$ et $N_n^+ := \{i \le n :\ X_i = \tau\}$. Elle est à comparer avec la vraisemblance du modèle sans censure, où l'on observe les Y_i directement. Dans ce cas, la vraisemblance par rapport à la mesure de Lebesgue s'écrit

$$\theta \mapsto \mathrm{L}_n(\theta, Y_1, \ldots, Y_n) = \theta^n\exp\left(-\theta\sum_{i=1}^n Y_i\right).$$

♢

Définition I-2.11 (Maximum de vraisemblance). *On appelle estimateur du maximum de vraisemblance tout estimateur $\hat{\theta}_n^{\mathrm{MV}}$ satisfaisant*

$$\mathrm{L}_n(\hat{\theta}_n^{\mathrm{MV}}, X_1, \ldots, X_n) = \max_{\theta\in\Theta} \mathrm{L}_n(\theta, X_1, \ldots, X_n)\,.$$

Autrement dit

$$\hat{\theta}_n^{\mathrm{MV}} \in \arg\max_{\theta\in\Theta} \mathrm{L}_n(\theta, X_1, \ldots, X_n)\,. \tag{I-2.14}$$

L'application

$$\theta \in \Theta \to \ell_n(\theta, X_1, \ldots, X_n) := \frac{1}{n} \log \mathrm{L}_n(\theta, X_1, \ldots, X_n) \,, \tag{I-2.15}$$

bien définie si $q(\theta, \cdot) > 0$, est appelée *fonction de log-vraisemblance*. En posant $\log 0 = -\infty$, on pourra parler de log-vraisemblance en toute généralité. On a aussi

$$\hat{\theta}_n^{\mathrm{MV}} \in \arg\max_{\theta \in \Theta} \ell_n(\theta, X_1, \ldots, X_n).$$

Si le maximum de $\theta \to \mathrm{L}_n(\theta)$, ou encore le maximum de $\theta \to \ell_n(\theta)$, est atteint dans l'intérieur de Θ, et si l'application $\theta \to \ell_n(\theta)$ est continûment différentiable, alors, l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$ vérifie

$$\nabla_\theta \, \ell_n(\theta, X_1, \ldots, X_n)|_{\theta = \hat{\theta}_n^{\mathrm{MV}}} = 0. \tag{I-2.16}$$

Ceci fournit un système de d équations si $\Theta \subseteq \mathbb{R}^d$.

Définition I-2.12 (Equations de vraisemblance). *L'équation* (I-2.16) *est appelée équation de vraisemblance si $d = 1$ et système d'équations de vraisemblance si $d > 1$.*

En résolvant (I-2.16), on obtient tous les points critiques de $\theta \to \ell_n(\theta)$, en particulier, tous ses maxima et minima locaux. On appelle racine de l'équation de vraisemblance tout (estimateur) $\widehat{\theta}_n^{\mathrm{rv}}$ solution de (I-2.16), c'est-à-dire tel que

$$\nabla_\theta \, \ell_n(\widehat{\theta}_n^{\mathrm{rv}}, X_1, \ldots, X_n) = 0 \,.$$

Dans certains cas, le problème d'optimisation a une solution unique, mais cette solution n'a pas de caractérisation variationnelle. C'est typiquement le cas lorsque la fonction de vraisemblance n'est pas dérivable. On ne peut pas dans un tel cas obtenir l'estimateur du maximum de vraisemblance comme solution d'un système d'équations de vraisemblance (voir Exemple I-2.15).

L'estimateur du maximum de vraisemblance n'est pas toujours défini (voir Exemple I-2.16).

Enfin, il n'y a pas nécessairement unicité de l'estimateur du maximum de vraisemblance (voir Exemple I-2.17).

Exemple I-2.13 (modèle gaussien standard). Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle statistique gaussien

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{ \mathrm{N}(\mu, \sigma^2) \,,\, (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^* \right\}) \,.$$

La densité de la loi $\mathrm{N}(\mu, \sigma^2)$ est donnée par

$$q_\theta(x) := (2\pi\sigma^2)^{-1/2} \exp\left(-\tfrac{1}{2\sigma^2}(x-\mu)^2 \right) \,, \quad \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^* \,.$$

La log-vraisemblance associée s'écrit

$$\theta \mapsto \ell_n\left((\mu, \sigma^2), X_1, \ldots, X_n\right) := -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \,.$$

Les équations de vraisemblance s'écrivent

$$\begin{cases} \frac{\partial \ell_n}{\partial \mu}\left((\mu, \sigma^2), X_1, \ldots, X_n\right) &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \,, \\ \frac{\partial \ell_n}{\partial \sigma^2}\left((\mu, \sigma^2), X_1, \ldots, X_n\right) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2 \,. \end{cases}$$

Pour $n \geq 2$, ceci nous fournit le point critique

$$\hat{\theta}_n = \left(\overline{X}_n, \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2 \right).$$

On vérifie que la log-vraisemblance est strictement concave et que le point critique est l'unique maximum global. Par conséquent, nous avons $\hat{\theta}_n^{\mathrm{MV}} = \hat{\theta}_n$. ♢

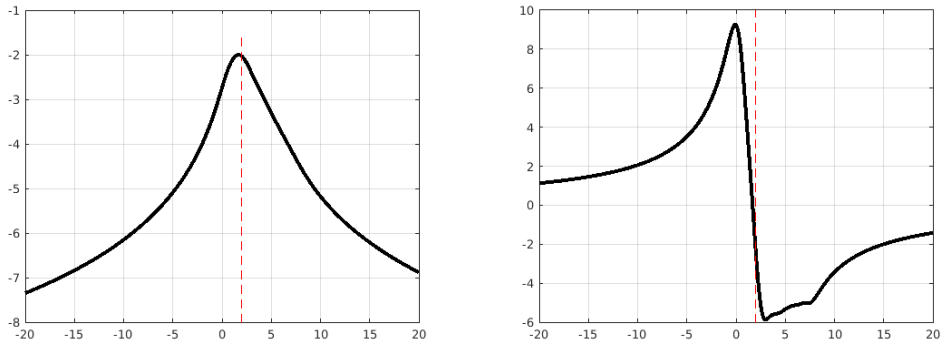

FIGURE I-2.2 – On dispose de $n = 25$ mesures réelles $x_1, \cdots, x_n$ (obtenues comme la réalisation de n v.a. i.i.d. de loi de Cauchy de paramètre $\theta_0 = 2$). On affiche une réalisation de la fonction de log-vraisemblance normalisée : $\theta \mapsto -\log \pi - \frac{1}{n}\sum_{i=1}^{n} \log\left(1+(x_i-\theta)^2\right)$ et une réalisation de sa dérivée (graphe de droite). On observera que le maximum de vraisemblance n'est pas θ_0 (le trait vertical en pointillé est la droite d'équation $x = \theta_0$).

Exemple I-2.14 (Loi de Cauchy). Soit $(X_1, \ldots, X_n)$ un n-échantillon d'un modèle de Cauchy

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\text{Cauchy}(\theta)\,,\ \theta \in \Theta := \mathbb{R}\})\ ,$$

où $\text{Cauchy}(\theta)$ est une loi de Cauchy de paramètre de translation θ et de paramètre d'échelle 1, de densité par rapport à la mesure de Lebesgue (voir exemple I-2.3)

$$q_\theta(x) := \frac{1}{\pi\left(1+(x-\theta)^2\right)}\ .$$

La vraisemblance s'écrit alors,

$$\theta \mapsto \mathrm{L}_n(\theta, X_1, \ldots, X_n) = \frac{1}{\pi^n}\prod_{i=1}^{n} \frac{1}{1+(X_i-\theta)^2}\ .$$

La log-vraisemblance est donnée par

$$\theta \mapsto \ell_n(\theta, X_1, \ldots, X_n) = -\log \pi - \frac{1}{n}\sum_{i=1}^{n} \log\left(1+(X_i-\theta)^2\right),$$

dont l'équation de vraisemblance associée est :

$$\sum_{i=1}^{n} \frac{X_i-\theta}{1+(X_i-\theta)^2} = 0. \tag{I-2.17}$$

Cette équation n'admet pas de solution explicite et peut d'ailleurs admettre plusieurs solutions. Pour maximiser la fonction de vraisemblance, il faut avoir recours à une procédure numérique. On peut par exemple utiliser dans cet exemple une méthode de gradient. Notons $\theta^{(k)}$ la valeur du paramètre à la k-ième itération de l'algorithme de gradient. La valeur à la $k+1$-ème itération est donnée par

$$\theta^{(k+1)} = \theta^{(k)} + \gamma \sum_{i=1}^{n} \frac{X_i-\theta^{(k)}}{1+(X_i-\theta^{(k)})^2}\ ,$$

où γ est un pas d'apprentissage. Nous avons représenté Figure I-2.2 la log-vraisemblance et sa dérivée pour $n = 50$ observations ; ces mesures observées ont été produites comme la réalisation de v.a. de Cauchy de paramètre $\theta_0 = 2$. ♢

Exemple I-2.15 (modèle uniforme). Soit $(X_1, \ldots, X_n)$ un n-échantillon d'un modèle statistique uniforme

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{\text{Unif}([0,\theta])\,,\ \theta \in \Theta := \mathbb{R}_+^*\right\})\ ,$$

où pour $\theta \in \Theta$, Unif$([0,\theta])$ est la loi uniforme de paramètre θ de densité par rapport à la mesure de Lebesgue donnée par

$$q_\theta(x) := \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x) \, .$$

La fonction de vraisemblance s'écrit

$$\theta \mapsto \mathrm{L}_n(\theta, X_1, \ldots, X_n) = \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{[0,\theta]}(X_i) = \theta^{-n} \mathbb{1}_{[0,\theta]}(X_{n:n}),$$

où $X_{n:n} = \max_{i=1,\ldots,n} X_i$ (voir Définition IV-2.14). La valeur maximale de $\mathrm{L}_n(\theta, X_1, \ldots, X_n)$ est obtenue pour $\theta = X_{n:n}$ et donc $\hat{\theta}_n^{\mathrm{MV}} = X_{n:n}$. Par contre, la vraisemblance n'est pas dérivable et la fonction de log-vraisemblance n'est pas définie pour toutes les valeurs de $\theta \in \Theta$. ♢

Exemple I-2.16 (absence d'estimateur du maximum de vraisemblance). Soit q_0 la densité par rapport à la mesure de Lebesgue définie par

$$q_0(x) = \frac{\mathrm{e}^{-\frac{|x|}{2}}}{2\sqrt{2\pi|x|}}, \ \ x \in \mathbb{R}.$$

Considérons le n-échantillon du modèle statistique $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{q_0(\cdot - \theta) \, , \ \theta \in \Theta := \mathbb{R}\})$. La fonction de vraisemblance s'écrit

$$\theta \mapsto \mathrm{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^n q_0(X_i - \theta).$$

On a $\lim_{\theta \to X_i} \mathrm{L}_n(\theta, X_1, \ldots, X_n) = +\infty$ pour tout $i = 1, \ldots, n$. Pour cette expérience statistique, l'estimateur du maximum de vraisemblance n'est pas défini. ♢

Exemple I-2.17 (modèle de Laplace). Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle statistique

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{q_\theta \cdot \lambda_{\mathrm{Leb}} \, , \ \theta \in \Theta := \mathbb{R}\})$$

où pour $\theta \in \Theta$, la densité q_θ est donnée par

$$q_\theta(x) := \frac{1}{2} \exp\big(-|x - \theta|\big) \, , x \in \mathbb{R} \, .$$

La log-vraisemblance est donnée par

$$\theta \mapsto \ell_n(\theta, X_1, \ldots, X_n) = -n \log(2) - \sum_{i=1}^n \big|X_i - \theta\big|.$$

Maximiser $\ell_n(\theta, X_1, \ldots, X_n)$ revient à minimiser la fonction $\theta \to \sum_{i=1}^n \big|X_i - \theta\big|$. Cette fonction est dérivable presque partout, de dérivée $-\sum_{i=1}^n \mathrm{sign}(X_i - \theta)$. La dérivée (définie presque partout) est constante par morceaux. Si n est impair, elle s'annule en un point unique $X_{(n+1)/2:2}$ (sous réserve que les X_i soient deux-à-deux distincts, ce qui est vrai presque sûrement car la loi de $(X_1, \ldots, X_n)$ est à densité par rapport à la mesure de Lebesgue), où $X_{1:n} \leq \ldots \leq X_{n:n}$ désigne la statistique d'ordre associée à l'échantillon (voir Définition IV-2.14). Si n est pair, il y a une infinité de solutions : tout point de l'intervalle $\big(X_{n/2:n}, X_{n/2+1:n}\big)$ est un estimateur du maximum de vraisemblance. ♢

I-2.3.1 Invariance par reparamétrisation

L'estimateur du maximum de vraisemblance est invariant par changement de paramétrisation. Cela signifie que, si $\hat{\theta}_n^{\mathrm{MV}}$ est l'estimateur du maximum de vraisemblance pour θ, alors $G(\hat{\theta}_n^{\mathrm{MV}})$ est l'estimateur du maximum de vraisemblance du paramètre $G(\theta)$ pour toute fonction G bijective. Pour le voir, considérons une famille de probabilités associée à une expérience statistique $\big\{\mathbb{P}_\theta, \, \theta \in \Theta\big\}$ et une bijection

$$G : \Theta \to G(\Theta)$$

de Θ sur son image $G(\Theta)$. On construit une nouvelle famille de probabilités $\big\{\mathbb{Q}_\tau, \tau \in G(\Theta)\big\}$ en posant

$$\mathbb{Q}_\tau := \mathbb{P}_{G^{-1}(\tau)} \, .$$

Proposition I-2.18. *Si $G : \Theta \to G(\Theta)$ est une bijection et si $\hat{\theta}_n^{\mathrm{MV}}$ désigne l'estimateur du maximum de vraisemblance pour l'expérience statistique associée à la famille de lois $\{\mathbb{P}_\theta, \theta \in \Theta\}$, alors $G(\hat{\theta}_n^{\mathrm{MV}})$ est l'estimateur du maximum de vraisemblance de $G(\theta)$, c'est-à-dire pour l'expérience statistique associée à la famille de lois $\{\mathbb{Q}_\tau, \tau \in G(\Theta)\}$.*

Démonstration. Posons $\widehat{\tau}_n := G(\hat{\theta}_n^{\mathrm{MV}})$. Alors $\hat{\theta}_n^{\mathrm{MV}} = G^{-1}(\widehat{\tau}_n)$. Pour tout $\tau \in G(\Theta)$, la vraisemblance $\widetilde{\mathrm{L}_n}(\tau, X_1, \ldots, X_n)$ associée à la famille $\{\mathbb{P}_{G^{-1}(\tau)}, \tau \in G(\Theta)\}$ s'écrit

$$\widetilde{\mathrm{L}_n}(\tau, X_1, \ldots, X_n) = \mathrm{L}_n(G^{-1}(\tau), X_1, \ldots, X_n) \, .$$

On a donc

$$\widetilde{\mathrm{L}_n}(\tau, X_1, \ldots, X_n) = \mathrm{L}_n(\theta, X_1, \ldots, X_n) \leq \mathrm{L}_n(\hat{\theta}_n^{\mathrm{MV}}, X_1, \ldots, X_n) = \widetilde{\mathrm{L}_n}(\widehat{\tau}_n, X_1, \ldots, X_n) \, . \qquad \square$$

Exemple I-2.19. Soit $(X_1, \ldots, X_n)$ un n-échantillon d'un modèle statistique exponentiel

$$\big(\mathbb{R}_+, \mathscr{B}(\mathbb{R}_+), \{\mathrm{Expo}(\theta) \, , \, \theta \in \Theta = \mathbb{R}_+^*\}\big) \, .$$

Rappelons que pour tout $\theta \in \Theta$, $\mathrm{Expo}(\theta)$ a une densité par rapport la mesure de Lebesgue sur $\mathbb{R}_+^*$:

$$q_\theta(x) = \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\{x \geq 0\}} \, .$$

La log-vraisemblance normalisée s'écrit

$$\theta \mapsto \ell_n(\theta) = \log \theta - \theta \bar{X}_n \, , \qquad \bar{X}_n := n^{-1} \sum_{i=1}^n X_i \, ,$$

donc $\ell_n'(\theta) = 0$ si et seulement si $\theta = 1/\bar{X}_n$. On vérifie que c'est un maximum global, donc $\hat{\theta}_n^{\mathrm{MV}} = \frac{1}{\bar{X}_n}$. Par invariance par changement de paramétrisation, on en déduit sans calcul que l'estimateur du maximum de vraisemblance pour un n-échantillon de loi exponentielle de paramètre $\tau = 1/\theta$, $\theta \in \Theta = \mathbb{R}_+^*$ est $\widehat{\tau}_n := \bar{X}_n$. $\diamondsuit$

I-2.3.2 Généralisation

Nous avons défini la vraisemblance pour un modèle statistique de type n-échantillon. Il est toutefois possible d'étendre l'estimateur du maximum de vraisemblance à des modèles plus généraux. Considérons à titre d'exemple le modèle de régression Exemple I-1.11.

Exemple I-2.20 (Régression linéaire).

$$(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{p_\theta \cdot \lambda_{\mathrm{Leb}}^{\otimes n} \, , \, \theta \in \Theta\})$$

où

$$p_\theta(y_1, \ldots, y_n) := \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^n \left\{ y_k - \sum_{\ell=0}^p \beta_\ell \varphi_\ell(\mathbf{x}_k) \right\}^2 \right) \, ,$$

et $\varphi_\ell : \mathbb{R}^d \mapsto \mathbb{R}$ sont des fonctions connues, $\ell = 0, \cdots, p$. Les paramètres à estimer sont donnés dans ce cas par :

$$\theta = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2) \in \Theta := \mathbb{R}^{p+1} \times \mathbb{R}_+^* \, .$$

Pour estimer le paramètre $\beta = (\beta_0, \ldots, \beta_p)$, l'estimateur du maximum de vraisemblance coïncide avec l'estimateur des moindres carrés défini comme le minimiseur par rapport aux paramètres inconnus $(\beta_0, \ldots, \beta_p)$ de la quantité

$$J_n(\beta_0, \ldots, \beta_p) := \sum_{k=1}^n \left\{ Y_k - \sum_{\ell=0}^p \beta_\ell \varphi_\ell(\mathbf{x}_k) \right\}^2 \, .$$

Ce problème d'optimisation admet une solution qui est caractérisée par le système d'équations

$$\nabla J_n(\beta_0, \ldots, \beta_p) = \left[\frac{\partial J_n}{\partial \beta_0}(\beta_0, \ldots, \beta_p), \ldots, \frac{\partial J_n}{\partial \beta_p}(\beta_0, \ldots, \beta_p) \right]^T = \mathbf{0}_{(p+1)\times 1} \, .$$

On obtient ainsi un système d'équations linéaires

$$\sum_{k=1}^{n} \left\{ Y_k - \sum_{\ell=0}^{p} \beta_\ell \varphi_\ell(\mathbf{x}_k) \right\} \varphi_j(\mathbf{x}_k) = 0 \,, \quad j = 0, \ldots, p \,. \tag{I-2.18}$$

En notant $\mathbf{Y}$ le vecteur de $\mathbb{R}^n$ de composantes $(Y_1, \cdots, Y_n)^T$, et Φ la matrice de régression $n \times (p+1)$ définie par

$$\Phi := [\varphi_0, \ldots, \varphi_p], \quad \text{où nous avons posé} \quad \varphi_i = [\varphi_i(\mathbf{x}_1), \varphi_i(\mathbf{x}_2), \ldots, \varphi_i(\mathbf{x}_n)]^T$$

le système d'équations (I-2.18) peut s'écrire,

$$\Phi^T \Phi \boldsymbol{\beta} = \Phi^T \mathbf{Y} \,.$$

Si la matrice $\Phi^T\Phi$ est inversible, ce système d'équations admet une solution unique donnée par

$$\hat{\boldsymbol{\beta}}_n := (\Phi^T \Phi)^{-1} \Phi^T \mathbf{Y} \,.$$

Cet estimateur des paramètres de régression $\boldsymbol{\beta}$ est appelé *estimateur des moindres carrés*. Pour estimer la variance σ^2, nous pouvons par exemple considérer

$$\hat{\sigma}_n^2 \; \frac{1}{n} \sum_{k=1}^{n} \left\{ Y_k - \sum_{\ell=0}^{p} \hat{\beta}_{n,\ell} \varphi_\ell(\mathbf{x}_k) \right\}^2 = n^{-1} \left\| \mathbf{Y} - \Phi \hat{\boldsymbol{\beta}}_n \right\|^2 \,. \qquad \diamondsuit$$

I-2.4 M-estimateurs

L'estimateur du maximum de vraisemblance, introduit section I-2.3, est défini comme la solution d'un problème de maximisation : c'est la valeur du paramètre qui maximise la vraisemblance des observations. C'est un cas particulier d'une très vaste classe de méthodes, où l'estimateur est obtenu comme la solution d'un problème d'optimisation.

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'un modèle statistique $(\mathsf{X}, \mathscr{X}, \{\mathbb{Q}_\theta \,,\, \theta \in \Theta\})$. Nous posons pour tout $\theta \in \Theta$, $\mathbb{P}_\theta := \mathbb{Q}_\theta^{\otimes n}$.

Considérons une fonction $m : \Theta \times \mathsf{X} \to \mathbb{R}$, $(\theta, x) \mapsto m(\theta, x)$ telle que pour tout $\theta, \theta_0 \in \Theta$, $\mathbb{E}_{\theta_0}[|m(\theta, X_1)|] < \infty$. Considérons, pour tout $\theta_0 \in \Theta$, la fonction $M_{\theta_0} : \Theta \to \mathbb{R}$, définie par

$$\theta \mapsto M_{\theta_0}(\theta) := \mathbb{E}_{\theta_0}[m(\theta, X_1)] \,. \tag{I-2.19}$$

Supposons de plus que m est telle que, pour tout $\theta_0 \in \Theta$, la fonction M_{θ_0} atteigne son maximum au point θ_0. Nous ne disposons que des observations $(X_1, \ldots, X_n)$ et la fonction M_{θ_0} n'est donc pas connue. Par contre, nous pouvons l'estimer par une moyenne empirique.

Soit $M_n : \Theta \to \mathbb{R}$ la fonction définie par

$$\theta \mapsto M_n(\theta) := n^{-1} \sum_{i=1}^{n} m(\theta, X_i) \,.$$

En supposant que, pour tout $(\theta, \theta_0) \in \Theta \times \Theta$, $\mathbb{E}_{\theta_0}[m^2(\theta, X_1)] < \infty$, l'inégalité de Bienayme-Tchebychev montre que pour tout $\delta > 0$,

$$\mathbb{P}_{\theta_0} \left(\left| M_n(\theta) - M_{\theta_0}(\theta) \right| \geq \delta \right) \leq \frac{\mathrm{Var}_{\theta_0}(m_\theta(X_1))}{n\delta^2} \,.$$

Donc, quand n est grand, la fonction $\theta \mapsto M_n(\theta)$ approche la fonction limite $\theta \mapsto M_{\theta_0}(\theta)$. Nous appelons M-estimateur toute solution du problème d'optimisation

$$\hat{\theta}_n \in \operatorname*{arg\,max}_{\theta \in \Theta} M_n(\theta) \,. \tag{I-2.20}$$

Dans de très nombreux cas, ce problème d'optimisation n'admet pas de solution "explicite". Le plus souvent, un M-estimateur sera donc obtenu en pratique en mettant en oeuvre une procédure numérique.

Si l'on suppose que *(i)* pour tout $x \in \mathsf{X}$ la fonction $\theta \mapsto m(\theta,x)$ est différentiable et *(ii)* si l'ensemble des paramètres Θ est ouvert, en posant

$$\psi(\theta,x) = \nabla_\theta m(\theta,x) := \left[\frac{\partial m}{\partial \theta^{(1)}}(\theta,x), \ldots, \frac{\partial m}{\partial \theta^{(d)}}(\theta,x) \right]^T ,$$

nous avons

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta m(\theta,X_i)|_{\theta=\hat{\theta}_n} = \frac{1}{n}\sum_{i=1}^{n} \psi(\hat{\theta}_n,X_i) = 0 \tag{I-2.21}$$

ce qui permet – dans ce cas – d'interpréter un M-estimateur comme un Z-estimateur. On peut bien entendu construire des Z-estimateurs qui ne soient pas naturellement associés à un problème d'optimisation. De même, on peut construire des M-estimateurs associés à des fonctions m qui ne sont pas différentiables et qui ne sont donc pas des Z-estimateurs.

Dans l'eq. (I-2.20), l'estimateur est obtenu comme la solution d'un problème de maximisation. Dans certains cas, il est plus naturel de considérer des estimateurs solutions de problèmes de *minimisation*. La méthode de construction de tels estimateurs ainsi que leurs propriétés sont similaires.

Exemple I-2.21 (Méthodes des moindres carrés pour le paramètre de translation). On veut estimer le paramètre de translation dans l'exemple I-2.6 (dont nous reprenons les notations). Supposons que F soit la fonction de répartition d'une loi symétrique ($F(x) = 1 - F(-x)$ pour tout $x \in \mathbb{R}$), admettant un moment d'ordre 2, $\sigma^2 = \int x^2 F(\mathrm{d}x) < \infty$. Posons

$$m(\theta,x) := -(x-\theta)^2 .$$

Pour tout $\theta, \theta_0 \in \Theta$, nous avons

$$M_{\theta_0}(\theta) = -(\theta_0 - \theta)^2 - \sigma^2 .$$

La fonction $\theta \mapsto M_{\theta_0}(\theta)$ admet un maximum unique au point θ_0. Le M-estimateur associé à la fonction m est donc la valeur du paramètre $\hat{\theta}_n$ qui maximise la fonction

$$\theta \mapsto M_n(\theta) = -n^{-1}\sum_{i=1}^{n} (X_i - \theta)^2 \tag{I-2.22}$$

Cette fonction a un maximum unique, qui est donné par $\hat{\theta}_n = n^{-1}\sum_{i=1}^n X_i$. La maximisation de (I-2.22) équivaut à la minimisation de la fonction

$$\theta \mapsto n^{-1}\sum_{i=1}^{n} (X_i - \theta)^2 . \tag{I-2.23}$$

Dans ce cas précis, il est plus naturel de considérer le problème de *minimisation* que de maximisation. L'estimateur minimisant (I-2.23) est appelé *estimateur des moindres carrés*. La fonction $\theta \mapsto m(\theta,x)$ est différentiable pour tout $x \in \mathbb{R}$. Le maximum de la fonction (I-2.22) est l'unique solution de l'équation

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

La moyenne empirique est à la fois un Z-estimateur et un M-estimateur.

La médiane empirique est aussi un M-estimateur : elle est solution du problème d'optimisation

$$M_n(\hat{\theta}_n) = \arg\max_{\theta\in\mathbb{R}} M_n(\theta), \qquad \text{où} \qquad M_n(\theta) = n^{-1}\sum_{i=1}^{n} |X_i - \theta| ;$$

voir par exemple la discussion dans l'Exemple I-2.6. ♢

I-2.4.1 Divergence de Kullback-Leibler

Définition I-2.22 (Divergence de Kullback-Leibler). *Soient $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités définies sur un espace mesurable $(\mathsf{X}, \mathscr{X})$. Soit μ une mesure positive telle que $\mathbb{P}_0 \ll \mu$ et $\mathbb{P}_1 \ll \mu$ (nous pouvons prendre par exemple $\mu = \mathbb{P}_0 + \mathbb{P}_1$). Notons p_0 et p_1 les densités de $\mathbb{P}_0$ et $\mathbb{P}_1$ par rapport à la mesure μ (voir le théorème A.47).*

On appelle divergence de Kullback–Leibler *(ou encore* entropie relative*) entre les distributions $\mathbb{P}_0$ et $\mathbb{P}_1$ la quantité*

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) := \int_{\{x:\, p_0(x)>0\}} p_0(x) \log \frac{p_0(x)}{p_1(x)} \mu(\mathrm{d}x)\,, \tag{I-2.24}$$

avec la convention $\log(1/0) = +\infty$.

Nous montrerons (Théorème I-2.23) que la définition de la divergence de Kullback-Leibler $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1)$ ne dépend pas du choix la mesure dominante μ.

L'intégrale (I-2.24) est toujours bien définie, bien qu'éventuellement infinie. En effet, l'intégrale de la partie négative $\left(p_0(x) \log \frac{p_0(x)}{p_1(x)}\right)_-$ où $y_- := -\max(-y, 0)$ est finie :

$$\begin{aligned}
\int_{\{x:\, p_0(x)>0\}} \left(p_0(x) \log \frac{p_0(x)}{p_1(x)}\right)_- \mu(\mathrm{d}x) &= \int p_0(x) \log \frac{p_1(x)}{p_0(x)} \mathbb{1}\left(p_0(x) > 0,\, p_0(x) \le p_1(x)\right) \mu(\mathrm{d}x) \\
&\le \int p_0(x) \frac{p_1(x)}{p_0(x)} \mathbb{1}(p_0(x) > 0)\, \mu(\mathrm{d}x) \\
&\le \int p_1(x)\, \mu(\mathrm{d}x) = 1,
\end{aligned}$$

où nous avons utilisé que $\max(\log(y), 0) \le y$ pour tout $y \ge 1$. Il s'en suit que $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1)$ est toujours définie et que :

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) > -\infty. \tag{I-2.25}$$

Notons que $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1)$ n'est pas une distance, puisqu'elle n'est pas symétrique en $\mathbb{P}_0$, $\mathbb{P}_1$ et qu'elle ne vérifie pas l'inégalité triangulaire. Nous avons rassemblé dans l'énoncé suivant les propriétés importantes de la divergence de Kullback–Leibler. Celles-ci reposent sur l'inégalité de Jensen (Lemme IV-1.4).

Théorème I-2.23. *Soient $(\mathsf{X}, \mathscr{X})$ un espace mesurable et $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités.*

(i) $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) < +\infty$ *implique* $\mathbb{P}_0 \ll \mathbb{P}_1$.

(ii) $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) \in [0, \infty]$ *et* $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$ *si et seulement si* $\mathbb{P}_0 = \mathbb{P}_1$.

(iii) $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1)$ *ne dépend pas du choix de la mesure de domination.*

Démonstration. (i) Remarquons en effet que, s'il existe $A \in \mathscr{X}$ tel que $\mathbb{P}_1(A) = 0$ et $\mathbb{P}_0(A) > 0$, c'est-à-dire si $p_1(x) = 0$ pour μ-presque tout $x \in A$ et que $\mu\{x \in A : p_0(x) > 0\} > 0$ alors

$$\int_{x:p_0(x)>0} \left(p_0(x) \log \frac{p_0(x)}{p_1(x)}\right)_+ \mu(\mathrm{d}x) \ge \int_{x\in A:p_0(x)>0} \infty\, \mu(\mathrm{d}x) = \infty.$$

Ceci montre en fait que $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) < \infty$ seulement si $\mathbb{P}_0 \ll \mathbb{P}_1$.

(ii) Remarquons tout d'abord que si $p_0(x) = p_1(x)$ $\mathbb{P}_0$–p.s., alors $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$. Il nous suffit donc de montrer $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) \ge 0$ et $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$ implique $p_0(x) = p_1(x)$ $\mathbb{P}_0$-p.s. On applique l'inégalité de Jensen (Lemme IV-1.4) à la variable aléatoire p_1/p_0, et on obtient, par convexité de $-\log$,

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) \ge -\log \int p_0(x) \left(\frac{p_1(x)}{p_0(x)}\right) \mu(\mathrm{d}x) = 0\,.$$

De plus l'égalité a lieu si et seulement si $p_1 = c_0 p_0$ $\mathbb{P}_0$-p.s., donc si $p_1 = p_0$ $\mathbb{P}_0$-p.s. puisque ce sont deux densités.

(iii) Supposons que $\mathbb{P}_0 \ll \mathbb{P}_1$. Par le théorème A.47, la probabilité $\mathbb{P}_0$ admet une densité par rapport à $\mathbb{P}_1$, que nous notons $\tilde{f}_0$ (et qui est unique à une $\mathbb{P}_1$ − p.s.-équivalence près). Nous allons démontrer que, dans ce cas,

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \int \tilde{f}_0(x) \log(\tilde{f}_0(x)) \mathbb{P}_1(\mathrm{d}x) \,, \tag{I-2.26}$$

avec la convention $0 \log(0) = 0$. Comme $\mathbb{P}_1 = p_1 \cdot \mu$ et $\mathbb{P}_0 = \tilde{f}_0 \cdot \mathbb{P}_1$, nous avons $\mathbb{P}_0 = \tilde{p_0} p_1 \cdot \mu$. Comme la densité est unique à une μ-équivalence près par le théorème A.47, nous avons $p_0 = \tilde{f}_0 p_1$ μ-presque partout. Il s'en suit

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) &= \int_{\{x:\, \tilde{f}_0(x) p_1(x) > 0\}} \tilde{f}_0(x) p_1(x) \log \tilde{f}_0(x) \, \mu(\mathrm{d}x) \\
&= \int_{\{x:\, \tilde{f}_0(x) > 0\}} \tilde{f}_0(x) \log(\tilde{f}_0(x)) \mathbb{P}_1(\mathrm{d}x) \,.
\end{aligned}$$

L'assertion (i) et (I-2.26) montre que $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1)$ ne dépend pas du choix de la mesure de domination : soit $\mathbb{P}_0 \ll \mathbb{P}_1$ et alors on a (I-2.26), dont le membre de droite ne dépend pas de μ, soit $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \infty$. □

I-2.4.2 Maximum de vraisemblance et *M*-estimation

Nous allons maintenant montrer que l'estimateur du maximum de vraisemblance est un M-estimateur. Soit $(X_1, \ldots, X_n)$ un n-échantillon d'un modèle statistique paramétrique dominé $(\mathsf{X}, \mathscr{X}, \{q_\theta \cdot \mu \,,\, \theta \in \Theta\})$. Notons pour tout $\theta \in \Theta$, $\mathbb{P}_\theta := \mathbb{Q}_\theta^{\otimes n}$ où $\mathbb{Q}_\theta := q_\theta \cdot \mu$.

Supposons que pour tout $(\theta, x) \in \Theta \times \mathsf{X}$, $q_\theta(x) > 0$ et posons

$$m(\theta, x) := \log q_\theta(x) \,. \tag{I-2.27}$$

Supposons que pour tout $\theta, \theta_0 \in \Theta$,

$$\mathbb{E}_{\theta_0} [|\log q_\theta(X_1)|] < \infty \,.$$

Considérons, pour tout $\theta_0 \in \Theta$, la fonction $M_{\theta_0} : \theta \to \mathbb{R}$ définie par

$$M_{\theta_0} : \theta \mapsto M_{\theta_0}(\theta) = \mathbb{E}_{\theta_0} [\log q_\theta(X_1)] \,.$$

Notons que

$$\begin{aligned}
M_{\theta_0}(\theta) = \mathbb{E}_{\theta_0} [\log q_\theta(X_1)] &= \int_{\mathsf{X}} \log q_\theta(x) q_{\theta_0}(x) \mu(\mathrm{d}x) \\
&= \int q_{\theta_0}(x) \log \frac{q_\theta(x)}{q_{\theta_0}(x)} \mu(\mathrm{d}x) + \int q_{\theta_0}(x) \log q_{\theta_0}(x) \mu(\mathrm{d}x) \\
&= -\mathrm{KL}(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta) + \mathrm{Ent}(\mathbb{Q}_{\theta_0}) \,,
\end{aligned}$$

où $\mathrm{Ent}(\mathbb{Q}(\theta_0))$ est l'entropie de la loi $\mathbb{Q}_{\theta_0}$, définie par

$$\mathrm{Ent}(\mathbb{Q}_{\theta_0}) = \int q_{\theta_0}(x) \log q_{\theta_0}(x) \mu(\mathrm{d}x) \,. \tag{I-2.28}$$

Par Théorème I-2.23, la fonction $\theta \mapsto \mathrm{KL}(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta)$ est finie et positive. Nous avons $\mathrm{KL}(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta) = 0$ si et seulement si $\mathbb{Q}_{\theta_0} = \mathbb{Q}_\theta$ ce qui équivalent à $\theta = \theta_0$. Par conséquent, pour tout $\theta \neq \theta_0$ nous avons

$$M_{\theta_0}(\theta) < M_{\theta_0}(\theta_0) \,,$$

ce qui montre que la fonction $\theta \mapsto M_{\theta_0}(\theta)$ admet un maximum strict en θ_0. Nous pouvons estimer cette quantité en calculant la moyenne empirique

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

qui est la log-vraisemblance de l'observation (voir (I-2.15)).

Chapitre I-3

Tests et régions de confiance

Le problème des tests en statistique est différent de celui de l'estimation. On fait une hypothèse a priori sur le paramètre inconnu et il s'agit de décider à l'aide des observations si cette hypothèse est vérifiée (dans un sens que nous préciserons) ou non. Contrairement aux problèmes d'estimation, le problème n'est pas d'approcher la valeur numérique d'un paramètre θ mais de répondre à une question sur ce paramètre en étant capable d'évaluer la "fiabilité" de la réponse.
La procédure consiste à définir l'hypothèse à tester (dite "hypothèse nulle") et une alternative, définir une statistique, puis définir un ensemble de valeurs de cette statistique qui conduiront à rejetter l'hypothèse nulle. Cette construction doit aussi garantir un taux d'erreur fixé a priori.

Les notions principales sont celles d'hypothèses nulle et alternative, de région d'acceptation et de rejet d'un test, de fonction puissance d'un test, de niveau et de taille d'un test, de p-valeur et de fonction pivotale.

I-3.1 Tests d'hypothèse

Exemple I-3.1. On dispose de mesures à valeur réelles, pour l'étude desquelles on considère le modèle statistique $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathrm{N}(\theta,1)\,,\ \theta\in\Theta\})$ où $\Theta=\mathbb{R}$. Sur la base de ces mesures, on souhaite répondre à la question : l'espérance est-elle inférieure ou égale à $\theta_0 = 2$? On veut aussi une procédure de test qui garantisse que l'erreur consistant à répondre "non" alors que l'espérance est bien inférieure ou égale à θ_0, soit inférieure à $\alpha\%$.
On veut donc tester l'hypothèse nulle $H_0 : \theta\in\Theta_0$ où $\Theta_0 :=]-\infty, 2]$.
On va se donner une hypothèse alternative H_1, par exemple $\theta\in\Theta_1$ où $\Theta_1 :=]2,+\infty[$. Il faut que $\Theta_0\cap\Theta_1=\emptyset$.
Une règle de décision naturelle est de se baser sur la moyenne empirique $\bar{X}_n$ des observations Z, et de considérer que dès que Z est dans la zone $\{\bar{X}_n \le c_\alpha\}$, on accepte l'hypothèse nulle H_0. Formellement, on définit la fonction $\phi : Z\mapsto \mathbb{1}_{\bar{X}_n > c_\alpha}$ qui retourne 0 si on accepte l'hypothèse H_0 et qui retourne 1 si on la rejette.
Ce seuil c_α est choisi pour garantir le niveau maximal d'erreur fixé i.e. : $\mathbb{P}_\theta(\bar{X}_n > c_\alpha)\le\alpha$ et ce pour tout $\theta\in\Theta_0$ (la condition "pour tout $\theta\in\Theta_0$" est en écho à "alors que l'espérance est bien inférieure à θ_0" et l'événement dont on calcule la probabilité correspond à "on répond non").
Il est attendu que c_α soit (légèrement) supérieur à θ_0. ♢

Plus généralement, étant donnée une expérience statistique $(\mathrm{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta\in\Theta\})$ et Θ_0 et Θ_1 deux sous-ensembles disjoints de Θ, le but d'un test statistique est de déterminer si $\theta\in\Theta_0$ ou $\theta\in\Theta_1$. On dit qu'on fait le *test d'hypothèses*

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

H_0 est dite *l'hypothèse nulle* et H_1 est *l'hypothèse alternative*. Pour effectuer un test, une première idée consiste à construire une statistique prenant deux valeurs 0 ou 1. On parle alors de tests purs.

Définition I-3.2 (Tests purs). *Un test pur de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ est une fonction mesurable $\phi : \mathsf{Z} \to \{0,1\}$ telle que*

— *Si $\phi(Z) = 0$ on accepte l'hypothèse nulle H_0 (de façon équivalente, on rejette l'hypothèse alternative).*

— *Si $\phi(Z) = 1$ on rejette l'hypothèse nulle H_0 (de façon équivalente, on accepte l'alternative H_1).*

L'ensemble

$$\mathscr{R} := \{z \in \mathsf{Z} : \ \phi(z) = 1\}$$

est la région de rejet du test - *on parle aussi de* région critique *du test. L'ensemble $\mathscr{A} = \mathsf{Z} \setminus \mathscr{R}$ est la* région d'acceptation du test.

Exemple I-3.3. Reprenons Exemple I-3.1. Nous avons $\Theta_0 :=]-\infty, \theta_0]$ et $\Theta_1 :=]\theta_0, +\infty[$. Le test pur est défini par la fonction $\phi : (x_1, \cdots, x_n) \mapsto \mathbb{1}_{n^{-1}\sum_{i=1}^n x_i > c_\alpha}$; la région de rejet est $\mathscr{R} := \{n^{-1}\sum_{i=1}^n x_i > c_\alpha\}$ et la région d'acceptation est $\mathscr{A} := \{n^{-1}\sum_{i=1}^n x_i \leq c_\alpha\}$. ♢

Lorsque Θ_0 (ou Θ_1) est réduit à un singleton, on parle d'*hypothèse simple*. Sinon, on parle d'*hypothèse composite*. Dans l'exemple I-3.3, nous avons ainsi à faire à un test entre deux hypothèses composites.

Exemple I-3.4 (Sondage). Considérons l'exemple I-1.1. Supposons que nous cherchions à tester $\theta \leq \theta_0$ où θ est la probabilité de voter pour le candidat A. De façon plus formelle, soit $(X_1, \ldots, X_n)$ un n-échantillon d'une expérience statistique de Bernoulli, $(\{0,1\}, \mathscr{P}(\{0,1\}), \{\mathrm{Ber}(\theta)\,,\ \theta \in \Theta :=]0,1[\})$. Fixons $\theta_0 \in \Theta$.

Considérons la partition de l'espace des paramètres $\Theta_0 := \{\theta \in \Theta : \ \theta \leq \theta_0\}$ et $\Theta_1 := \{\theta \in \Theta : \ \theta > \theta_0\}$. Pour $t \geq 0$, il est naturel de considérer un test de la forme suivante

$$\phi(X_1, \ldots, X_n) = \mathbb{1}\left\{ n^{-1} \sum_{i=1}^n X_i > t \right\}.$$

On accepte l'hypothèse nulle si le nombre de votes exprimés en faveur de A est inférieur à un seuil t à choisir. On rejette l'hypothèse nulle dans le cas contraire. Dans ce cas, la région de rejet est donnée par

$$\mathscr{R} = \left\{ (x_1, \ldots, x_n) \in \{0,1\}^n : \ n^{-1} \sum_{i=1}^n x_i > t \right\}. \qquad \diamondsuit$$

Dans le cas où $\Theta \subseteq \mathbb{R}$, les tests suivants sont couramment utilisés.

— $H_0 : \ \theta = \theta_0$, contre $H_1 : \ \theta = \theta_1$. C'est un test d'une hypothèse simple contre une alternative simple.

— $H_0 : \ \theta = \theta_0$, contre $H_1 : \ \theta \neq \theta_0$, qui est test d'une hypothèse simple contre une alternative composite.

— $H_0 : \ \theta \leq \theta_0$, contre $H_1 : \ \theta > \theta_1$, qui est un test d'une hypothèse composite contre une alternative composite.

Il est fréquent que la région de rejet $\mathscr{R}$ se mette sous la forme $\mathscr{R} = \{z \in \mathsf{Z} : \ T(z) > c\}$ pour une certaine statistique $T : \mathsf{Z} \to \mathbb{R}$ et un réel c (l'inégalité peut aussi être large). La variable aléatoire T est alors appelée la *statistique de test* et c la *valeur critique* du test. On dira aussi de manière équivalente qu'on fait le test de statistique T et de valeur critique c. Dans ce cas, on rejette H_0 si, ayant observé Z on a $T(Z) > c$.

Nous utiliserons la fonction puissance pour mesurer les performances ou la fiabilité d'un test statistique ϕ.

Définition I-3.5 (Fonction puissance). *Soit $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ un modèle statistique paramétrique. La* fonction puissance $\beta_\phi : \Theta \to [0,1]$ *d'un test pur ϕ est définie par*

$$\beta_\phi : \theta \mapsto \mathbb{P}_\theta\left(\phi(Z) = 1\right) = \mathbb{P}_\theta\left(\mathscr{R}\right) \ .$$

Lorsqu'on effectue un test, c'est à dire qu'on choisit entre une hypothèse H_0 et une alternative H_1, on s'expose à **deux types d'erreur**.

— *Erreur de* 1*ère espèce :* rejeter H_0 alors que cette hypothèse est satisfaite.

— *Erreur de* 2*ème espèce :* accepter H_0 alors que cette hypothèse est erronée.

Ces erreurs sont quantifiées à l'aide de la fonction puissance de la manière suivante.

Définition I-3.6 (Taille d'un test, niveau d'un test). *La taille du test ϕ est définie par*

$$\overline{\alpha}(\phi) := \sup_{\theta\in\Theta_0} \beta_\phi(\theta) \ .$$

Soit $\alpha \in]0,1[$. Un test ϕ est de de niveau α *si sa taille est inférieure ou égale à α.*

La taille du test est la probabilité maximale de rejeter H_0 alors qu'elle était vraie. Etant donnée deux hypothèses H_0, H_1, ou de façon équivalent, Θ_0, Θ_1, il n'y a pas unicité de la fonction de test ϕ. Il est donc naturel de se poser la question du choix optimal de la fonction de test. Il est clair que si ϕ_1 et ϕ_2 sont deux tests et si

$$\begin{cases} \beta_{\phi_1}(\theta) \le \beta_{\phi_2}(\theta) & \text{pour tout } \theta \in \Theta_0 \\ \beta_{\phi_1}(\theta) \ge \beta_{\phi_2}(\theta) & \text{pour tout } \theta \in \Theta_1 \end{cases}$$

alors le test ϕ_1 doit être préféré au test ϕ_2. Cette relation définit un préordre dans l'ensemble des tests. Il faut faire attention toutefois que deux procédures ne sont pas nécessairement comparables.

Exemple I-3.7. Reprenons l'Exemple I-3.1 avec $\Theta_0 = \{\theta \le \theta_0\}$ et $\Theta_1 = \{\theta > \theta_0\}$. Donnons nous deux réels $c_1 > c_2$ et définissons deux tests $\phi_1(X_1,\cdots,X_n) = \mathbb{1}_{\{n^{-1}\sum_{i=1}^n X_i > c_1\}}$ et $\phi_2(Z) = \mathbb{1}_{\{n^{-1}\sum_{i=1}^n X_i > c_2\}}$; la statistique de test est ici $T(Z) = n^{-1}\sum_{i=1}^n X_i$ et c_1 et c_2 sont deux valeurs critiques. Les régions de rejet de ces tests sont données par $\mathscr{R}_1 := \left\{(x_1,\cdots,x_n) \in \mathbb{R}^n : n^{-1}\sum_{i=1}^n x_i > c_1\right\}$, $\mathscr{R}_2 := \left\{(x_1,\cdots,x_n) \in \mathbb{R}^n : n^{-1}\sum_{i=1}^n x_i > c_2\right\}$. Il est clair que $\mathscr{R}_1 \subset \mathscr{R}_2$, ce qui implique que, pour tout $\theta \in \Theta := \mathbb{R}$,

$$\beta_{\phi_1}(\theta) < \beta_{\phi_2}(\theta) \ .$$

Plus précisément, puisque sous $\mathbb{P}_\theta$, $n^{-1}\sum_{i=1}^n X_i \sim \mathrm{N}(\theta, 1/n)$, nous avons

$$\beta_{\phi_i}(\theta) = 1 - \Phi((c_i - \theta)\sqrt{n}) \ ,$$

où Φ désigne la fonction de répartition de la loi $\mathrm{N}(0,1)$. Par suite, la taille du test ϕ_1 est

$$\sup_{\theta\in\Theta_0} \beta_{\phi_1}(\theta) = 1 - \inf_{\theta\le\theta_0} \Phi((c_1-\theta)\sqrt{n}) = 1 - \Phi((c_1-\theta_0)\sqrt{n}) \ .$$

De même, celle de ϕ_2 est $1-\Phi((c_2-\theta_0)\sqrt{n})$. ϕ_1 a une taille strictement inférieure à celle de ϕ_2. Par contre, pour tout $\theta \in \Theta_1$, nous avons $\beta_{\phi_1}(\theta) < \beta_{\phi_2}(\theta)$: le risque de seconde espèce est donc plus faible pour le test ϕ_2 que pour le test ϕ_1.

Sur la Figure fig. I-3.1, nous traçons à gauche l'évolution de la fonction puissance du test ϕ_1 lorsque $\theta_0 = 2, n = 50$ et c_1 est choisi de sorte que

$$1 - \Phi((c_1-\theta_0)\sqrt{n}) = 0.05 \ .$$

Nous trouvons $c_1 = 2.23$. Observons qu'en $\theta = \theta_0$, la fonction puissance est égale à la taille du test et vaut 0.05 (conséquence de la définition de la valeur critique c_1). Nous représentons aussi la fonction puissance du test ϕ_2 dans le cas où $c_2 = 0.9\,c_1$: la taille de ce test est supérieure à celle de ϕ_1 mais (voir le comportement sur $]\theta_0, +\infty[$) le risque de seconde espèce est inférieur.

A droite, nous traçons la taille du test en fonction de la valeur critique c_1 : plus c est grand, plus cette erreur peut être réduite. Néanmoins, une telle stratégie revient à augmenter la taille (au sens de l'inclusion) de la région d'acceptation et donc augmente la valeur du risque de seconde espèce. Observons enfin qu'en $c = \theta_0$, la taille vaut $1/2$; et que pour atteindre une taille de test égale à 5%, la valeur critique se situe dans l'intervalle $[2.2, 2.4]$ (la valeur exacte est 2.23).♢

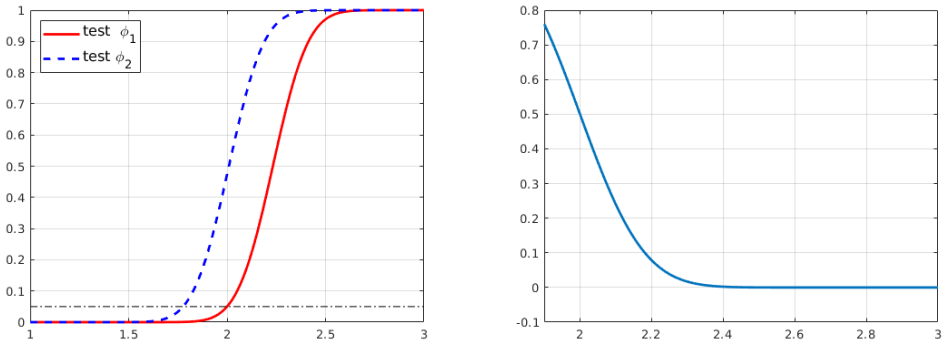

FIGURE I-3.1 – Cas $\theta_0 = 2$ et $n = 50$. Gauche : fonction puissance des tests ϕ_1 et ϕ_2 ; le seuil c_1 a été choisi pour garantir une taille de test égale à 5% (le trait horizontal est la droite d'équation $y = 0.05$) et on a pris $c_2 = 0.9\,c_1$. Droite : taille du test ϕ_1 en fonction de la valeur critique c_1.

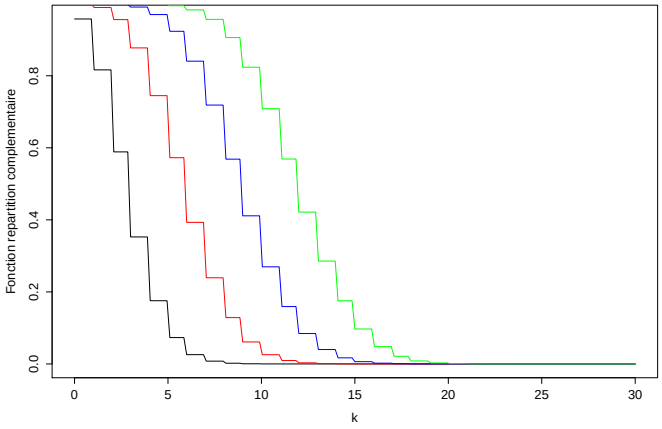

FIGURE I-3.2 – Fonction de répartition complémentaire de la loi binômiale pour $n = 30$ et $\theta = 0.1$ (noir), 0.2 (rouge), 0.3 (bleu) et 0.4 (vert).

Exemple I-3.8 (Sondage). Considérons le cas du sondage (voir exemple I-1.1). Nous allons considérer comme statistique de test $T(X_1,\dots,X_n) = S_n = \sum_{i=1}^n X_i$; nous allons tout d'abord chercher à déterminer la taille du test. La situation est ici élémentaire car la loi de $S_n = \sum_{i=1}^n X_i$ est connue de façon explicite : pour tout $k \in \{0,\dots,n\}$ et $\theta \in \Theta = [0,1]$, nous avons

$$\mathbb{P}_\theta(S_n = k) = p_\theta(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k} .$$

• Nous allons tout d'abord montrer que, pour tout seuil c fixé, la fonction $\theta \mapsto \mathbb{P}_\theta(S_n \geq c)$ est *croissante*, autrement dit, que

$$\text{pour tout } 0 \leq \theta \leq \vartheta \leq 1, \qquad \mathbb{P}_\theta(S_n \geq c) \leq \mathbb{P}_\vartheta(S_n \geq c) . \tag{I-3.1}$$

Ceci est illustré dans fig. I-3.2. Soient $U_1,\dots,U_n$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$, $0 \leq \theta \leq \vartheta \leq 1$ et $c \in \mathbb{R}$. Pour tout $i \in \{1,\dots,n\}$, on définit $A_i = \mathbb{1}_{\{U_i \leq \theta\}}$, $B_i = \mathbb{1}_{\{U_i \leq \vartheta\}}$ de sorte que $A_i \leq B_i$. On a alors $\sum_{i=1}^n A_i$ suit la loi binomiale de paramètre n et θ, $\sum_{i=1}^n B_i$ la loi binomiale de paramètre n et ϑ et $\sum_{i=1}^n A_i \leq \sum_{i=1}^n B_i$. Par conséquent,

$$\mathbb{P}_\theta(S_n \geq c) = \mathbb{P}\left(\sum_{i=1}^n A_i \geq c\right) \leq \mathbb{P}\left(\sum_{i=1}^n B_i \geq c\right) = \mathbb{P}_\vartheta(S_n \geq c) \quad .$$

Ainsi, pour tout c, nous avons

$$\sup_{\theta \leq \theta_0} \mathbb{P}_\theta(A_n > c) = \mathbb{P}_{\theta_0}(A_n > c) .$$

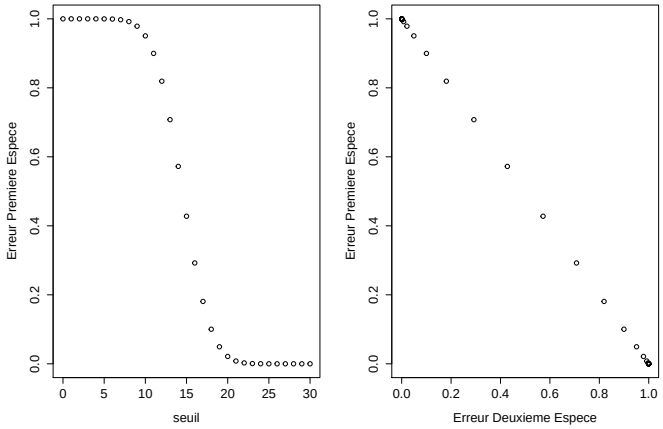

FIGURE I-3.3 – Figure de gauche : Erreur de première espèce en fonction du seuil ; Figure de droite : Erreur de première espèce en fonction de l'erreur de deuxième espèce

• Pour un seuil c donné, nous pouvons maintenant aisément calculer la taille du test. Elle est donnée par

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(A_n > c) = \mathbb{P}_{\theta_0}(A_n > c) \ .$$

Nous avons visualisé dans la fig. I-3.3 l'évolution de la taille du test en fonction du seuil c. Puisque S_n ne prend que des valeurs entières, on voit que pour un niveau $\alpha \in]0,1[$ donné, il n'existe pas forcément de seuil c_α garantissant que la taille du test soit exactement égale à ce niveau α. Nous poursuivrons cette remarque dans le chapitre **??**, lorsque nous parlerons des tests randomisés.

Pour obtenir un test dont la taille soit inférieure ou égale à un niveau α fixé, il suffit donc de choisir le seuil c_α comme le plus petit entier pour lequel $\mathbb{P}_{\theta_0}(A_n > c_\alpha) \le \alpha$. On remarque que, pour tout $\theta \in \Theta_1$, nous avons

$$\mathbb{P}_\theta(A_n > c_\alpha) \ge \mathbb{P}_{\theta_0}(A_n > c_\alpha) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(A_n > c_\alpha) \ . \qquad \diamondsuit$$

De manière plus générale, chercher à minimiser la taille d'un test (parmi une famille de tests donnée) nous conduit à choisir la région de rejet la plus petite possible. A l'inverse, la maximisation de la fonction de puissance du test sur Θ_1 nous conduit à choisir la région de rejet la plus grande possible. La détermination de la région de rejet nous oblige donc à réaliser un compromis entre deux objectifs qui sont contradictoires. Pour résoudre cette difficulté, nous sommes conduits à convenir d'une règle. La plus communément admise est celle de *Neyman-Pearson* qui se fonde sur une dissymétrie entre les hypothèses H_0 et H_1. Mathématiquement, on se donne un niveau $\alpha \in]0,1[$ typiquement petit, de l'ordre de 5%, 10^{-3} ou 10^{-6} selon les applications - et on construit le test ϕ de façon à garantir que sa taille $\overline{\alpha}(\phi)$ soit inférieure ou égale à α. On dit alors que le test est de *niveau* α. Ayant borné la taille du test par α, il est naturel de chercher à maximiser la fonction puissance du test $\beta_\phi(\theta)$ pour tout $\theta \in \Theta_1$.

Exemple I-3.9. Dans l'exemple I-3.1, on garantit un niveau α si on choisit de rejeter H_0 lorsque $\sqrt{n}(\bar{X}_n - \theta_0) > z_{1-\alpha}$, où $z_{1-\alpha}$ est le $(1-\alpha)$-quantile de la loi $\mathrm{N}(0,1)$. En effet, la région de rejet du test est alors

$$\mathscr{R} = \left\{ (x_1,\cdots,x_n) \in \mathbb{R}^n \,:\, n^{-1}\sum_{i=1}^n x_i > \theta_0 + z_{1-\alpha}/\sqrt{n} \right\}$$

et on a bien

$$\mathbb{P}_{\theta_0}\left(n^{-1}\sum_{i=1}^n X_i \in \mathscr{R} \right) = \mathbb{P}_{\theta_0}\left(\sqrt{n}(\bar{X}_n - \theta_0) > z_{1-\alpha} \right) = \mathbb{P}(\mathrm{N}(0,1) > z_{1-\alpha}) = \alpha \ .$$

Permutons maintenant les hypothèses H_0 et H_1, c'est à dire que l'on fait le test

$$H_0 : \ \theta > \theta_0, \quad \text{contre} \quad H_1 : \ \theta \le \theta_0 \ .$$

Un calcul élémentaire montre que la fonction puissance du test de la forme $\mathbb{1}_{\bar{X}_n \le c}$ est

$$\theta \mapsto \Phi(\sqrt{n}(c-\theta))$$

où Φ est la fonction de répartition d'une loi $\mathrm{N}(0,1)$; et par suite, la taille du test est

$$\Phi(\sqrt{n}(c-\theta_0)) \, .$$

La valeur critique c peut être choisie pour que la taille atteigne le niveau : on prend $c = \theta_0 + z_\alpha/\sqrt{n}$ où z_α est le quantile d'ordre α de la loi $\mathrm{N}(0,1)$. En conclusion, pour tester $\theta \le \theta_0$ contre $\theta > \theta_0$ (resp. pour tester $\theta > \theta_0$ contre $\theta \le \theta_0$) nous avons obtenu les régions de rejet

$$\mathscr{R} = \{\sqrt{n}(\bar{X}_n - \theta_0) > z_{1-\alpha}\} \qquad \text{resp.} \quad \mathscr{R}' = \{\sqrt{n}(\bar{X}_n - \theta_0) \le z_\alpha\} = \{\sqrt{n}(\bar{X}_n - \theta_0) \le -z_{1-\alpha}\} \, .$$

Nous observons que les zones de rejet ne sont pas complémentaires ! ♢

Poursuivons cette analyse de la dissymétrie entre les hypothèses H_0 et H_1.

Exemple I-3.10. Nous disposons d'un n-échantillon du modèle gaussien $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathrm{N}(\theta,1)\,,\ \theta \in \Theta\})$ où $\Theta := \mathbb{R}$, et nous voulons tester

$$H_0 : \theta = 0 \qquad H_1 : \theta = \theta_0$$

pour une valeur $\theta_0 > 0$ fixée. On prend la moyenne empirique $\bar{X}_n$ comme statistique de test, et on cherche une région de rejet de la forme $\{n^{-1}\sum_{i=1}^n x_i > c\}$. Pour garantir une taille (et un niveau) de test $\alpha \in]0,1[$, on prend pour région de rejet

$$\mathscr{R} := \{n^{-1}\sum_{i=1}^n x_i > \frac{z_{1-\alpha}}{\sqrt{n}}\} \, ;$$

z_u désigne le quantile d'ordre u de la loi $\mathrm{N}(0,1)$. Maintenant, permutons les hypothèses : pour tester

$$H_0' : \theta = \theta_0 \qquad H_1' : \theta = 0$$

nous considérons le test de région de rejet

$$\mathscr{R}' := \{n^{-1}\sum_{i=1}^n x_i \le \theta_0 + \frac{z_\alpha}{\sqrt{n}}\} = \{n^{-1}\sum_{i=1}^n x_i \le \theta_0 - \frac{z_{1-\alpha}}{\sqrt{n}}\} \, ;$$

Pour comprendre la dissymétrie entre les hypothèses H_0 et H_1, prenons $0 < \theta_0 < 2z_{1-\alpha}/\sqrt{n}$ ce qui entraine que $z_{1-\alpha}/\sqrt{n} > \theta_0 - z_{1-\alpha}/\sqrt{n}$. Dans ce cas, on peut découper l'ensemble des valeurs prises par $\bar{X}_n$ en trois régions.

- $\mathscr{R}_0 = \{\bar{X}_n < \theta_0 - z_{1-\alpha}/\sqrt{n}\}$ est la région dans laquelle le premier test accepte H_0 et le second rejette H_0'.
- $\mathscr{R}_1 = \{\bar{X}_n > z_{1-\alpha}/\sqrt{n}\}$ est la région dans laquelle le premier test rejette H_0 et le second accepte H_0'.
- $\mathscr{R}_2 = \{\theta_0 - z_{1-\alpha}/\sqrt{n} \le \bar{X}_n \le z_{1-\alpha}/\sqrt{n}\}$ est la région dans laquelle une même valeur de la statistique va conduire à accepter H_0 et accepter H_0'.

Ainsi, la troisième région décrit une expérience (observer $\bar{X}_n$) qui conduira à répondre que les deux valeurs sont acceptables selon que l'on mettra en place le premier test ou le second test. Les données ne permettent pas de trancher entre $\theta = 0$ et $\theta = \theta_0$. Le point de vue de Neyman-Pearson consiste à décider l'hypothèse nulle dans ce cas litigieux.♢

Cet exemple illustre le principe général selon lequel il est plus facile d'accepter l'hypothèse *nulle* H_0 plutôt que l'*alternative* H_1. Dans la pratique, une question importante est donc de bien poser le problème de test, c'est-à-dire de bien choisir l'hypothèse nulle H_0. Voici quelques heuristiques fréquemment utilisées pour faire ce choix.

- Choisir comme hypothèse H_0 celle qui est en notre défaveur : par exemple, si on veut proposer un nouveau médicament et qu'on dispose d'essais cliniques sur des patients, on mettra en priorité comme hypothèse H_0 que ce médicament est inefficace. Si cette hypothèse est rejetée, le test sera plus probant que si la même hypothèse était acceptée en tant qu'hypothèse nulle.
- Choisir comme hypothèse H_0 celle qui est la plus dangereuse : si on souhaite tester la sécurité d'un lieu avant l'implantation d'une centrale nucléaire, il est absolument crucial de détecter un risque sismique. Il est beaucoup moins grave de rejeter un lieu sans risque !

En vertu du second principe, deux groupes avec des visées et intérêts différents auront souvent des couples H_0 et H_1 inversés. Ainsi, industriels et associations de consommateurs partent rarement avec la même hypothèse nulle.

I-3.2 p-valeur

L'approche consistant à se fixer à l'avance un niveau α a l'avantage de la simplicité mais conduit à une information binaire sur l'expérience, rejet ou non de l'hypothèse nulle. La *p-valeur* permet de quantifier plus précisément le niveau d'incertitude de l'hypothèse.

Définition I-3.11 (p-valeur). *Soit $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ un modèle statistique. Soit $\{\phi_\alpha\ :\ \alpha \in [0,1]\}$ une famille de tests purs vérifiant*

(a) pour tout $\alpha \in [0,1]$, le test ϕ_α est de niveau α i.e.

$$\overline{\alpha}(\phi_\alpha) := \sup_{\theta\in\Theta_0} \beta_{\phi_\alpha}(\theta) \leq \alpha,$$

(b) pour tout $0 \leq \alpha \leq \alpha' < 1$, $\mathscr{R}_\alpha \subset \mathscr{R}_{\alpha'}$ où $\mathscr{R}_\alpha := \{z \in \mathsf{Z}\,,\ \phi_\alpha(z) = 1\}$ est la région de rejet du test.

La p-valeur de l'observation Z pour la famille de tests $\{\phi_\alpha\ :\ \alpha \in [0,1]\}$ est définie par

$$\hat{\alpha}(Z) := \inf\{\alpha \in [0,1]\ :\ Z \in \mathscr{R}_\alpha\}\ .$$

Cette définition s'applique à une famille de tests telle que l'augmentation de la taille du test (i.e. l'augmentation du risque maximal de première espèce $\sup_{\theta\in\Theta_0}\beta_\phi(\theta)$) est équivalent à la croissance (au sens de l'inclusion) des zones de rejet ; la p-valeur est la valeur minimale de la taille de test qui conduit à rejeter l'hypothèse nulle. Tout test de taille supérieure à la p-valeur de l'observation Z rejetera l'hypothèse nulle.

Etant donnée l'observation Z, la p-valeur du test est la valeur $\hat{\alpha} = \hat{\alpha}(Z)$ telle que H_0 est rejetée pour tout $\alpha > \hat{\alpha}(Z)$ et acceptée pour tout $\alpha < \hat{\alpha}$. En particulier, plus la p-valeur du test est faible, plus l'évidence suggère de rejeter l'hypothèse nulle.

Exemple I-3.12. Nous avons déjà vu que l'on obtient un test de niveau α dans l'exemple I-3.9 lorsqu'on choisit pour région de rejet $\mathscr{R}_\alpha = \{n^{-1}\sum_{i=1}^n x_i > \theta_0 + z_{1-\alpha}/\sqrt{n}\}$. Ainsi, ayant observé $Z = (X_1, \cdots, X_n)$, on va rejeter H_0 pour tout niveau α tel que $\theta_0 + z_{1-\alpha}/\sqrt{n} < \bar{X}_n$ et accepter H_0 pour tout α tel que $\bar{X}_n \leq \theta_0 + z_{1-\alpha}/\sqrt{n}$.

Pour $0 < \alpha < \alpha' < 1$, nous avons $z_{1-\alpha} > z_{1-\alpha'}$ et donc $\mathscr{R}_\alpha \subset \mathscr{R}_{\alpha'}$. Il est donc légitime de parler de p-valeur associée à cette famille de test : la p-valeur $\hat{\alpha}(Z)$ de cette famille de test est donnée par

$$\bar{X}_n = \theta_0 + z_{1-\hat{\alpha}(Z)}/\sqrt{n} \qquad \text{i.e.} \quad \sqrt{n}(\bar{X}_n - \theta_0) = z_{1-\hat{\alpha}(Z)} \quad \text{i.e.} \quad \hat{\alpha}(Z) = 1 - \Phi\left(\sqrt{n}(\bar{X}_n - \theta_0)\right)\ ,$$

où Φ est la fonction de répartition de la loi $\mathrm{N}(0,1)$.

Cet exemple illustre deux propriétés que nous allons démontrer dans un cadre plus général. La première consiste à remarquer que puisque nous considérons des tests ϕ_α de niveau α et dont la zone de rejet est de la forme $\{T(Z) \geq c_\alpha\}$ où $\alpha \mapsto c_\alpha$ est une fonction décroissante sur $]0,1[$, alors la p-valeur vérifie

$$\hat{\alpha}(Z) = \inf\{\alpha \in]0,1[\,, T(Z) \geq c_\alpha\}\ .$$

La seconde concerne la loi de la p-valeur sous H_0. Pour tout $\theta \in \Theta_0$, nous avons

$$\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) = \mathbb{P}_\theta(1 - \Phi\left(\sqrt{n}(\bar{X}_n - \theta_0)\right) \leq u) = \mathbb{P}_\theta(\Phi\left(\sqrt{n}(\bar{X}_n - \theta_0)\right) \geq 1 - u)\ .$$

Puisque Φ est une fonction de répartition d'une loi continue, il vient

$$\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) = \mathbb{P}_\theta(\sqrt{n}(\bar{X}_n - \theta_0) \geq \Phi^{-1}(1-u)) = \mathbb{P}_\theta(\sqrt{n}(\bar{X}_n - \theta) \geq \Phi^{-1}(1-u) + \sqrt{n}(\theta_0 - \theta))\ .$$

Puisque $\theta \in \Theta_0$, nous avons $\theta_0 - \theta \geq 0$ et donc

$$\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) \leq \mathbb{P}_\theta(\sqrt{n}(\bar{X}_n - \theta) \geq \Phi^{-1}(1-u))\ .$$

Enfin, sous $\mathbb{P}_\theta$, $\sqrt{n}(\bar{X}_n - \theta) \sim \mathrm{N}(0,1)$ et donc

$$\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) \leq 1 - \Phi(\Phi^{-1}(1-u)) = u\ .$$ ♢

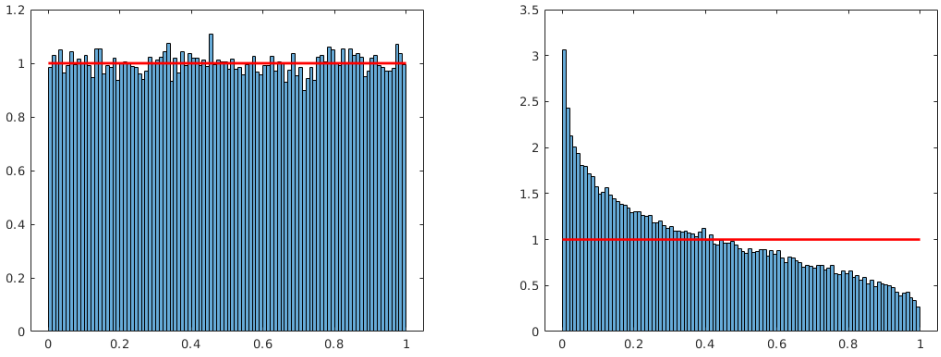

FIGURE I-3.4 – Histogramme de $N = 10^5$ p-valeurs associées à N observations indépendantes pour le test de région de rejet $\{\bar{X}_n > \theta_0 + z_{1-\alpha}/\sqrt{n}\}$ dans le cas $\theta_0 = 2$ et $n = 20$. Les observations $Z = (X_1, \cdots, X_n)$ sont un n-échantillon de $\mathrm{N}(\theta, 1)$. A gauche, on prend le cas $\theta = 2$ et à droite, le cas $\theta = 2.1$. On trace aussi en trait horizontal la droite d'équation $y = 1$ qui correspond à l'histogramme de la loi uniforme sur $[0, 1]$.

Nous venons de vérifier que pour tout $\theta \in \Theta_0$ et tout $u \in]0,1[$, $\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) \leq u$.

Dans le cas où $n = 20$ et $\theta_0 = 2$, nous visualisons N p-valeurs associées à N expériences. Les observations $Z^{(1)}, \ldots, Z^{(N)}$ ont été obtenues en simulant indépendamment N vecteurs de taille n de v.a. i.i.d. de loi $\mathrm{N}(\theta, 1)$. A gauche sur la fig. I-3.4, nous représentons l'histogramme de ces $N = 10^5$ p-valeurs dans le cas où $\theta = 2$ (donc la condition $\theta \in \Theta_0$ est vérifiée) ; et à droite, nous représentons l'histogramme de ces $N = 10^5$ p-valeurs dans le cas où $\theta = 2.1$ (donc la condition $\theta \in \Theta_0$ n'est pas vérifiée).

La proposition suivante généralise l' exemple I-3.12 et donne un moyen simple de calculer la p-valeur dans le cas important où le test est donné par une statistique T et une valeur critique c.

Proposition I-3.13 (Calcul pratique de la p-valeur). *Soit une famille de tests purs $\{\phi_\alpha, \alpha \in [0,1]\}$ de niveau α et de la forme $\phi_\alpha(Z) = \mathbb{1}_{\{T(Z) \geq c_\alpha\}}$. Supposons que la fonction $\alpha \mapsto c_\alpha$ soit décroissante sur $[0,1]$. Alors*

$$\hat{\alpha}(Z) = \inf\{\alpha \in]0,1[\,:\, T(Z) \geq c_\alpha\}.$$

Démonstration. Dans ce cas $\mathscr{R}_\alpha = \{z \in \mathsf{Z} \,:\, T(z) \geq c_\alpha\}$. La preuve découle d'une application directe des hypothèses. □

Une propriété importante de la p-valeur est que, sous certaines conditions, elle suit sous l'hypothèse nulle, une loi uniforme. Plus précisément, on a la proposition suivante.

Proposition I-3.14. *Considérons $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta \,,\, \theta \in \Theta\})$ un modèle statistique paramétrique. Considérons le test*

$$H_0 : \theta \in \Theta_0, \quad \textit{contre} \quad H_1 : \theta \in \Theta_1$$

où $\Theta_0 \cup \Theta_1 = \Theta$ et $\Theta_0 \cap \Theta_1 = \emptyset$. Soit $\{\phi_\alpha \,:\, \alpha \in [0,1]\}$ une famille de tests purs vérifiant

(a) pour tout $\alpha \in [0,1]$, le test ϕ_α est de niveau α, i.e.

$$\overline{\alpha}(\phi_\alpha) := \sup_{\theta \in \Theta_0} \beta_{\phi_\alpha}(\theta) \leq \alpha,$$

(b) pour tout $0 < \alpha \leq \alpha' < 1$, $\mathscr{R}_\alpha \subset \mathscr{R}_{\alpha'}$ où $\mathscr{R}_\alpha := \{z \in \mathsf{Z} \,,\, \phi_\alpha(z) = 1\}$ est la zone de rejet.

Alors la distribution de la statistique "p-valeur" $\hat{\alpha}(Z)$ *vérifie la propriété suivante : pour tout* $\theta \in \Theta_0$*,*

$$\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) \leq u \quad \textit{pour tout } 0 \leq u \leq 1\,. \tag{I-3.2}$$

Si pour tout $\theta \in \Theta_0$*,*

$$\mathbb{P}_\theta(Z \in \mathscr{R}_\alpha) = \alpha \quad \textit{pour tout } \alpha \in]0,1[\,, \tag{I-3.3}$$

alors

$$\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) = u \quad \textit{pour tout } u \in]0,1[\,,$$

i.e. la statistique $\hat{\alpha}(Z)$ *est uniformément distribuée sur* $(0,1)$*.*

Démonstration. (i) Soit $u \in [0,1]$. L'événement $\{\hat{\alpha}(Z) \leq u\}$ est inclus dans l'événement $\{Z \in \mathscr{R}_v\}$ pour tout $u < v$ ce qui implique

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(Z \in \mathscr{R}_u) \leq \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(Z \in \mathscr{R}_v) \leq v\,.$$

Le résultat en découle en prenant la limite $v \downarrow u$.

(ii) Comme $\{Z \in \mathscr{R}_u\} \subset \{\hat{\alpha}(Z) \leq u\}$, nous avons

$$\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) \geq \mathbb{P}_\theta(Z \in \mathscr{R}_u)\,.$$

Par conséquent, si (I-3.3) est satisfaite, alors $\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) \geq u$. Par conséquent, puisque nous avons (I-3.2), il vient que pour tout $\theta \in \Theta_0$, et $u \in [0,1]$, $\mathbb{P}_\theta(\hat{\alpha}(Z) \leq u) = u$. □

I-3.3 Régions de confiance, Fonction pivotale

Un estimateur ponctuel ne nous fournit pas directement une information “calculable” sur la qualité de cet estimateur (i.e. une quantification de l'erreur commise en utilisant $T(X_1,\ldots,X_n)$ pour estimer la fonction du paramètre $g(\theta) \in \mathbb{R}^p$).

Le concept de *région de confiance* que nous présentons dans ce chapitre donne un moyen de quantifier la précision d'estimation dans un problème statistique. Dans cette approche, l'idée d'utiliser un estimateur ponctuel $T : \mathsf{Z} \to \mathbb{R}^p$ est abandonnée en faveur d'une *région d'estimation*, c'est-à-dire une fonction $\mathscr{C}$, définie sur l'espace des observations Z, à valeurs dans l'ensemble des parties de $\mathbb{R}^p$. Pour des raisons de mesurabilité, nous supposerons que, pour tout $\theta \in \Theta$,

$$\{z \in \mathsf{Z} :\ g(\theta) \in \mathscr{C}(z)\} \in \mathscr{Z}\,.$$

Étant donnée l'observation de Z, la partie $\mathscr{C}(Z)$ doit être interprétée comme la "région" à laquelle la fonction du paramètre $g(\theta)$ appartient. Pour que $\mathscr{C}(Z)$ puisse être appelée région de confiance, on doit être capable de montrer une borne inférieure sur la probabilité que $g(\theta) \in \mathscr{C}(Z)$ (Définition I-3.15). Lorsque $g(\theta)$ est scalaire, on choisit généralement comme régions de confiance des intervalles, dont les bornes sont des fonctions mesurables de z.

Définition I-3.15 (Région de confiance). *Soient* $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ *un modèle statistique,* $g : \Theta \to \mathbb{R}^p$ *une fonction et* $\alpha \in]0,1[$*. Une fonction* $\mathscr{C} : \mathsf{Z} \to \mathscr{B}(\mathbb{R}^p)$ *est une région de confiance au niveau (de couverture)* $1-\alpha$ *pour* $g(\theta)$ *si, pour tout* $\theta \in \Theta$*,* $\{z \in \mathsf{Z} :\ g(\theta) \in \mathscr{C}(z)\} \in \mathscr{Z}$ *et*

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\left(\{z \in \mathsf{Z} :\ g(\theta) \in \mathscr{C}(z)\}\right) \geq 1-\alpha\,.$$

Si $\mathscr{C}(Z)$ est une région de confiance de niveau $1-\alpha$ pour $g(\theta)$, l'interprétation souvent proposée est : "la fonction $g(\theta)$ a une probabilité $1-\alpha$ d'appartenir à $\mathscr{C}(Z)$". Cette interprétation est **erronée**, puisque le paramètre θ est déterministe et fixé, et que l'aléa porte sur l'observation Z : une région de confiance de niveau $1-\alpha$ correspond à une région aléatoire (la statistique $\mathscr{C} : \mathsf{Z} \to \mathscr{B}(\mathbb{R}^p)$ est à valeur ensemble) qui sous $\mathbb{P}_\theta$ a une probabilité au moins égale à $(1-\alpha)$ de contenir la fonction $g(\theta)$ et ceci quelque soit la valeur $\theta \in \Theta$ du paramètre. On peut saisir le sens de cette assertion par une expérience de pensée : si l'on répétait l'expérience indépendamment un nombre N de fois (pour la même valeur de $\theta \in \Theta$) pour collecter $Z^{(1)}, \cdots, Z^{(N)}$, la proportion des régions de confiance obtenues $\mathscr{C}(Z^{(1)}), \cdots, \mathscr{C}(Z^{(N)})$ qui contiendraient la quantité inconnue $g(\theta)$ serait d'au moins $1-\alpha$.

Dans le cas où $p = 1$, la région de confiance $\mathscr{C}(Z)$ est souvent un *intervalle de confiance* qui prend l'une des trois formes suivantes :

(i) $\mathscr{C}(Z) = [m(Z), \infty[$: $m(Z)$ est une *borne inférieure de confiance*.

(ii) $\mathscr{C}(Z) =]-\infty, M(Z)]$: $M(Z)$ est une *borne supérieure de confiance*

(iii) $\mathscr{C}(Z) = [m(Z), M(Z)]$ est un *intervalle de confiance bilatéral*.

Avant de procéder à une construction plus formelle, nous allons considérer un exemple.

Exemple I-3.16 (Intervalle de confiance pour le modèle de sondage). Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle de Bernoulli

$$(\{0,1\}, \mathscr{P}(\{0,1\}), \{\mathrm{Ber}(\theta)\,,\ \theta \in \Theta := [0,1]\})\ .$$

Pour tout $\theta \in \Theta$, nous posons $\mathbb{P}_\theta := \mathrm{Ber}^{\otimes n}(\theta)$: sous $\mathbb{P}_\theta$, les v.a. $(X_1, \cdots, X_n)$ sont i.i.d. de loi $\mathrm{Ber}(\theta)$. Soit $\hat{\theta}_n := n^{-1}\sum_{i=1}^n X_i$, estimateur qui représente le nombre moyen de succès. Nous allons considérer successivement deux méthodes de construction d'un intervalle de confiance qui exploitent toutes la statistique $\hat{\theta}_n$.

• Nous avons pour tout $\theta \in \Theta$,

$$\mathrm{Var}_\theta\left(\hat{\theta}_n\right) = \mathbb{E}_\theta\left[(\hat{\theta}_n - \theta)^2\right] = \frac{\theta(1-\theta)}{n}\ ,$$

ce qui implique que

$$\sup_{\theta\in\Theta} \mathrm{Var}_\theta\left(\hat{\theta}_n\right) = 1/(4n)\ . \tag{I-3.4}$$

On peut construire un intervalle de confiance à partir de la borne (I-3.4) : pour tout $\delta > 0$ et $\theta \in \Theta$, par l'inégalité de Bienayme-Tchebychev (Lemme IV-1.2) nous avons

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

Soit $\alpha \in (0,1)$. Prenons $\delta = \delta_{n,\alpha}$ solution de l'équation $1/(4n\delta^2) = \alpha$, i.e.

$$\delta_{n,\alpha} = \frac{1}{2\sqrt{n\alpha}}.$$

On pose [1]

$$\mathscr{I}_{n,\alpha} = \left[\hat{\theta}_n \pm \frac{1}{2\sqrt{n\alpha}}\right]\ .$$

Alors, pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta\left(\theta \in \mathscr{I}_{n,\alpha}\right) \geq 1-\alpha\ .$$

Ainsi, l'intervalle $\mathscr{I}_{n,\alpha}$ est un *intervalle de confiance* pour θ de niveau $1-\alpha$.

L'interprétation de $\mathscr{I}_{n,\alpha}$ est claire : on garantit que avec une probabilité au moins égale à $(1-\alpha)$, l'intervalle aléatoire $\mathscr{I}_{n,\alpha}$ contient la quantité d'intérêt inconnnue θ. La qualité de cet intervalle se mesure à sa longueur, notée $|\mathscr{I}_{n,\alpha}|$, qui est ici égale à

$$|\mathscr{I}_{n,\alpha}| = \frac{1}{\sqrt{n\alpha}}\ .$$

L'ordre de grandeur de $\mathscr{I}_{n,\alpha}$ comme une fonction de n est $1/\sqrt{n}$; il est le même que celui de l'erreur quadratique de l'estimateur $\hat{\theta}_n$ (l'erreur quadratique est définie comme $\mathbb{E}_\theta\left[(\hat{\theta}_n - \theta)^2\right]$ et nous avons vu qu'elle dépend de n comme $O(1/n)$). Pour n fixé, on a aussi $|\mathscr{I}_{n,\alpha}| \to +\infty$ lorsque $\alpha \to 0$, cette asymptotique correspondant à un grand niveau de

1. La notation $[a \pm b]$ désigne l'intervalle $[a-b, a+b]$.

précision de l'intervalle de confiance. Il s'agit d'un compromis inévitable entre précision d'estimation (vouloir $|\mathscr{I}_{n,\alpha}|$ petit) et risque (vouloir α petit) qui sont antagonistes.

• On peut raffiner ce résultat en utilisant les inégalités de déviation exponentielle (Section IV-1.4). Comme établi au corollaire IV-1.10, pour tout $\delta > 0$,

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta \left(\left| \hat{\theta}_n - \theta \right| > \delta \right) \leq 2\exp\left(-2n\delta^2 \right) .$$

En prenant $\delta = \delta(\alpha, n)$ solution de l'équation $2\exp(-2n\delta^2) = \alpha$, i.e.

$$\delta(\alpha, n) := \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} ,$$

on définit, pour tout $\alpha > 0$,

$$\mathscr{I}^\star_{n,\alpha} := \left[\hat{\theta}_n \pm \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \right] .$$

Par construction $\mathscr{I}^\star_{n,\alpha}$ est un intervalle de confiance pour θ de niveau $1-\alpha$. On a

$$\frac{|\mathscr{I}^\star_{n,\alpha}|}{|\mathscr{I}_{n,\alpha}|} = \sqrt{2\alpha \log(2/\alpha)} \to 0 \quad \text{lorsque} \quad \alpha \to 0 .$$

De plus, pour $\alpha = 5\%$, on a un rapport de

$$\frac{|\mathscr{I}^\star_{n,\alpha}|}{|\mathscr{I}_{n,\alpha}|} = 0.61.$$

Pour $\alpha = 1\%$, le rapport devient 0.33, soit une précision 3 fois meilleure ! Ainsi, utiliser l'intervalle issu de l'inégalité d'Hoeffding améliore significativement l'intervalle de confiance, même si les ordres de grandeur de $\mathscr{I}_{n,\alpha}$ et $\mathscr{I}^\star_{n,\alpha}$ sont comparables en n. ♢

Exemple I-3.17 (Echantillon gaussien variance connue, Intervalle de confiance pour l'espérance.). Soit $(X_1, \ldots, X_n)$ un n-échantillon d'un modèle gaussien

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{ \mathrm{N}(\mu, \sigma^2) ,\ \mu \in \mathbb{R} \right\} \right) ;$$

la variance $\sigma^2 > 0$ est *connue*. Nous cherchons à construire un intervalle de confiance bilatéral $[m(X_1, \ldots, X_n), M(X_1, \ldots, X_n)]$ pour l'espérance μ, de niveau de confiance $1-\alpha$: on veut que pour tout $\mu \in \mathbb{R}$,

$$\mathbb{P}_\mu \left(\mu \in [m(X_1, \ldots, X_n), M(X_1, \ldots, X_n)] \right) \geq 1 - \alpha . \tag{I-3.5}$$

Nous allons exploiter la statistique $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$. Nous savons que pour tout $\mu \in \mathbb{R}$, sous $\mathbb{P}_\mu$,

$$G(X_1, \ldots, X_n, \mu) := \sqrt{n}(\bar{X}_n - \mu)/\sigma \sim \mathrm{N}(0,1) .$$

Nous venons d'exhiber une fonction G des observations $(X_1, \cdots, X_n)$ et du paramètre μ dont la loi sous $\mathbb{P}_\mu$ est indépendante de μ, et ce, pour tout $\mu \in \mathbb{R}$. Une telle fonction est dite *fonction pivotale*. L'intérêt de ces fonctions pour la construction d'intervalles de confiance est que, étant donné un risque $\alpha \in]0,1[$, il suffit de trouver une intervalle $[a,b]$ tel que pour tout $\mu \in \mathbb{R}$, on a $\mathbb{P}_\mu(G(X_1, \ldots, X_n, \mu) \in [a,b]) \geq 1 - \alpha$. Dans le cas présent, il suffit de prendre $b = -a = z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ d'une loi $\mathrm{N}(0,1)$; et on a l'égalité pour tout $\mu \in \mathbb{R}$:

$$\mathbb{P}_\mu \left(\sqrt{n}(\bar{X}_n - \mu)/\sigma \in \left[-z_{1-\alpha/2}, z_{1-\alpha/2} \right] \right) = 1 - \alpha .$$

Par exemple, nous avons $z_{1-\alpha/2} = 1.96$ si $\alpha = 0.05$ et $z_{1-\alpha/2} = 2.57$ si $\alpha = 0.01$. Nous en déduisons

$$\mathbb{P}_\mu \left(\mu \in \bar{X}_n \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

ce qui définit l'intervalle de confiance suivant

$$\mathscr{I}_{n,\alpha} := \left[\bar{X}_n \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right] .$$

L'intervalle de confiance bilatéral a pour diamètre $2\sigma z_{1-\alpha/2}/\sqrt{n}$ qui tend vers 0 quand $n \to \infty$, pour tout niveau de confiance $1-\alpha$ donné, à la vitesse $O(1/\sqrt{n})$. A n fixé, lorsque $\alpha \to 0$ nous avons $z_{1-\alpha} \to +\infty$ et la taille de l'intervalle de confiance tend vers $+\infty$. Nous retrouvons le même compromis entre précision d'estimation (largeur

étroite de l'intervalle de confiance) et faible risque (α petit) que dans l'exemple I-3.16.
Nous avons choisi de prendre l'intervalle $[a,b]$ de la forme $\left[-z_{1-\alpha/2}, z_{1-\alpha/2}\right]$ (intervalle symétrique), mais nous aurions pu prendre n'importe quels quantiles $z_\beta < z_\gamma$ tels que $\gamma - \beta = 1-\alpha$; il n'y a bien sûr pas unicité de la paire (β,γ) ayant cette propriété, et le choix d'une paire est guidé par exemple par le souhait d'avoir un intervalle de confiance de longueur minimale (c'est ici ce que réalise le choix $\beta = \alpha/2$ et $\gamma = 1-\alpha/2$). ♢

L'objet *fonction pivotale* (dit aussi *pivot*) que nous venons de voir dans l'exemple I-3.17 est maintenant introduit de façon générale.

Définition I-3.18 (Fonction Pivotale). *Soit* $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta\in\Theta\})$ *un modèle statistique paramétrique où* $\Theta \in \mathscr{B}(\mathbb{R}^d)$. *On dit qu'une* fonction *mesurable*

$$\begin{aligned} G : (\mathsf{Z}\times\Theta, \mathscr{Z}\otimes\mathscr{B}(\Theta)) &\to (\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p)) \\ (z,\theta) &\mapsto G(z,\theta) \end{aligned}$$

est pivotale *si,*

(i) pour tout $\theta\in\Theta$, *la fonction* $z\mapsto G(z,\theta)$ *est mesurable*

(ii) pour tout $\theta,\vartheta\in\Theta$, *la loi image de* $\mathbb{P}_\theta$ *par* $G(\cdot,\theta)$ *coïncide avec la loi image de* $\mathbb{P}_\vartheta$ *par* $G(\cdot,\vartheta)$, *i.e.*

$$\mathbb{P}_\theta(G(Z,\theta)\in A) = \mathbb{P}_\vartheta(G(Z,\vartheta)\in A), \quad \textit{pour tout } A\in\mathscr{B}(\mathbb{R}^p)\,.$$

Une fonction pivotale permet alors de construire une région de confiance $\mathscr{C}(Z)$ pour la valeur θ ou la valeur $g(\theta)$, de niveau de confiance $(1-\alpha)\in\,]0,1[$ donné. Soit $A_\alpha\in\mathscr{B}(\mathbb{R}^p)$ un ensemble tel que

$$\mathbb{P}_\theta\left(G(Z,\theta)\in A_\alpha\right) \geq 1-\alpha, \quad \text{pour tout} \quad \theta\in\Theta\,. \tag{I-3.6}$$

Rappelons que le membre de gauche de l'inégalité précédente ne dépend pas de $\theta\in\Theta$. Il s'en suit que, pour tout ensemble $A_\alpha\in\mathscr{B}(\mathbb{R}^p)$ ainsi choisi, la région définie pour tout $z\in\mathsf{Z}$ par

$$\mathscr{C}(z) = \{\theta\in\Theta :\ G(z,\theta)\in A_\alpha\}$$

est une région de confiance, de niveau de confiance $1-\alpha$.

Considérons le cas où la fonction pivotale est à valeur réelle ($p=1$). Pour $\alpha\in\,]0,1[$, notons q_α le quantile d'ordre α de la loi de la fonction pivotale $G(Z,\theta)$ (q_α ne dépend pas de θ). Il y a de nombreuses façons de satisfaire (I-3.6). On peut par exemple prendre $A_\alpha = \,]-\infty, q_{1-\alpha}]$ ou $A_\alpha = \,]q_\alpha,\infty[$. Si la loi de $G(Z,\theta)$ est symétrique, on a $q_{1-\alpha} = -q_\alpha$ il peut être plus approprié de choisir un intervalle symétrique $A_\alpha = \left]-q_{1-\alpha/2}, q_{1-\alpha/2}\right]$. Plus généralement, pour tout $0\leq p_1(\alpha) < p_2(\alpha)\leq 1$ tels que $1-\alpha = p_2(\alpha)-p_1(\alpha)$, on peut choisir $A_\alpha = \left]q_{p_1(\alpha)}, q_{p_2(\alpha)}\right]$. Le choix de A_α est souvent guidé soit par un objectif particulier soit par la volonté de minimiser le "volume" de la région de confiance, dans un sens à préciser.

Exemple I-3.19 (Echantillon gaussien variance inconnue, Intervalle de confiance pour l'espérance.). A la différence de ce que nous avons fait dans l'exemple I-3.17, nous allons maintenant déterminer des intervalles de confiance lorsque *la variance est inconnue*. Soit $(X_1,\ldots,X_n)$ un n-échantillon du modèle gaussien

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{\mathrm{N}(\mu,\sigma^2)\,,\ (\mu,\sigma^2)\in\Theta := \mathbb{R}\times\mathbb{R}_+^\star\right\}\right)\,.$$

Considérons la fonction

$$G(X_1,\ldots,X_n,\mu) := \sqrt{n}(\bar{X}_n-\mu)/S_n\,,$$

où S_n^2 est un estimateur de la variance de l'échantillon,

$$S_n^2 := (n-1)^{-1}\sum_{i=1}^n (X_i-\bar{X}_n)^2. \tag{I-3.7}$$

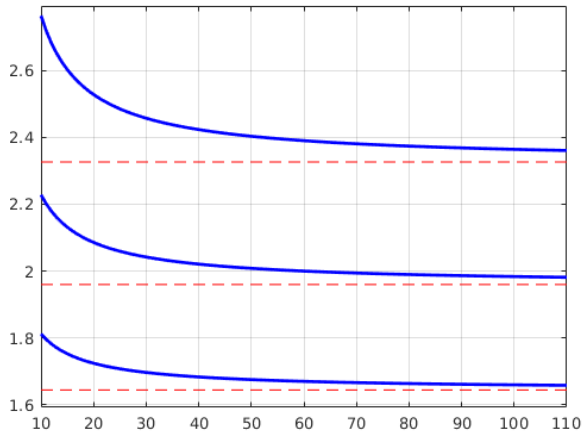

FIGURE I-3.5 – En traits pleins, quantiles $p = 0.95, 0.975$ et 0.99 d'une loi de Student à n degrés de liberté en fonction de n. Pour référence, on indique en traits pointillés la valeur des quantiles de même ordre d'une loi $\mathrm{N}(0,1)$.

Le théorème IV-3.24 montre que $G(X_1,\ldots,X_n,\mu)$ est une fonction pivotale distribuée suivant une loi de Student à $(n-1)$ degrés de liberté : pour tout $A \in \mathbb{R}$ et $\theta = (\mu,\sigma^2) \in \Theta$,

$$\mathbb{P}_\theta\left(G(X_1,\ldots,X_n,\mu) \in A\right) = \mathrm{t}_{n-1}(A)\ .$$

Soit le risque $\alpha \in]0,1[$ fixé. Notons $t_\alpha^{(n-1)}$ le quantile d'ordre α de la loi t_{n-1}. La densité de Student est une fonction paire et unimodale (de mode en 0) ce qui justifie de considérer des intervalles symétriques. Nous avons

$$\mathbb{P}_\theta\left(-\mathrm{t}_{1-\alpha/2}^{(n-1)} \le G(X_1,\ldots,X_n,\mu) \le \mathrm{t}_{1-\alpha/2}^{(n-1)}\right) = 1-\alpha.$$

En résolvant la relation précédente par rapport à μ, nous obtenons, pour tout $\theta = (\mu,\sigma^2) \in \Theta$,

$$\mathbb{P}_\theta\left(\mu \in [m(X_1,\ldots,X_n), M(X_1,\ldots,X_n)]\right) = 1-\alpha\ ,$$

avec

$$m(X_1,\ldots,X_n) := \bar{X}_n - S_n\frac{\mathrm{t}_{1-\alpha/2}^{(n-1)}}{\sqrt{n}} \quad \text{et} \quad M(X_1,\ldots,X_n) := \bar{X}_n + S_n\frac{\mathrm{t}_{1-\alpha/2}^{(n-1)}}{\sqrt{n}}\ . \qquad \diamondsuit$$

Nous avons représenté dans la fig. I-3.5 les quantiles d'ordre $p = 0.95$, $p = 0.975$ et $p = 0.99$ de la loi de Student à n degrés de liberté (traits pleins) en fonction de n. Pour comparaison, nous indiquons par un trait horizontal pointillé la valeur des quantiles de même ordre de la loi gaussienne centrée réduite. Nous voyons sur ce graphique que les quantiles d'une loi de Student à grand nombre de degrés de liberté ($n \ge 100$ par exemple) sont assez bien approchés par les quantiles d'une loi $\mathrm{N}(0,1)$.

Exemple I-3.20 (Echantillon gaussien espérance inconnue, Intervalle de confiance pour la variance). Nous allons maintenant construire un intervalle de confiance pour la variance, lorsque la moyenne est inconnue. Soit $(X_1,\ldots,X_n)$ un n-échantillon du modèle gaussien

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{\mathrm{N}(\mu,\sigma^2),\ (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+^\star\right\}\right)\ .$$

La fonction

$$G(X_1,\ldots,X_n,\sigma^2) := (n-1)\frac{S_n^2}{\sigma^2}$$

où S_n^2 est un estimateur de la variance de l'échantillon ((I-3.7)) est distribuée suivant une loi de χ^2 à $n-1$ degrés de liberté (voir Section IV-3.4). Cette fonction est donc pivotale. Si nous notons $\chi^2_\alpha(n-1)$ le quantile d'ordre α de la loi $\chi^2(n-1)$ et si nous prenons $\alpha_1+\alpha_2 = \alpha$, alors, pour tout (μ,σ^2),

$$\mathbb{P}_{\mu,\sigma^2}\left(\chi^2_{\alpha_1}(n-1) \le G(X_1,\ldots,X_n,\sigma^2) \le \chi^2_{1-\alpha_2}(n-1)\right) = 1-\alpha\ .$$

En résolvant l'équation précédente par rapport à σ^2, nous obtenons donc que :

$$\left[(n-1)S_n^2/\chi^2_{1-\alpha_2}(n-1),(n-1)S_n^2/\chi^2_{\alpha_1}(n-1)\right]$$

est un intervalle de confiance pour σ^2 de niveau de confiance $1-\alpha$.

Il n'y a pas unicité de la paire d'ordre de quantile (α_1,α_2) définissant un intervalle de confiance de niveau $1-\alpha$. Notons $L_n(S_n^2,\alpha_1,\alpha_2)$ la longueur de cet intervalle (qui est une statistique). Il est possible de montrer qu'il existe α_1^* et α_2^*, $0<\alpha_1^*<\alpha_2^*$, $\alpha_1^*+\alpha_2^*=\alpha$, tels que,

$$\mathbb{E}_{\mu,\sigma^2}[L_n(S_n^2,\alpha_1^*,\alpha_2^*)] \leq \mathbb{E}_{\mu,\sigma^2}[L_n(S_n^2,\alpha_1,\alpha_2)]$$

pour tout $(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+$ et tout (α_1,α_2) tels que $0<\alpha_1<\alpha_2$ et $\alpha_1+\alpha_2=\alpha$. On peut montrer que, lorsque n est grand, $\alpha_1^*\simeq\alpha_2^*\simeq\alpha/2$. ♢

Exemple I-3.21 (Echantillon gaussien, Région de confiance pour la moyenne et la variance). Soit $(X_1,\ldots,X_n)$ un n-échantillon du modèle gaussien

$$\left(\mathbb{R},\mathscr{B}(\mathbb{R}),\left\{\mathrm{N}(\mu,\sigma^2),\ (\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^\star\right\}\right).$$

Cette fois, nous cherchons à construire une *région de confiance* pour (μ,σ^2) de niveau de confiance $(1-\alpha)$. Notons les intervalles de confiance précédemment utilisés par

$$I_1(X_1,\ldots,X_n) := \left[\bar{X}_n - S_n\frac{\mathrm{t}^{(n-1)}_{1-\alpha/4}}{\sqrt{n}}, \bar{X}_n + S_n\frac{\mathrm{t}^{(n-1)}_{1-\alpha/4}}{\sqrt{n}}\right],$$

pour l'intervalle de confiance pour la moyenne μ de niveau de confiance $1-\alpha/2$ dans le cas de variance inconnue (voir exemple I-3.19) et :

$$I_2(X_1,\ldots,X_n) := \left[\frac{(n-1)S_n^2}{\chi^2_{1-\alpha/4}(n-1)}, \frac{(n-1)S_n^2}{\chi^2_{\alpha/4}(n-1)}\right],$$

pour l'intervalle de confiance pour la variance σ^2 de niveau de confiance $1-\alpha/2$ dans le cas d'espérance inconnue (voir exemple I-3.20). Nous avons :

$$\begin{aligned}\mathbb{P}_{\mu,\sigma^2}\left((\mu,\sigma^2)\notin I_1(X_1,\ldots,X_n)\times I_2(X_1,\ldots,X_n)\right)&\\ \leq \mathbb{P}_{\mu,\sigma^2}(\mu\notin I_1(X_1,\ldots,X_n)) + \mathbb{P}_{\mu,\sigma^2}(\sigma^2\notin I_2(X_1,\ldots,X_n)) &\leq \alpha,\end{aligned}$$

et donc $I(X_1,\ldots,X_n)=I_1(X_1,\ldots,X_n)\times I_2(X_1,\ldots,X_n)$ est une région de confiance de niveau de confiance supérieur à $1-\alpha$. En fait, on peut montrer en utilisant le théorème IV-3.24 que le niveau de confiance de cette région est exactement égal à $(1-\alpha/2)^2$. ♢

I-3.4 Construction de tests par la méthode de pivot

Les fonctions pivotales $(z,\theta)\mapsto G(z,\theta)$ ont été introduites à la définition I-3.18. Étant donné une fonction pivotale G, on peut déterminer, pour tout $\alpha\in[0,1]$, des réels t_α^- et t_α^+ tels que

$$\text{pour tout } \theta\in\Theta,\qquad \mathbb{P}_\theta\left(G(Z,\theta)\in[t_\alpha^-,t_\alpha^+]\right)\geq 1-\alpha\ .$$

En d'autres termes, l'ensemble $\hat{I}_\alpha := \{\theta\in\Theta:\ G(Z,\theta)\in[t_\alpha^-,t_\alpha^+]\}$ est une région de confiance pour θ de niveau $1-\alpha$. Attention, cette région de confiance est en général implicite et il n'est pas toujours facile d'en déduire un test. Cependant, l'exemple suivant décrit des situations classiques dans lesquelles cette résolution est en fait possible.

Exemple I-3.22 (Test sur la moyenne d'une loi $\mathrm{N}(\mu,\ \sigma^2)$ *avec* σ^2 *connue).* Soit $(X_1,\ldots,X_n)$ un n-échantillon du modèle $\left\{\mathrm{N}(\mu,\sigma^2),\ \mu\in\mathbb{R}\right\}$ où la variance σ^2 connue. La fonction

$$G(\bar{X}_n,\mu) := \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$$

est pivotale (voir exemple I-3.17) : pour tout $\mu \in \mathbb{R}$, la loi de $G(\bar{X}_n, \mu)$ est $\mathrm{N}(0,1)$.

Considérons le test

$$H_0 : \mu = \mu_0, \quad \text{contre} \quad H_1 : \mu \neq \mu_0$$

pour $\mu_0 \in \mathbb{R}$ fixé. Puisque G est pivotale, on a sous H_0 :

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1) \, .$$

On en déduit que pour un niveau de test $\alpha \in]0,1[$ donné,

$$\begin{aligned} \mathbb{P}_{\mu_0}\left(-z_{1-\alpha/2} < \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) &= \mathbb{P}_{\mu_0}\left(\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}_n < \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= 1 - \alpha \end{aligned}$$

de sorte que la région suivante

$$\mathscr{A}_\alpha := \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < n^{-1}\sum_{i=1}^n x_i < \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \right\} ,$$

définit une région d'acceptation pour un test de niveau (et de taille) α.

Dans cet exemple, la fonction puissance $\mu \mapsto \beta(\mu)$ peut être déterminée facilement, puisque la loi de $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ sous $\mathbb{P}_\mu$ reste $\mathrm{N}(0,1)$ pour tout μ. Il vient

$$\begin{aligned} \beta(\mu) &= 1 - \mathbb{P}_\mu\left(-z_{1-\alpha/2} < \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \\ &= 1 - \mathbb{P}_\mu\left(-z_{1-\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \\ &= 1 - \mathbb{P}_\mu\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) \\ &= 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) \end{aligned}$$ ♢

où Φ est la fonction de répartition de la loi de Gauss centrée réduite. La fonction $\mu \mapsto \beta(\mu)$ a un minimum au point μ_0 (qui vaut précisément α puisque la taille du test est α). Par ailleurs on vérifie aisément que $\beta(\mu) \to 1$ quand $n \to \infty$ aussi bien lorsque $\mu > \mu_0$ que lorsque $\mu < \mu_0$. Sur la fig. I-3.6, on trace la fonction puissance sur $[-1,5]$ pour différentes valeurs de n. On prend $\sigma = 1$, $\mu_0 = 2$et $\alpha = 0.05$. Le trait horizontal est la droite d'équation $y = \alpha$.

Exemple I-3.23 (Test sur la moyenne d'une loi $\mathrm{N}(\mu,\ \sigma^2)$ *avec* σ^2 *inconnue).* Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle gaussien

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{\mathrm{N}(\mu, \sigma^2),\ (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\right\}\right) .$$

On considère la fonction

$$G(X_1, \ldots, X_n; \mu) := \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \tag{I-3.8}$$

où S_n^2 est un estimateur de la variance σ^2. La fonction $G(X_1, \ldots, X_n; \mu)$ est une fonction pivotale (voir aussi l'exemple I-3.19) : Le théorème de Gosset (Théorème IV-3.24) montre que, pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$, sous $\mathbb{P}_{(\mu,\sigma^2)}$

(i) La moyenne empirique $\bar{X}_n$ et la variance empirique S_n^2 sont indépendantes.

(ii) La moyenne empirique $\bar{X}_n$ suit une loi normale $\mathrm{N}(\mu, \sigma^2/n)$.

(iii) $(n-1)S_n^2/\sigma^2$ suit une loi du χ^2 centrée à $(n-1)$ degrés de liberté.

Par conséquent, la distribution de $G(X_1, \ldots, X_n; \mu)$ suit une loi de Student $\mathrm{t}(n-1)$ (voir Section IV-3.5 pour la définition et les propriétés). Considérons le test

$$H_0 : \mu = \mu_0, \quad \text{contre} \quad H_1 : \mu \neq \mu_0$$

Sous l'hypothèse nulle (i.e. sous $\mathbb{P}_{\mu_0}$), nous avons donc

$$\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1) \, .$$

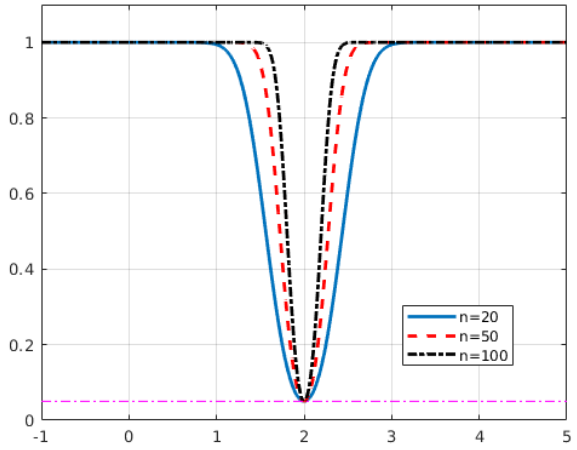

FIGURE I-3.6 – Fonction puissance dans le cas $n = 20$, $n = 50$ et $n = 100$. On a pris $\alpha = 0.05$, $\mu_0 = 2$ et $\sigma = 1$. Le trait horizontal est la droite d'équation $y = \alpha$.

Comme, par définition des quantiles $t_u^{(n-1)}$, nous avons

$$1-\alpha = \mathbb{P}_{\mu_0,\sigma^2}\left(-\mathrm{t}_{1-\alpha/2}^{(n-1)} < \frac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}} < \mathrm{t}_{1-\alpha/2}^{(n-1)}\right),$$

le test

$$\phi_\alpha(X_1,\ldots,X_n) := \mathbb{1}\left\{\left|\frac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}}\right| > t_{1-\alpha/2}(n-1)\right\}$$

est de taille α. La fonction puissance de ce test est donnée par

$$(\mu,\sigma^2) \mapsto \beta(\mu,\sigma^2) := \mathbb{P}_{\mu,\sigma^2}\left(-\mathrm{t}_{1-\alpha/2}^{(n-1)} < \frac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}} < \mathrm{t}_{1-\alpha/2}^{(n-1)}\right).$$

La loi de $(\bar{X}_n-\mu_0)/(S_n/\sqrt{n})$ ne suit plus une loi de Student car, sous $\mathbb{P}_{(\mu,\sigma^2)}$, μ_0 n'est plus l'espérance de $\bar{X}_n$. Nous écrivons

$$\frac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n-\mu)/\sigma + \sqrt{n}(\mu-\mu_0)/\sigma}{S_n/\sigma},$$

ce qui montre que cette statistique de test suit alors une loi de Student *non centrale* de paramètre de non centralité $\sqrt{n}(\mu-\mu_0)/\sigma$. ♢

Exemple I-3.24 (Test sur la variance d'une loi $\mathrm{N}(\mu,\sigma^2)$ *avec* μ *inconnu).* Soit $(X_1,\ldots,X_n)$ un n-échantillon du modèle gaussien

$$\left(\mathbb{R},\mathscr{B}(\mathbb{R}),\left\{\mathrm{N}(\mu,\sigma^2),\ (\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*\right\}\right).$$

Nous nous intéressons au test

$$H_0:\ \sigma^2=\sigma_0^2, \quad \text{contre} \quad H_1:\ \sigma^2\neq\sigma_0^2$$

où $\sigma_0^2 > 0$ est une valeur donnée. Considérons la fonction

$$G(X_1,\ldots,X_n;\sigma^2) := \frac{(n-1)S_n^2}{\sigma^2}.$$

Cette fonction est pivotale (voir aussi exemple I-3.20). Pour tout $\mu\in\mathbb{R}$ et $\sigma^2>0$, $G(X_1,\ldots,X_n;\sigma^2)\sim\chi^2(n-1)$. Donc, sous H_0 on a :

$$\mathbb{P}_{\mu,\sigma_0^2}\left(\chi^2_{\alpha/2}(n-1) < \frac{(n-1)S_n^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}(n-1)\right) = 1-\alpha,$$

ce qui conduit à considérer le test

$$\phi_\alpha(X_1,\ldots,X_n) := \mathbb{1}\left\{\frac{(n-1)S_n^2}{\sigma_0^2} \notin \left]\chi^2_{\alpha/2}(n-1),\chi^2_{1-\alpha/2}(n-1)\right[\right\}$$

dont le niveau et la taille sont α. ♢

I-3.5 Pour aller plus loin : Dualité entre régions de confiance et tests d'hypothèse de base simple

Il existe des liens étroits entre tests statistiques et région de confiance. Nous illustrons tout d'abord ces liens par un exemple.

Exemple I-3.25 (Test bilatéral pour la moyenne d'une gaussienne). Soient $(X_1,\dots,X_n)$ un n-échantillon d'un modèle gaussien

$$\left(\mathbb{R},\mathscr{B}(\mathbb{R}),\left\{\mathrm{N}(\mu,\sigma^2),\ (\mu,\sigma^2)\in\Theta:=\mathbb{R}\times\mathbb{R}_+^*\right\}\right)\ .$$

Soit $\mu_0\in\mathbb{R}$; nous cherchons à tester

$$H_0:\ \mu=\mu_0,\quad \text{contre}\quad H_1:\ \mu\neq\mu_0\ . \tag{I-3.9}$$

Nous avons construit dans l'exemple I-3.19 un intervalle de confiance pour μ de niveau de couverture $(1-\alpha)$, $\alpha\in\,]0,1[$,

$$I(Z):=\left[\bar X_n\pm S_n\frac{\mathrm{t}_{1-\alpha/2}^{(n-1)}}{\sqrt{n}}\right]\ ,\quad S_n^2:=(n-1)^{-1}\sum_{i=1}^n(X_i-\bar X_n)^2$$

où $\mathrm{t}_\beta^{(n-1)}$ est le quantile d'ordre β de la loi de Student à $n-1$ degrés de liberté. Considérons le test :

$$\phi_{\mu_0}(Z)=\mathbb{1}_{\{\mu_0\notin I(Z)\}}=\mathbb{1}_{\{|G(Z;\mu_0)|>\mathrm{t}_{1-\alpha/2}^{(n-1)}\}}\ , \tag{I-3.10}$$

où nous avons noté, pour $\nu\in\mathbb{R}$,

$$G(Z;\nu)=\sqrt{n}(\bar X_n-\nu)/S_n\ ,$$

La taille du test $\phi_{\mu_0}(Z)$ est

$$\mathbb{P}_{\mu_0}\left(\mu_o\notin I(Z)\right)=\alpha\ .$$

Nous avons associé à un intervalle de confiance de probabilité de couverture $1-\alpha$, un test de niveau α.

Nous aurions pu procéder en partant de la famille de tests (I-3.10) en définissant cette fois une région de confiance

$$\tilde I(Z):=\left\{\mu_0\in\mathbb{R}:\ \phi_{\mu_0}(Z)=0\right\}\ ,$$

qui donne, pour une observation donnée Z, l'ensemble des valeurs de la moyenne μ_0 pour lesquelles le test (I-3.10) est accepté. Il est élémentaire de voir ici que $I(Z)$ et $\tilde I(Z)$ coïncident.

On peut donc construire des tests d'hypothèses à partir d'un intervalle de confiance ou construire des intervalles de confiance à partir de famille de tests. Nous allons voir dans la suite comment ce procédé se généralise. ♢

Cet exemple est un cas particulier du principe de dualité entre intervalles de confiance et tests que nous présentons maintenant. Soit $(\mathsf{Z},\mathscr{Z},\{\mathbb{P}_\theta\ ,\ \theta\in\Theta\})$ une expérience statistique. Soit $g:\Theta\to\mathbb{R}^q$. Dans l'exemple I-3.25, nous avons simplement $q=1$, $\theta=(\mu,\sigma^2)$ et $g(\theta)=\mu$.

Pour $\alpha\in\,]0,1[$ et $\nu\in\mathbb{R}^q$, soit ϕ_ν une procédure de test de niveau α pour

$$H_0:\ g(\theta)=\nu,\quad \text{contre}\quad H_1:\ g(\theta)\neq\nu\ .$$

Notons

$$\Theta_0^\nu=\{\theta\in\Theta:\ g(\theta)=\nu\}\ . \tag{I-3.11}$$

Puisque le test est de niveau α, pour tout $\theta\in\Theta_0^\nu$, nous avons

$$\mathbb{P}_\theta(\phi_\nu(Z)=1)\le\alpha\ . \tag{I-3.12}$$

On définit la région de confiance *duale* $S(z)\subseteq\mathbb{R}^q$ associée à $z\in\mathsf{Z}$ pour $g(\theta)$ par

$$S(z)=\{\nu\in\mathbb{R}^q:\ \phi_\nu(z)=0\}\ . \tag{I-3.13}$$

Par définition, la probabilité de couverture de la région S pour le paramètre $g(\theta)$ est donnée (en utilisant (I-3.12)) par

$$\begin{aligned}\inf_{\theta\in\Theta}\mathbb{P}_\theta(\{z\in\mathsf{Z},\ g(\theta)\in S(z)\})&=\inf_{\nu\in\mathbb{R}^q}\inf_{\theta\in\Theta_0^\nu}\mathbb{P}_\theta(\phi_\nu(Z)=0)\\&=1-\sup_{\nu\in\mathbb{R}^q}\sup_{\theta\in\Theta_0^\nu}\mathbb{P}_\theta(\phi_\nu(Z)=1)\ge1-\alpha\ .\end{aligned}$$

Réciproquement, supposons maintenant que l'on dispose d'une région de confiance S' de niveau de confiance $1-\alpha$ pour la fonction du paramètre $g(\theta)$, i.e. pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(\{z \in \mathsf{Z},\ g(\theta) \in S(z)\}) \geq 1-\alpha\ .$$

Pour tout $\nu \in \mathbb{R}^q$, considérons le test ϕ'_ν défini pour tout $z \in \mathsf{Z}$ par $\phi'_\nu(z) = \mathbb{1}\{\nu \notin S'(z)\}$.

Pour tout $\nu \in \mathbb{R}^q$, ϕ'_ν est une procédure de test pour l'hypothèse

$$H_0 :\ g(\theta) = \nu, \quad \text{contre} \quad H_1 :\ g(\theta) \neq \nu$$

de niveau α. En effet, pour tout $\theta \in \Theta_0^\nu = \{\theta \in \Theta :\ g(\theta) = \nu\}$

$$\mathbb{P}_\theta\left(\phi'_\nu(Z) = 1\right) = \mathbb{P}_\theta\left(\nu \notin S'(Z)\right) = 1 - \mathbb{P}_\theta\left(g(\theta) \in S'(Z)\right) \leq \alpha\ .$$

I-3.6 Pour aller plus loin : Utilisation d'inégalités de déviation

Dans les exemples vus jusqu'ici de construction d'intervalles de confiance ou de construction de tests à partir d'une fonction pivotale, nous pouvions obtenir explicitement les quantiles de la loi du pivot G. Toutefois, il est en pratique assez rare d'avoir de telles fonctions pivotales à disposition. Il existe alors au moins deux stratégies classiques permettant de contourner cette difficulté. La première, que nous décrivons dans cette section, est d'utiliser des inégalités de déviations pour construire des bornes sur les quantiles inconnus. La seconde, que nous aborderons dans la section II-1.4 consiste à utiliser une approche asymptotique.

Pour fixer les idées, considérons $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta,\ \theta \in \Theta\})$ une expérience statistique. On s'intéresse maintenant à un test portant sur une fonction $g(\theta)$ du paramètre à valeurs réelles. Nous considérons la procédure de test

$$H_0 :\ g(\theta) \leq g_0, \quad \text{contre} \quad H_1 :\ g(\theta) > g_0\ \ ,$$

où $g_0 \in \mathbb{R}$. Ceci correspond à une partition de l'espace des paramètres en deux ensembles disjoints, $\Theta_0 = \{\theta \in \Theta :\ g(\theta) \leq g_0\}$ et $\Theta_1 = \{\theta \in \Theta :\ g(\theta) > g_0\}$. On suppose qu'on dispose d'un estimateur $T : z \mapsto T(z)$ de $g(\theta)$ que nous allons utiliser comme statistique de test. Il est naturel à partir de ces éléments de chercher une région de rejet de la forme suivante,

$$\mathscr{R} = \{z \in \mathsf{Z} :\ T(z) \geq c\}$$

et il s'agit de déterminer la valeur critique c permettant de garantir un niveau α.

Une *inégalité de déviation* est une borne sur les probabilités $\mathbb{P}_\theta\left(T(Z) - g(\theta) > \varepsilon\right)$ de la forme : pour tout $\varepsilon > 0$,

$$\mathbb{P}_\theta\left(T(Z) - g(\theta) > \varepsilon\right) \leq \gamma_\theta(\varepsilon)\ . \tag{I-3.14}$$

En supposant qu'une telle fonction $\gamma_\theta(\varepsilon)$ existe, on a donc en utilisant le fait que $g_0 - g(\theta) \geq 0$ pour $\theta \in \Theta_0$,

$$\begin{aligned}
\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(T(Z) > c\right) &= \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(T(Z) - g(\theta) > c - g(\theta) + g_0 - g_0\right) && \text{(I-3.15)}\\
&\leq \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(T(Z) - g(\theta) > c - g_0\right) && \text{(I-3.16)}\\
&\leq \gamma(c-g_0) := \sup_{\theta\in\Theta_0} \gamma_\theta(c-g_0)\ \ .
\end{aligned}$$

En supposant que $\gamma(\varepsilon) \xrightarrow[\varepsilon\to+\infty]{} 0$, on peut prendre $c_\alpha = g_0 + y_\alpha$, où

$$y_\alpha := \inf\{\varepsilon > 0 :\ \gamma(\varepsilon) \leq \alpha\}$$

et on obtient ainsi, en posant $\mathscr{R} = \{z \in \mathsf{Z} :\ T(z) \geq g_0 + y_\alpha\}$,

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(Z \in \mathscr{R}_\alpha\right) \leq \gamma(y_\alpha) \leq \alpha\ \ . \tag{I-3.17}$$

Le test de région de rejet $\mathscr{R} := \{z \in \mathsf{Z} : T(z) \geq g_0 + y_\alpha\}$ est de niveau α.

Supposons que $Z := (X_1, \ldots, X_n)$ est un n-échantillon d'un modèle statistique

$$(\mathsf{X}, \mathscr{X}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})\ .$$

Les inégalités de déviation que nous utiliserons fréquemment portent sur les estimateurs de la forme

$$T(Z) = \frac{1}{n}\sum_{i=1}^{n} t(X_i)$$

d'un paramètre $g(\theta) = \mathbb{E}_\theta[t(X_1)]$. C'est le cas par exemple de l'inégalité de Bienayme-Tchebychev rappelée au lemme IV-1.2 :

$$\text{pour tout } \varepsilon > 0, \qquad \mathbb{P}_\theta\left(\left|n^{-1}\sum_{i=1}^{n} t(X_i) - g(\theta)\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}_\theta(t(X_1))}{n\varepsilon^2}\ .$$

On déduit de cette inégalité qu'on peut utiliser dans la majoration (I-3.14) la fonction

$$\gamma(\varepsilon) := v^2(\Theta_0)/(n\varepsilon^2)\,, \qquad v^2(\Theta_0) := \sup_{\theta\in\Theta_0} \mathrm{Var}_\theta(t(X_1))\,.$$

Donc, si $v^2(\Theta_0) < \infty$, on peut prendre $y_\alpha := v(\Theta_0)/\sqrt{n\alpha}$ et on en déduit que le test de région de rejet

$$\mathscr{R}_\alpha := \left\{z \in \mathsf{Z} : T(z) \geq g_0 + \frac{v(\Theta_0)}{\sqrt{n\alpha}}\right\} = \left\{T(Z) \geq g_0 + \frac{v(\Theta_0)}{\sqrt{n\alpha}}\right\}\ ,$$

est de niveau α.

Exemple I-3.26. Soit $Z = (X_1, \ldots, X_n)$ un n-échantillon de Bernoulli

$$(\{0,1\}, \mathscr{P}(\{0,1\}), \{\mathrm{Ber}(\theta)\,,\ \theta \in \Theta = [0,1]\})\ .$$

On cherche à construire un test de l'hypothèse

$$H_0 :\ \theta \leq \frac{1}{2}, \quad \text{contre} \quad H_1 :\ \theta > \frac{1}{2}\ .$$

On considère l'estimateur de la proportion θ donné par $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Pour appliquer la construction précédente, il reste à évaluer

$$v^2(\Theta_0) = \sup_{\theta\leq 1/2} \mathrm{Var}_\theta(X_1) = \sup_{\theta\leq 1/2} \mathrm{Var}_\theta(X_1) = \sup_{\theta\leq 1/2} \theta(1-\theta) = \frac{1}{4}\ .$$

On en déduit que le test de région de rejet

$$\mathscr{R} = \left\{(x_1, \ldots, x_n) \in \{0,1\}^n :\ n^{-1}\sum_{i=1}^{n} x_i > \frac{1}{2} + \frac{1}{2\sqrt{n\alpha}}\right\}\ ,$$

est de niveau α. ♢

Lorsque les variables sont bornées, on peut remplacer l'inégalité de Tchebycheff par celle d'Hoeffding (Théorème IV-1.9). Plus précisément, supposons que l'observation $Z = (X_1, \ldots, X_n)$ où $X_1, \ldots, X_n$ sont des variables aléatoires i.i.d. à valeurs dans $(\mathsf{X}, \mathscr{X})$. Supposons qu'il existe une fonction $t : \mathsf{X} \to \mathbb{R}$ mesurable telle que pour tout $\inf_{x\in\mathsf{X}} t(x) = a > -\infty$ et $b = \sup_{x\in\mathsf{X}} t(x) < \infty$ et $g(\theta) = \mathbb{E}_\theta[t(X_1)]$ pour tout $\theta \in \Theta$. L'inégalité de Hoeffding montre que pour tout $\varepsilon > 0$, nous avons

$$\mathbb{P}_\theta\left(n^{-1}\sum_{i=1}^{n} t(X_i) - g(\theta) > \varepsilon\right) \leq \mathrm{e}^{-2n\varepsilon^2/(b-a)^2}\ .$$

Notons que la relation qui précède correspond à une borne de type (I-3.14) dans laquelle $\gamma_\theta(x)$ ne dépend pas de θ (elle dépend par contre de a et de b). On peut donc prendre directement $\gamma(\varepsilon) = \mathrm{e}^{-2n\varepsilon^2/(b-a)^2}$ dans (I-3.15), ce qui mène au choix suivant de y_α dans (I-3.17).

$$y_\alpha = (b-a)\sqrt{\frac{1}{2n}\ln\left(\frac{1}{\alpha}\right)} \ .$$

On en déduit que le test de région de rejet

$$\mathscr{R} := \left\{ (x_1,\dots,x_n) \in \mathsf{X}^n \,:\, n^{-1}\sum_{i=1}^n t(x_i) > g_0 + (b-a)\sqrt{\frac{1}{2n}\ln\left(\frac{1}{\alpha}\right)} \right\} \ ,$$

est de niveau α.

Exemple I-3.27. Reprenons l'exemple de Bernoulli, comme les variables sont à valeurs dans $[0,1]$, on peut appliquer le test basé sur l'inégalité d'Hoeffding, le test de région de rejet

$$\mathscr{R} = \left\{ z \in \{0,1\}^n \,:\, \bar{x}_n > \frac{1}{2} + \sqrt{\frac{1}{2n}\ln\left(\frac{1}{\alpha}\right)} \right\} \ ,$$

est de niveau α. En reprenant les calculs de la section I-3.3, on peut comparer ce test avec celui basé sur l'inégalité de Tchebycheff. Celui qui a la meilleure puissance est celui dont la valeur critique est la plus petite. Il s'agit donc de comparer les valeurs de y_α obtenues par ces deux inégalités, autrement dit, il s'agit de comparer $1/(2\alpha)$ et $\ln(1/\alpha)$. Or, la dérivée de la fonction $u \mapsto \ln(u) - u/2$ est égale à $1/u - 1/2$, cette fonction atteint donc son maximum pour $u = 2$ et ce maximum vaut $\ln(2) - 1 < 0$. On a donc toujours $\ln(u) < u/2$, en particulier $\ln(1/\alpha) < 1/(2\alpha)$, et donc, dans cet exemple, le test basé sur l'inégalité d'Hoeffding est toujours plus puissant que celui basé sur celle de Tchebycheff. Ce n'est pas toujours le cas en général ! ♢

Chapitre I-4

Bases de la théorie de la décision

Dans les chapitres précédents, nous avons défini un estimateur comme une statistique arbitraire. Afin d'être utile, un estimateur doit cependant être "proche", en un certain sens, de la valeur de la fonction à estimer. Il est également naturel de se demander, étant donnés deux estimateurs T_1, T_2, lequel est préférable à l'autre, voire de chercher des estimateurs "optimaux" selon différents critères. L'objectif de ce chapitre est de définir des bornes quantitatives de la performance d'estimateurs.

Les notions principales sont celles de risque d'une règle de décision, de biais d'un estimateur, de décomposition biais-variance, de modèle régulier, d'information de Fisher, de borne de Cramer-Rao et d'estimateur efficace.

I-4.1 Règles de décision, pertes et risques

I-4.1.1 Perte et risque d'une règle de décision

Les notions de *perte* et de *risque* permettent de définir des mesures quantitatives de la performance d'estimateurs. Ces notions se formulent naturellement dans le cadre général de la *théorie de la décision*, qui permet de répondre à ces questions pour des *règles de décision* dont les estimateurs ponctuels sont un exemple.

Définition I-4.1 (Risque d'une règle de décision). *Soit* $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ *un modèle statistique paramétrique et* $\ell : \Theta \times \mathsf{A} \to \mathbb{R}_+$ *une* fonction de perte*, où* A *est un ensemble d'*actions*. On suppose* A *muni d'une tribu* $\mathscr{A}$ *telle que* $\ell(\theta, \cdot) : (\mathsf{A}, \mathscr{A}) \to \mathbb{R}$ *soit mesurable pour tout* $\theta \in \Theta$*.*

Une règle de décision δ *est une fonction mesurable* $\delta : (\mathsf{Z}, \mathscr{Z}) \to (\mathsf{A}, \mathscr{A})$ *qui associe une action* $\delta(Z)$ *à une observation* Z *; le* risque *d'une règle de décision* δ *est défini comme l'espérance*

$$R_\ell(\theta, \delta) := \mathbb{E}_\theta[\ell(\theta, \delta(Z))]\,. \tag{I-4.1}$$

L'estimation ponctuelle peut être naturellement formulée comme un problème de décision. En effet, soit $g : \Theta \to \mathsf{Y}$ une fonction du paramètre à estimer, où $(\mathsf{Y}, \mathscr{Y})$ est un espace mesurable. En prenant pour espace d'actions $\mathsf{A} = \mathsf{Y}$, un estimateur $T : (\mathsf{Z}, \mathscr{Z}) \to (\mathsf{Y}, \mathscr{Y})$ de $g : \theta \mapsto g(\theta)$ n'est rien d'autre qu'une règle de décision au sens de la définition I-4.1. De plus, à une fonction de distance $W : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}_+$ mesurable, on peut naturellement associer la *fonction de perte*, définie pour tout $a \in \mathsf{A}$ et $\theta \in \Theta$:

$$(\theta, a) \mapsto W(g(\theta), a)\,;$$

elle représente la perte réalisée lorsque l'on estime $g(\theta)$ par l'action a. Ainsi, toutes les notions que nous allons définir par la suite pour évaluer les règles de décision s'appliquent en particulier aux estimateurs.

Exemple I-4.2. Il est fréquent de considérer les fonctions de perte suivantes pour l'estimation d'une fonction $g : \Theta \to \mathbb{R}^p$, $\theta \mapsto g(\theta)$,

— la *perte quadratique* $\ell(\theta, a) = \|a - g(\theta)\|_2^2$ associée au carré de la norme euclidienne

$$(u,v) \mapsto \|u-v\|_2^2 = \sum_{i=1}^{p} (u^{(i)} - v^{(i)})^2 \; ; \qquad u = (u^{(1)}, \cdots, u^{(p)}) \in \mathbb{R}^p \; ,$$

— la *perte absolue* $\ell(\theta, a) = \|a - g(\theta)\|_1$ associée à la norme 1

$$(u,v) \mapsto \|u-v\|_1 = \sum_{i=1}^{p} |u^{(i)} - v^{(i)}| \; .$$

♢

Le test d'hypothèse peut également se formuler comme un problème de décision : écrivons $\Theta = \Theta_0 \cup \Theta_1$ où les ensembles Θ_0, Θ_1 - disjoints - décrivent resp. l'hypothèse nulle H_0 et l'hypothèse aletrnative H_1. Mettre en oeuvre un test c'est définir une action à valeur dans $\mathsf{A} := \{0,1\}$. Une fonction de perte adaptée à la situation de "test" est de la forme

$$\ell(\theta,0) = \begin{cases} \ell_0 & \text{si } \theta \in \Theta_1 \\ 0 & \text{sinon.} \end{cases} \qquad \ell(\theta,1) = \begin{cases} \ell_1 & \text{si } \theta \in \Theta_0 \\ 0 & \text{sinon.} \end{cases}$$

pour des réels ℓ_0, ℓ_1 strictement positifs. On exprime ainsi que si l'on décide d'accepter H_0 (on prend la décision "0") alors que $\theta \in \Theta_0$, la perte est nulle ; si on décide de rejeter H_0 (on prend la décision "1") alors que $\theta \in \Theta_0$, la perte est ℓ_1. Le risque d'un test ϕ associé à cette fonction de perte vaut

$$R_\ell(\theta, \phi) = \ell_0 \, \mathbb{P}_\theta(\phi(Z) = 0) \mathbb{1}_{\Theta_1}(\theta) + \ell_1 \, \mathbb{P}_\theta(\phi(Z) = 1) \mathbb{1}_{\Theta_0}(\theta) = \begin{cases} \ell_0 \, \mathbb{P}_\theta(\phi(Z) = 0) & \text{si } \theta \in \Theta_1 \\ \ell_1 \, \mathbb{P}_\theta(\phi(Z) = 1) & \text{si } \theta \in \Theta_0 \end{cases} .$$

I-4.1.2 Admissibilité

La fonction de risque définit un ordre partiel dans l'espace des règles de décision : la règle δ_1 est *préférable* à la règle δ_2 si, pour tout $\theta \in \Theta$,

$$R_\ell(\theta, \delta_1) \leq R_\ell(\theta, \delta_2) \; .$$

Cela suggère la définition suivante :

Définition I-4.3 (Admissibilité). *Une règle de décision $\delta : \mathsf{Z} \to \mathsf{A}$ est dite* inadmissible *(pour la perte ℓ) s'il existe une règle δ^* telle que*

$$R_\ell(\theta, \delta^*) \leq R_\ell(\theta, \delta) \qquad \text{pour tout } \theta \in \Theta$$
$$R_\ell(\theta, \delta^*) < R_\ell(\theta, \delta) \qquad \text{pour au moins un } \theta \in \Theta$$

Une règle qui n'est pas inadmissible *est dite* admissible*. Autrement dit, une règle de décision est admissible si elle est maximale pour l'ordre partiel induit par la fonction de risque.*

La propriété d'admissibilité est un prérequis minimal. En revanche, il existe en général beaucoup de règles incomparables ainsi que des règles admissibles indésirables, comme le montre l'exemple suivant.

Exemple I-4.4 (Estimateurs admissibles indésirables). Considérons un n-échantillon du modèle de Bernoulli,

$$\left(\{0,1\}^n, \mathscr{P}(\{0,1\}^n), \left\{\mathrm{Ber}_\theta^{\otimes n}, \; \theta \in \Theta := [0,1]\right\}\right) .$$

On cherche à estimer le paramètre θ. Soit $\vartheta \in \Theta$; on considère l'estimateur constant $\hat{\theta}_n : Z \mapsto \vartheta$. Déterminons son risque pour la perte quadratique $\ell(\theta, a) = (\theta - a)^2$. Nous avons

$$R_\ell(\theta, \hat{\theta}_n) = (\theta - \vartheta)^2 \; , \theta \in \Theta.$$

Notons que ce risque est nul en $\theta = \vartheta$. Par suite, pour toute règle d'estimation $\delta^\star$, nous avons

$$R_\ell(\vartheta, \delta^\star) \geq R_\ell(\vartheta, \hat{\theta}_n) = 0 \,.$$

Supposons que $\hat{\theta}_n$ soit inadmissible. Alors pour toute règle de décision $\delta^\star$, nous avons $R_\ell(\theta, \delta^\star) \leq R_\ell(\theta, \hat{\theta}_n)$ pour tout θ. Donc en particulier en $\theta = \vartheta$, nous avons $R_\ell(\vartheta, \delta^\star) = 0$. Compte-tenu de la fonction de perte, cela signifie que $\delta^\star(Z) = \vartheta = \hat{\theta}_n$. Par suite, $\hat{\theta}_n$ est admissible. Néanmoins, cet estimateur n'est pas très intéressant, puisqu'il ne sait estimer (parfaitement) qu'un seul paramètre. ♢

I-4.1.3 Optimalité sous contraintes

Une dernière approche de l'optimalité consiste à se restreindre à certaines "classes". Ainsi, il est parfois possible d'obtenir des estimateurs dont le risque est minimal au sein de leur classe, et ce *quelle que soit* la valeur du paramètre θ. Nous étudierons en particulier la classe des estimateurs *sans biais* dans le cas de la perte quadratique dans la section I-4.2.

Une autre classe remarquable est constituée des estimateurs invariants sous l'action de certains groupes de transformations.

Exemple I-4.5 (Estimation équivariante du paramètre de translation). Soit p une densité de probabilité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. Pour $\theta \in \Theta := \mathbb{R}^d$ et $x \in \mathbb{R}^d$, nous notons

$$p_\theta(x) := p(x - \theta) \,.$$

Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle

$$\Big(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d), \{\mathbb{P}_\theta = p_\theta \cdot \lambda_{\text{Leb}}\,,\ \theta \in \Theta\}\Big) \,.$$

Considérons le problème de l'estimation du paramètre de translation θ, et la recherche d'un estimateur optimal au sens du risque associé à la fonction de perte $(u, v) \mapsto w(u - v)$ où $w : \mathbb{R}^d \to \mathbb{R}_+$ est une fonction mesurable positive.

Remarquons que, pour tout $c \in \mathbb{R}^d$ et $\theta \in \Theta$, et toute fonction h mesurable positive,

$$\mathbb{E}_\theta[h(X_j + c)] = \int_{\mathbb{R}^d} h(x + c) p_\theta(x) \mathrm{d}x = \int_{\mathbb{R}^d} h(y) p_\theta(y - c) \mathrm{d}y = \int_{\mathbb{R}^d} h(y) p_{\theta + c}(y) \mathrm{d}y = \mathbb{E}_{\theta + c}[h(X_j)] \,;$$

nous avons utilisé $p_\theta(u) = p(u - \theta) = p_{\theta + c}(u + c)$. Il est donc naturel de se restreindre à la classe d'estimateurs satisfaisant la propriété d'invariance par translation, i.e. pour tout $c \in \mathbb{R}^d$,

$$T(x_1 + c,\ \ldots\ ,\ x_n + c) = T(x_1,\ \ldots\ ,\ x_n) + c, \quad \text{pour tous } x_1, \ldots, x_n \in \mathbb{R}^d. \tag{I-4.2}$$

Les estimateurs qui satisfont (I-4.2) sont dits *équivariants*. Un estimateur T est équivariant si et seulement si, pour tous $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$T(x_1,\ \ldots\ ,\ x_n) = x_1 + T(0, x_2 - x_1,\ \ldots\ ,\ x_n - x_1) \,. \tag{I-4.3}$$

Si T est un estimateur équivariant, alors le risque de T pour une perte qui ne dépendrait que de la différence $(u, v) \mapsto w(u - v)$ ne dépend pas de $\theta \in \Theta$. En effet

$$\begin{aligned} R_w(\theta, T) &= \mathbb{E}_\theta[w(T(X_1, \ldots, X_n) - \theta)] = \mathbb{E}_\theta[w(T(0, X_2 - X_1, \ldots, X_n - X_1) + X_1 - \theta)] \\ &= \mathbb{E}_\theta[w(T(X_1 - \theta, \ldots, X_n - \theta))] \end{aligned}$$

en utilisant l'équivariance de T. Puis, comme on est dans un modèle de translation, nous avons

$$R_w(\theta, T) = \mathbb{E}_0[w(T(X_1, \ldots, X_n))] \,.$$

On en déduit que

$$R_w(\theta, T) = R_w(0, T) \qquad \forall \theta \,.$$

Il suffit donc de prendre un estimateur équivariant T_0 pour lequel $\inf_T R_w(0, T)$ est réalisé ; nous venons de voir qu'il est optimal dans la classe des estimateurs équivariants. ♢

I-4.2 Risque quadratique et estimateurs sans biais

Dans toute cette section, nous considérons le problème de l'estimation d'une fonction vectorielle $g : \Theta \to \mathbb{R}^p$ pour la perte quadratique $\ell(u, v) = \|u - v\|_2^2$ sur $\mathbb{R}^p$. Nous notons simplement $R = R_\ell$ le risque quadratique.

I-4.2.1 Décomposition biais-variance, exemples

Définition I-4.6 (Biais). *Soient* $(\mathrm{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ *une expérience statistique,* $g : \Theta \to \mathbb{R}^p$ *une fonction mesurable et* T *un estimateur de la fonction* $g : \theta \mapsto g(\theta)$ *tel que* $\mathbb{E}_\theta[\|T(Z)\|] < +\infty$ *pour tout* $\theta \in \Theta$. *Le* biais *de l'estimateur* T *est défini par :*

$$\theta \mapsto \mathbb{E}_\theta[T(Z)] - g(\theta)\ . \tag{I-4.4}$$

T *est un estimateur* sans biais *de* $g : \theta \mapsto g(\theta)$ *si pour tout* $\theta \in \Theta$, $\mathbb{E}_\theta[T(Z)] = g(\theta)$.

Exemple I-4.7 (Moyenne empirique). Soit $(X_1, \ldots, X_n)$ un n-échantillon de $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$. Supposons que, pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[|X_1|] < \infty$. Alors, la moyenne empirique

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

est un estimateur sans biais de l'espérance $g : \theta \mapsto g(\theta) := \mathbb{E}_\theta[X_1]$. ♢

L'exemple I-4.8 montre que, pour un n-échantillon d'un modèle paramétrique $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$, la variance empirique doit être correctement normalisée pour être sans biais.

Exemple I-4.8 (Variance empirique). Reprenons le cadre de l'exemple précédent en supposant de plus que $\mathbb{E}_\theta[X_1^2] < \infty$ pour tout $\theta \in \Theta$, et en cherchant cette fois-ci à estimer la variance $g : \theta \mapsto g(\theta) := \mathrm{Var}_\theta(X_1)$. Une idée naturelle est de considérer l'estimateur

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2\ . \tag{I-4.5}$$

Il s'avère que cet estimateur n'est pas sans biais. En effet, pour tout $a \in \mathbb{R}^p$ nous écrivons

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 &= \frac{1}{n}\sum_{i=1}^n [(X_i - a) - (\bar{X}_n - a)]^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - a)^2 - 2(\bar{X}_n - a)\frac{1}{n}\sum_{i=1}^n (X_i - a) + (\bar{X}_n - a)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - a)^2 - 2(\bar{X}_n - a)^2 + (\bar{X}_n - a)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - a)^2 - (\bar{X}_n - a)^2\ .
\end{aligned}$$

Appliquons cette relation avec $a = \mathbb{E}_\theta[X_i]$ de sorte que, en appliquant l'espérance $\mathbb{E}_\theta$ il vient

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

Dans la dernière relation, nous avons utilisé $\mathbb{E}_\theta[\bar{X}_n] = \mathbb{E}_\theta[X_1]$ de sorte que $\mathbb{E}_\theta\left[(\bar{X}_n - \mathbb{E}_\theta[X_1])^2\right] = \mathrm{Var}_\theta(\bar{X}_n)$. Par suite, puisque $\mathrm{Var}_\theta(\bar{X}_n) = n^{-1}\,\mathrm{Var}_\theta(X_1)$, il vient

$$\mathbb{E}_\theta[\hat{\sigma}_n^2] = \mathrm{Var}_\theta(X_1)\left(\frac{n-1}{n}\right)\ ,$$

ce qui implique

$$\mathbb{E}_\theta[\hat{\sigma}_n^2] - \mathrm{Var}_\theta(X_1) = -\frac{1}{n}\,\mathrm{Var}_\theta(X_1)\ .$$ ♢

L'exemple I-4.9 montre qu'il n'existe pas toujours d'estimateurs sans biais ; c'est donc une propriété qui peut être restrictive.

Exemple I-4.9 (Inexistence d'un estimateur sans biais). Cherchons un estimateur de $g : \theta \mapsto 1/\theta$ dans un modèle statistique d'une famille de lois binomiale $\mathrm{Bin}(n,\theta)$ où $\theta \in \Theta :=]0,1]$. Puisque sous $\mathbb{P}_\theta$, l'observation Z est de loi $\mathrm{Bin}(n,\theta)$, le biais d'un estimateur $T(Z)$ de g vérifie

$$\mathbb{E}_\theta[T(Z)] = \sum_{k=0}^{n} \binom{n}{k} T(k)\theta^k(1-\theta)^{n-k} . \tag{I-4.6}$$

Vu comme une fonction de θ, c'est une quantité bornée. Or, $g(\theta)$ n'est pas une fonction bornée (quand $\theta \to 0$, $g(\theta) \to +\infty$). Par suite, il n'est pas possible qu'existe une fonction mesurable T telle que pour tout $\theta \in]0,1[$ on ait

$$\mathbb{E}_\theta[T(Z)] - g(\theta) = 0 .$$

Plus précisément, la relation (I-4.6) montre que une fonction $g : \Theta \to \mathbb{R}$ admet un estimateur sans biais si et seulement si g est un polynôme (en θ) de degré inférieur ou égal à n, et dans ce cas l'estimateur sans biais de g est unique. En effet, d'après l'équation (I-4.6), pour toute fonction $T : \{0,\dots,n\} \to \mathbb{R}$, l'application $\Phi(T) : \theta \mapsto \mathbb{E}_\theta[T(X)]$ est un polynôme de degré inférieur ou égal à n. De plus, Φ est une application linéaire de l'espace des fonctions $\{0,\dots,n\} \to \mathbb{R}$ vers celui des polynômes de degré inférieur ou égal à n. Ces deux espaces vectoriels étant de même dimension finie $n+1$, pour montrer que Φ est bijective, il suffit de montrer qu'elle est injective. Or, si $T \in \ker(\Phi)$, on a pour tout $\theta \in [0,1[$:

$$(1-\theta)^n \sum_{k=0}^{n} \binom{n}{k} T(k) \left(\frac{\theta}{1-\theta}\right)^k = 0 ,$$

donc (en prenant $t = \theta/(1-\theta)$ et en notant que $(1-\theta)^n \neq 0$ pour $\theta \neq 1$) pour tout $t \geq 0$,

$$F(t) := \sum_{k=0}^{n} \binom{n}{k} T(k) t^k = 0$$

d'où, pour $0 \leq k \leq n$, $0 = F^{(k)}(t) = n!/(n-k)!\,T(k)$, et donc $T = 0$. Donc Φ est bijective, ce qui conclut. ♢

Le biais joue un rôle important dans l'étude du risque quadratique, en raison de la *décomposition biais-variance*

Rappelons que par convention, les vecteurs sont des vecteurs-colonne ; et nous notons $u^\top$ le vecteur transposé de u. Si $T(Z)$ est un estimateur de la fonction $g : \theta \mapsto g(\theta)$ à valeur dans $\mathbb{R}^p$, la quantité

$$\mathbb{E}_\theta\left[(T(Z)-g(\theta))\,(T(Z)-g(\theta))^\top\right] \tag{I-4.7}$$

est une matrice de taille $p \times p$ dont l'élément (i,j) est

$$\mathbb{E}_\theta\left[(T_i(Z)-g_i(\theta))\,(T_j(Z)-g_j(\theta))^\top\right] .$$

Dans le cas $p = 1$, cette matrice devient le scalaire

$$\mathbb{E}_\theta\left[(T(Z)-g(\theta))^2\right]$$

et elle mesure l'erreur quadratique (on parle aussi de *risque quadratique*) de l'estimateur $T(Z)$ de la fonction du paramètre $g(\theta)$. Lorsque l'estimateur est sans biais i.e. $g(\theta) = \mathbb{E}_\theta[T(Z)]$ pour tout $\theta \in \Theta$, l'erreur quadratique est la variance de l'estimateur $T(Z)$ (et dans le cas vetoriel $p > 1$, la matrice (I-4.7) est la matrice de variance-covariance de $T(Z)$). Lorsque l'estimateur est biaisé, l'erreur quadratique et la variance de $T(Z)$ (et de façon analogue, la matrice (I-4.7) et la matrice de variance-covariance de $T(Z)$) sont liées par une relation dite *décomposition biais-variance*.

Proposition I-4.10 (Décomposition biais-variance vectorielle). *Soient* $(\mathrm{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ *un modèle statistique,* $g : \Theta \to \mathbb{R}^p$ *une fonction mesurable et* T *un estimateur de la fonction* $g : \theta \mapsto g(\theta)$ *tel que* $\mathbb{E}_\theta[\|T(Z)\|_2^2] < +\infty$ *pour tout* $\theta \in \Theta$.

Pour tout $\theta \in \Theta$*, nous avons*

$$\mathbb{E}_\theta[(T(Z)-g(\theta))\,(T(Z)-g(\theta))^\top] = \mathrm{Var}_\theta(T(Z)) + (\mathbb{E}_\theta[T(Z)-g(\theta)])\,(\mathbb{E}_\theta[T(Z)-g(\theta)])^\top \ ; \tag{I-4.8}$$

$\mathrm{Var}_\theta(T(Z)) \in \mathrm{Mat}_p(\mathbb{R})$ *désigne la matrice de covariance de* $T(Z)$ *sous* $\mathbb{P}_\theta$.

Démonstration. On écrit

$$T(Z) - g(\theta) = T(Z) - \mathbb{E}_\theta[T(Z)] + \mathbb{E}_\theta[T(Z)] - g(\theta)\,;$$

puis

$$\begin{aligned}(T(Z)-g(\theta))\,(T(Z)-g(\theta))^\top &= (T(Z)-\mathbb{E}_\theta[T(Z)])\,(T(Z)-\mathbb{E}_\theta[T(Z)])^\top \\ &+ (\mathbb{E}_\theta[T(Z)]-g(\theta))\,(\mathbb{E}_\theta[T(Z)]-g(\theta))^\top \\ &+ (T(Z)-\mathbb{E}_\theta[T(Z)])\,(\mathbb{E}_\theta[T(Z)]-g(\theta))^\top \\ &+ (\mathbb{E}_\theta[T(Z)]-g(\theta))\,(T(Z)-\mathbb{E}_\theta[T(Z)])^\top\,.\end{aligned}$$

L'équation (I-4.8) s'en déduit en prenant l'espérance sous $\mathbb{P}_\theta$: le premier terme de droite va donner la matrice de variance-covariance de $T(Z)$, notée $\mathrm{Var}_\theta(T(Z))$. L'espérance du second terme est lui-même, et donne les termes de biais. L'espérance des deux derniers termes est nulle. □

En prenant la trace dans l'équation (I-4.8) et en notant que, pour tout $a \in \mathbb{R}^p$, la trace de $aa^\top$ est $\|a\|^2$, nous prouvons le corollaire suivant qui établit une décomposition du risque quadratique de l'estimateur $T(Z)$ de la quantité $g(\theta)$.

Corollaire I-4.11 (Décomposition biais-variance). *Reprenons les hypothèses de la proposition I-4.10. Pour tout $\theta \in \Theta$, on a*

$$\mathbb{E}_\theta\left[\|T(Z)-g(\theta)\|_2^2\right] = \mathbb{E}_\theta[\|T(Z)-\mathbb{E}_\theta[T(Z)]\|_2^2] + \|\mathbb{E}_\theta\left[T(Z)\right]-g(\theta)\|_2^2\,.$$

Dans le cas scalaire ($p=1$), nous obtenons

$$\mathbb{E}_\theta\left[(T(Z)-g(\theta))^2\right] = \mathbb{E}_\theta\left[(T(Z)-\mathbb{E}_\theta[T(Z)])^2\right] + (\mathbb{E}_\theta\left[T(Z)\right]-g(\theta))^2\,,$$

qui décompose le risque quadratique (terme de gauche) en un terme de variance, et un terme de biais au carré.

En utilisant cette décomposition biais-variance, nous montrons dans l'exemple I-4.12 que le meilleur estimateur linéaire de l'espérance, au sens de la minimisation du risque quadratique, est la moyenne empirique.

Exemple I-4.12 (Meilleur estimateur linéaire sans biais de l'espérance). Soit $(X_1,\dots,X_n)$ un n-échantillon de

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathbb{P}_\theta\,,\ \theta\in\Theta\})\,.$$

Supposons que, pour tout $\theta\in\Theta$, $\mathbb{E}_\theta[X_1^2] < \infty$. On cherche à estimer l'espérance $g(\theta) := \mathbb{E}_\theta[X_1]$ en se limitant à la classe des estimateurs de la forme

$$\hat{\mu}_\alpha := \sum_{i=1}^n \alpha_i X_i$$

où $\alpha = (\alpha_1,\dots,\alpha_n) \in \mathbb{R}^n$ tel que $\sum_{i=1}^n \alpha_i = 1$. Notons $\sigma^2(\theta) := \mathrm{Var}_\theta(X_1)$.

Pusque $\sum_{i=1}^n \alpha_i = 1$, $\hat{\mu}_\alpha$ est un estimateur sans biais de $g(\theta)$. Par suite, par la décomposition biais-variance (I-4.9), il vient

$$R(\theta,\hat{\mu}_\alpha) = \mathrm{Var}_\theta(\hat{\mu}_\alpha) = \sigma^2(\theta)\left(\sum_{i=1}^n \alpha_i^2\right)\,. \tag{I-4.9}$$

Ainsi, le risque de $\hat{\mu}_\alpha$ est minimisé *pour tout* $\theta\in\Theta$ par les poids $\alpha\in\mathbb{R}^n$ de somme 1 qui minimisent $\sum_{i=1}^n \alpha_i^2$. Par l'inégalité de Cauchy-Schwarz (ou par stricte convexité de la fonction carré), on a

$$\frac{1}{n}\sum_{i=1}^n \alpha_i^2 \geq \left(\frac{1}{n}\sum_{i=1}^n \alpha_i\right)^2\,,$$

avec égalité si et seulement $\alpha_1 = \cdots = \alpha_n$; puisque la somme vaut 1, le minimum est atteint en $\alpha_1^\star = \cdots = \alpha_n^\star = 1/n$.

Ainsi, la moyenne empirique

$$\hat{\mu}_{\alpha^\star} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est le meilleur estimateur linéaire sans biais de l'espérance, meilleur au sens du risque quadratique. Si de plus $\sigma^2(\theta) > 0$ pour au moins un $\theta \in \Theta$, l'estimateur $\hat{\mu}_\alpha$ est inadmissible pour tout $\alpha \in \mathbb{R}^n$ de somme 1 distinct de $(n^{-1}, \ldots, n^{-1})$. $\diamondsuit$

Dans l'exemple I-4.8, nous avons vu deux estimateurs de la variance dont un était sans biais. L'exemple suivant compare des ceux estimateurs en terme de risque quadratique : il est montré que l'estimateur biaisé est meilleur que le non biaisé au sens de ce critère.

Exemple I-4.13 (Estimation de la variance dans le cas gaussien). Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{ \mathrm{N}(\mu, \sigma^2)\,,\ \theta = (\mu, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}_+^* \right\}\right).$$

Comparons le risque quadratique des estimateurs S_n^2 et $\hat{\sigma}_n^2$ (cf. l'exemple I-4.8) de la variance σ^2

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

Nous avons vu dans l' exemple I-4.8 que S_n^2 est un estimateur sans biais de σ^2. Par le Théorème IV-3.24, la variance de l'estimateur S_n^2 est donnée par

$$\mathrm{Var}_\theta(S_n^2) = \frac{2\sigma^4}{n-1} \ .$$

Par ailleurs, en notant que $\hat{\sigma}_n^2 = \big\{(n-1)/n\big\} S_n^2$ et en utilisant la décomposition biais-variance, il vient pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$,

$$\begin{aligned}
\mathbb{E}_\theta[\hat{\sigma}_n^2] &= \frac{n-1}{n}\sigma^2 \\
\mathrm{Var}_\theta\left(\hat{\sigma}_n^2\right) &= \frac{(n-1)^2}{n^2}\,\mathrm{Var}_\theta\left(S_n^2\right) = \frac{2(n-1)\sigma^4}{n^2} \\
\mathbb{E}_\theta\left[\left(\hat{\sigma}_n^2 - \sigma^2\right)^2\right] &= \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 + \frac{2(n-1)\sigma^4}{n^2} \\
&= \frac{1}{n^2}\sigma^4 + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.
\end{aligned}$$

La différence des risques quadratiques $R(\theta, S_n^2) - R(\theta, \hat{\sigma}_n^2)$ est donc égale à :

$$\mathbb{E}_\theta\left[\left(S_n^2 - \sigma^2\right)^2\right] - \mathbb{E}_\theta\left[\left(\hat{\sigma}_n^2 - \sigma^2\right)^2\right] = \frac{2\sigma^4}{n-1} - \frac{(2n-1)\sigma^4}{n^2} = \frac{(3n-1)\sigma^4}{(n-1)n^2}$$

qui est toujours positif. L'estimateur $\hat{\sigma}_n^2$ est biaisé, mais son risque quadratique est plus faible car le carré du biais de cet estimateur est compensé par une variance plus faible. Pour le risque quadratique, S_n^2 n'est donc pas un estimateur admissible de σ^2. $\diamondsuit$

Dans les deux exemples ci-dessus, nous avons exhibé des estimateurs sans biais dont la variance décroît en n^{-1}. Nous verrons plus tard que cette vitesse de convergence est optimale sous des hypothèses de régularité. L'exemple suivant montre qu'il est toutefois possible pour certains modèles d'obtenir une convergence plus rapide ; nous obtenons ici une vitesse en $1/n^2$.

Exemple I-4.14 (Estimation du support d'une loi uniforme). Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon du modèle

$$\left(\mathbb{R}_+, \mathscr{B}(\mathbb{R}_+), \{\mathrm{Unif}([0,\theta]), \theta \in \Theta := \mathbb{R}_+^*\}\right) \ .$$

Considérons pour estimateur du support la statistique d'ordre

$$X_{n:n} = \max(X_1, \ldots, X_n) \ .$$

• Afin d'étudier son biais, commençons par calculer sa loi. Pour tout $\theta \in \Theta$ et $x \in [0, \theta]$, nous avons

$$\mathbb{P}_\theta(X_{n:n} \le x) = \mathbb{P}_\theta(X_1 \le x, \ldots, X_n \le x) = (x/\theta)^n \ .$$

On en déduit que la densité de la variable $X_{n:n}$ est donnée pour $x \in [0,\theta]$, par

$$p_\theta(x) = \frac{1}{\theta} n \left(\frac{x}{\theta}\right)^{n-1} = n \frac{x^{n-1}}{\theta^n} \ . \tag{I-4.10}$$

• Déterminons son biais : pour tout $\theta > 0$,

$$\mathbb{E}_\theta[X_{n:n}] = \int_0^\theta x\, n \frac{x^{n-1}}{\theta^n} \mathrm{d}x = \frac{n}{n+1} \frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1}\theta \ .$$

Par conséquent, l'estimateur

$$T_n^{(1)}(X_1,\ldots,X_n) = \frac{n+1}{n} X_{n:n} \tag{I-4.11}$$

est un estimateur sans biais du paramètre θ.

• Cet estimateur est toutefois inadmissible pour le risque quadratique. En effet, considérons l'estimateur $T_n^{(2)}(X_1,\ldots,X_n) = a_n X_{n:n}$ pour un scalaire a_n. Calculons le risque quadratique de cet estimateur. Tout d'abord, pour tout $\theta > 0$, nous avons

$$\mathbb{E}_\theta[X_{n:n}^2] = \int_0^\theta x^2\, n \frac{x^{n-1}}{\theta^n} \mathrm{d}x = \frac{n}{n+2} \frac{\theta^{n+2}}{\theta^n} = \frac{n}{n+2}\theta^2 \ .$$

On en déduit que

$$\begin{aligned}\mathbb{E}_\theta[(a_n X_{n:n} - \theta)^2] &= a_n^2 \mathbb{E}_\theta[X_{n:n}^2] - 2a_n\theta \mathbb{E}_\theta[X_{n:n}] + \theta^2 \\ &= \frac{n a_n^2}{n+2}\theta^2 - \frac{2a_n n}{n+1}\theta^2 + \theta^2 = \theta^2 \left\{ \frac{n a_n^2}{n+2} - \frac{2a_n n}{n+1} + 1 \right\} \ .\end{aligned}$$

Quel que soit $\theta > 0$, le risque quadratique atteint son unique minimum en $a_n = (n+2)/(n+1)$, donc $T_n^{(1)}$ est inadmissible. Pour le choix optimal $a_n = (n+2)/(n+1)$, le risque quadratique de $T_n^{(2)}$ vaut

$$R(\theta, T_n^{(2)}) = \mathbb{E}_\theta[(T_n^{(2)}(X_1,\ldots,X_n) - \theta)^2] = \frac{\theta^2}{(n+1)^2} \ . \qquad \diamondsuit$$

I-4.2.2 Optimalité parmi les estimateurs sans biais

Nous nous intéressons au problème général de la recherche d'estimateurs sans biais optimaux T de $g : \theta \mapsto g(\theta) \in \mathbb{R}$. Par la décomposition biais-variance (Corollaire I-4.11), le risque d'un tel estimateur vaut $R(\theta, T) = \mathrm{Var}_\theta(T(Z))$. Cela conduit naturellement à la définition suivante :

Définition I-4.15. *Soient $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ un modèle statistique, $g : \Theta \to \mathbb{R}$, $\theta \mapsto g(\theta)$ une fonction et T, T' des estimateurs sans biais de $g(\theta)$.*

— *On dit que T est* uniformément meilleur que *T' si $\mathrm{Var}_\theta(T(Z)) \le \mathrm{Var}_\theta(T'(Z))$ pour tout $\theta \in \Theta$.*

— *On dit que T est un estimateur E.S.B.V.M. (*Estimateur Sans Biais de Variance Minimale*) s'il est uniformément meilleur que tout estimateur sans biais de g.*

Autrement dit, un estimateur E.S.B.V.M. est optimal au sein de la classe des estimateurs sans biais de $g(\theta)$. Nous allons par la suite établir une borne inférieure sur la variance de n'importe quel estimateur sans biais. En particulier, tout estimateur sans biais atteignant cette borne inférieure est automatiquement E.S.B.V.M..

Tout d'abord, commençons par définir les modèles que nous étudierons. Ces modèles, dits *réguliers*, dépendent de manière régulière du paramètre $\theta \in \Theta$, et satisfont des hypothèses techniques nécessaires aux définitions ultérieures et à la dérivation sous le signe intégral. Soit $F : \Theta \times \mathsf{Z} \to \mathbb{R}$, $(\theta, z) \mapsto F(\theta, z)$ une fonction. Si pour $z \in \mathsf{Z}$, la fonction $\theta \mapsto F(\theta, z)$ est deux fois dérivable par rapport à θ, nous notons :

$$\mathrm{H}_F(\theta, z) := \left[\frac{\partial^2 F}{\partial \theta^{(i)} \partial \theta^{(j)}}(\theta, z) \right]_{1 \le i,j \le d} \ . \tag{I-4.12}$$

Si $A = (a_{ij})_{1\le i,j\le d} \in \mathrm{Mat}_d(\mathbb{R})$, nous notons $\|A\|$ sa *norme de Frobenius*, définie par :

$$\|A\|^2 := \mathrm{Tr}\left(A^T A\right) = \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}^2 \,.$$

Définition I-4.16 (Modèle régulier). *Considérons un modèle paramétrique $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta\in\Theta\})$, où Θ est un ouvert de $\mathbb{R}^d$. Ce modèle est dit* régulier *s'il satisfait les propriétés suivantes :*

(i) Le modèle est dominé par une mesure σ-finie μ sur $(\mathsf{Z},\mathscr{Z})$, de densité p_θ pour tout $\theta\in\Theta$, i.e. $\mathbb{P}_\theta = p_\theta\cdot\mu$. On pose alors, pour tout $\theta\in\Theta$ et $z\in\mathsf{Z}$:

$$\ell(\theta, z) := \log p_\theta(z)\,. \tag{I-4.13}$$

(ii) Pour μ-presque tout $z\in\mathsf{Z}$, la fonction $\theta\mapsto p_\theta(z)$ est deux fois continûment différentiable sur Θ.

(iii) Pour tout $\theta_0\in\Theta$, il existe un voisinage $\mathscr{V}_0$ de θ_0 et deux fonctions mesurables positives g,h : $(\mathsf{Z},\mathscr{Z})\to\mathbb{R}$ telles que

$$\int_{\mathsf{Z}} \{1+h(z)\}\, g(z)\mu(\mathrm{d}z) < \infty \tag{I-4.14}$$

et, pour tout $\theta\in\mathscr{V}_0$ et μ-presque tout $z\in\mathsf{Z}$,

$$\|\nabla\ell(\theta,z)\|^2 \le h(z)\,, \qquad \|\mathrm{H}_\ell(\theta,z)\| \le h(z)\,, \qquad p_\theta(z)\le g(z) \tag{I-4.15}$$

(iv) Pour tout $\theta\in\Theta$, la matrice $\mathbb{E}_\theta[\mathrm{H}_\ell(\theta,Z)]$ (bien définie par l'hypothèse (iii)) est inversible.

Un exemple de modèle non régulier est donné par le modèle d'estimation de support d'une loi uniforme ; considérons la densité $p_\theta(x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^+$ de la loi uniforme sur $[0,\theta]$. Elle est égale à $p_(x) = \mathbb{1}(x\le\theta)$; pour tout $x>0$, la fonction $\theta\mapsto p_\theta(x)$ est discontinue en x.

Lemme I-4.17. *Soit $(\mathsf{Z},\mathscr{Z},\{p_\theta\cdot\mu\,,\ \theta\in\Theta\})$ un modèle statistique régulier. La matrice*

$$\mathbb{I}(\theta) := \mathbb{E}_\theta\left[\nabla\ell(\theta,Z)\,\left(\nabla\ell(\theta,Z)\right)^\top\right]\,,$$

est bien définie et ne dépend pas du choix de la mesure de domination.

Démonstration. Notons que, pour tout $a\in\mathbb{R}^d$, $\left\|aa^\top\right\| = \mathrm{Tr}\left(aa^\top aa^\top\right)^{1/2} = \mathrm{Tr}\left(a^\top aa^\top a\right)^{1/2} = \|a\|^2$. Ainsi, pour tout $\theta_0\in\Theta$, on a

$$\mathbb{E}_{\theta_0}\left[\left\|\nabla\ell(\theta_0,Z)\ \left(\nabla\ell(\theta_0,Z)\right)^\top\right\|\right] = \mathbb{E}_{\theta_0}[\|\nabla\ell(\theta_0,Z)\|^2] \le \int_{\mathsf{Z}} h(z)g(z)\mu(\mathrm{d}z) < \infty$$

où g,h sont comme dans l'hypothèse (iii) de régularité. Ainsi, l'intégrale (I-4.16) définissant $\mathbb{I}(\theta)$ est bien définie.

Pour voir que $\mathbb{I}(\theta)$ ne dépend pas du choix de μ, soit ν une autre mesure de domination, i.e. pour tout $\theta\in\Theta$, $\mathbb{P}_\theta = q_\theta\cdot\nu$. Posons $\lambda = \mu+\nu$. Il existe (voir Théorème A.47) une fonction mesurable positive h telle que

$$p_\theta(z) = h(z)q_\theta(z)\,, \quad \lambda-\text{p.p.}$$

On en déduit que $\nabla(\log p_\theta(z)) = \nabla(\log q_\theta(z))$ $\lambda-$p.p., ce qui conclut la démonstration en substituant dans (I-4.16). □

Définition I-4.18 (Information de Fisher). *L'*information de Fisher *du modèle régulier* $(\mathsf{Z}, \mathscr{Z}, \{p_\theta \cdot \mu\,,\ \theta \in \Theta\})$ *est la fonction* $\mathbb{I} : \theta \mapsto \mathbb{I}(\theta)$ *définie par*

$$\mathbb{I}(\theta) := \mathbb{E}_\theta \left[\nabla \ell(\theta, Z) \left(\nabla \ell(\theta, Z) \right)^\top \right] , \tag{I-4.16}$$

où $\ell(\theta, z) := \log p_\theta(z)$.

Proposition I-4.19. *Soit* $(\mathsf{Z}, \mathscr{Z}, \{p_\theta \cdot \mu\,,\ \theta \in \Theta\})$ *un modèle statistique régulier. Les propriétés suivantes sont vérifiées :*

(i) Pour tout $\theta \in \Theta$*,*

$$\mathbb{E}_\theta \left[\nabla \ell(\theta, Z) \right] = 0 . \tag{I-4.17}$$

(ii) L'information de Fisher admet l'expression alternative suivante :

$$\mathbb{I}(\theta) = -\mathbb{E}_\theta \left[\mathrm{H}_\ell(\theta, Z) \right] . \tag{I-4.18}$$

(iii) Soient deux modèles statistiques réguliers $(\mathsf{Z}_1, \mathscr{Z}_1, \{\mathbb{P}_{1,\theta}, \theta \in \Theta\})$ *et* $(\mathsf{Z}_2, \mathscr{Z}_2, \{\mathbb{P}_{2,\theta}, \theta \in \Theta\})$ *avec pour ensemble de paramètres* Θ*. En notant* $\mathbb{I}_1$ *et* $\mathbb{I}_2$ *les informations de Fisher respectives de ces modèles, ainsi que* $\mathbb{I}_{12}$ *l'information de Fisher du modèle produit :*

$$\left(\mathsf{Z}_1 \times \mathsf{Z}_2, \mathscr{Z}_1 \otimes \mathscr{Z}_2, \left\{ \mathbb{P}_{1,\theta} \otimes \mathbb{P}_{2,\theta}\,,\ \theta \in \Theta \right\} \right) ,$$

il vient pour tout $\theta \in \Theta$ *:*

$$\mathbb{I}_{12}(\theta) = \mathbb{I}_1(\theta) + \mathbb{I}_2(\theta) . \tag{I-4.19}$$

En particulier, si $(X_1, \ldots, X_n)$ *est un n-échantillon du modèle statistique régulier*

$$\mathscr{E} = (\mathsf{X}, \mathscr{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$$

alors, pour tout $\theta \in \Theta$ *et* $n \geq 1$*,*

$$\mathbb{I}_{\mathscr{E}^n}(\theta) = n \mathbb{I}_{\mathscr{E}}(\theta),$$

où $\mathbb{I}_{\mathscr{E}}(\theta)$ *et* $\mathbb{I}_{\mathscr{E}^n}(\theta)$ *sont les informations de Fisher associée aux modèles* $\mathscr{E}$ *et au modèle* $\mathscr{E}^n$ *induit par le n-échantillon.*

Remarque I-4.20. Soit $(\mathsf{Z}, \mathscr{Z}, \{p_\theta \cdot \mu\,,\ \theta \in \Theta\})$ un modèe statistique régulier. Notons que, pour tout $z \in \mathsf{Z}$, la matrice $\nabla \ell(\theta, z) \nabla \ell(\theta, z)^\top$ est symétrique et semi-définie positive. Comme

$$\mathbb{I}(\theta) = \mathbb{E}_\theta [(\nabla \ell(\theta, Z)) \, (\nabla \ell(\theta, Z))^\top] = \int (\nabla \ell(\theta, z)) \, (\nabla \ell(\theta, z))^\top \, \mathbb{P}_\theta(\mathrm{d}z)$$

est symétrique et semi-définie positive pour tout $\theta \in \Theta$. Par l'égalité I-4.18 et Définition I-4.16-(iv), $\mathbb{I}(\theta)$ est de plus inversible, donc définie positive. ♢

Démonstration. Commençons par noter que, pour tout $\theta \in \Theta$ et $z \in \mathsf{Z}$,

$$\nabla \ell(\theta, z) = \frac{1}{p(\theta, z)} \nabla p(\theta, z) ; \tag{I-4.20}$$

en différenciant à nouveau, il vient :

$$\begin{aligned} \mathrm{H}_\ell(\theta, z) &= \frac{1}{p(\theta, z)} \mathrm{H}_p(\theta, z) - \frac{1}{p^2(\theta, z)} \nabla p(\theta, z) \nabla p(\theta, z)^\top \\ &= \frac{1}{p(\theta, z)} \mathrm{H}_p(\theta, z) - \nabla \ell(\theta, z) \nabla \ell(\theta, z)^\top . \end{aligned} \tag{I-4.21}$$

Nous pouvons alors écrire, pour tout $\theta \in \Theta$,

$$\begin{aligned}\mathbb{E}_\theta\big[\nabla\ell(\theta,Z)\big] &= \int_{\mathsf{Z}} \nabla\ell(\theta,z)\, p(\theta,z)\mu(\mathrm{d}z) = \int_{\mathsf{Z}} \nabla\, p(\theta,z)\mu(\mathrm{d}z) \\ &\overset{(*)}{=} \nabla \int_{\mathsf{Z}} p(\theta,z)\mu(\mathrm{d}z) = \nabla 1 = 0\,,\end{aligned}$$

ce qui établit le premier point, sous réserve que l'interversion $(*)$ soit justifiée. De même, on a

$$\begin{aligned}\mathbb{E}_\theta[\mathrm{H}_\ell(\theta,Z)] &= \int_{\mathsf{Z}} \mathrm{H}_\ell(\theta,z) p(\theta,z)\mu(\mathrm{d}z) = \int_{\mathsf{Z}} \mathrm{H}_p(\theta,z)\mu(\mathrm{d}z) - \int_{\mathsf{Z}} \nabla\ell(\theta,z)\,\nabla\ell(\theta,z)^\top p(\theta,z)\mu(\mathrm{d}z) \\ &\overset{(**)}{=} \nabla \int_{\mathsf{Z}} \nabla p(\theta,Z)^\top \mu(\mathrm{d}z) - \mathbb{I}(\theta) = \nabla 0 - \mathbb{I}(\theta) = -\mathbb{I}(\theta)\,,\end{aligned}$$

ce qui établit le second point, sous réserve que l'interversion $(**)$ soit valide.

Il reste à justifier les dérivations sous le signe intégral $(*)$ et $(**)$. Soit $\theta_0 \in \Theta$, $\mathscr{V}_0$ un voisinage de θ_0 dans Θ et g, h des fonctions comme dans l'hypothèse (iii) de régularité. On a pour tout $\theta \in \mathscr{V}_0$,

$$\|\nabla p(\theta,z)\| = \|\nabla\ell(\theta,z)\|\, p(\theta,z) \le (1 + \|\nabla\ell(\theta,z)\|^2) p(\theta,z) \le (1+h(z))g(z)\,,$$

avec $\int_{\mathsf{Z}}(1+h(z))g(z)\mu(\mathrm{d}z) < \infty$ par (I-4.14). La proposition A.42 de dérivation sous le signe intégral (pour chaque coordonnée du gradient selon θ) appliquée à la fonction $(\theta,z) \mapsto p(\theta,z)$ garantit alors que l'interversion $(*)$ est correcte. De même, pour tout $\theta \in \mathscr{V}_0$ et tout $z \in \mathsf{Z}$, l'égalité (I-4.21) implique

$$\begin{aligned}\big\|\mathrm{H}_p(\theta,z)\big\| &\le \|\mathrm{H}_\ell(\theta,z)\|\, p(\theta,z) + \Big\|\nabla\ell(\theta,z)\ (\nabla\ell(\theta,z))^\top\Big\|\, p(\theta,z) \\ &\le \Big(\|\mathrm{H}_\ell(\theta,z)\| + \Big\|\nabla\ell(\theta,z)\ (\nabla\ell(\theta,z))^\top\Big\|\Big)\, p(\theta,z) \\ &\le 2h(z)g(z)\end{aligned}$$

qui est intégrable sous μ. Comme $\mathrm{H}_p(\theta,z) = \nabla(\nabla p(\theta,z)^\top)$, la proposition A.42 appliquée à la fonction $(\theta,z) \mapsto \nabla\ell(\theta,z)$ (coordonnée par coordonnée) montre que l'interversion $(**)$ est valide.

Enfin, le troisième point découle directement du second et du fait que, pour le modèle produit, la log-vraisemblance se décompose en $\ell_{12}(\theta,(z_1,z_2)) = \ell_1(\theta,z_1) + \ell_2(\theta,z_2)$. □

Exemple I-4.21. Soit $(X_1,\dots,X_n)$ un n-échantillon de Bernoulli

$$(\{0,1\}, \mathscr{P}(\{0,1\}), \{\mathrm{Ber}(\theta)\,,\ \theta \in \Theta :=]0,1[\}\,.$$

Par la Proposition I-4.19, son information de Fisher vaut $\mathbb{I}(\theta) = n\mathbb{I}_1(\theta)$, où $\mathbb{I}_1$ désigne l'information de Fisher associée à une observation $X_1 \sim \mathrm{Ber}(\theta)$. Calculons cette quantité. Le modèle de Bernoulli est dominé par la mesure de comptage, de densité $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ pour $\theta \in]0,1[$ et $x \in \{0,1\}$. Ainsi,

$$\begin{aligned}\ell(\theta,x) &= x\log\theta + (1-x)\log(1-\theta) \\ \ell'(\theta,x) &= \frac{x}{\theta} - \frac{1-x}{1-\theta} \\ \ell''(\theta,x) &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}\,,\end{aligned}$$

d'où, par la Proposition I-4.19 :

$$\begin{aligned}\mathbb{I}_1(\theta) &= -\mathbb{E}_\theta[\ell''(\theta,X)] = -\theta\ell''(\theta,1) - (1-\theta)\ell''(\theta,0) \\ &= \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}\,.\end{aligned}$$

L'information de Fisher du n-échantillon de Bernoulli est donc $\mathbb{I}(\theta) = n/(\theta(1-\theta))$. ♢

Nous pouvons maintenant énoncer le principal résultat de cette section, qui est une borne inférieure sur la variance d'estimateurs sans biais dans un modèle régulier. Commençons par une définition technique :

Définition I-4.22 (Estimateur régulier). *Soit $(\mathsf{Z}, \mathscr{Z}, \{p_\theta \cdot \mu\,,\ \theta \in \Theta\})$, où Θ est un ouvert de $\mathbb{R}^d$, un modèle régulier. Un estimateur $T : (\mathsf{Z}, \mathscr{Z}) \to \mathbb{R}$ d'une fonction $g : \Theta \to \mathbb{R}$ est dit* régulier *si :*

(i) $T(Z)$ est de carré intégrable, i.e. pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[\{T(Z)\}^2] < \infty$.

(ii) La fonction $\theta \mapsto \mathbb{E}_\theta[T(Z)]$ est différentiable sur Θ, et pour tout $\theta \in \Theta$

$$\nabla \mathbb{E}_\theta[T(Z)] = \int_{\mathsf{Z}} T(z)\ \nabla p_\theta(z) \mu(\mathrm{d}z)\,. \tag{I-4.22}$$

Remarque I-4.23. Notons que, par la proposition A.42 de dérivation sous le signe intégral, la seconde condition de régularité est assurée dès lors que la condition suivante est vérifiée : pour tout $\theta_0 \in \Theta$, il existe un voisinage $\mathscr{V}_0$ de θ_0 dans Θ et une fonction mesurable positive g telle que

$$\int_{\mathsf{Z}} |T(z)|\, g(z) \mu(\mathrm{d}z) < \infty$$

et il existe un ensemble $\mathsf{Z}_0 \in \mathscr{Z}$ tel que $\mu(\mathsf{Z}_0^c) = 0$ pour tout $z \in \mathsf{Z}_0$ la fonction $\theta \mapsto p(\theta, z)$ est différentiable sur Θ et pour tout $\theta \in \mathscr{V}_0$ et $z \in \mathsf{Z}_0$,

$$\|\nabla p_\theta(z)\| \le g(z)\,. \qquad \diamondsuit$$

Théorème I-4.24 (Cramér-Rao). *Soit $(\mathsf{Z}, \mathscr{Z}, \{p_\theta \cdot \mu\,,\ \theta \in \Theta\})$ un modèle régulier où Θ est un ouvert de $\mathbb{R}^d$; $g : \Theta \to \mathbb{R}$ une fonction dérivable à valeurs réelles ; et T un estimateur sans biais et régulier de $g : \theta \mapsto g(\theta)$.*

Pour tout $\theta \in \Theta$,

$$\mathrm{Var}_\theta(T(Z)) \ge (\nabla g(\theta))^\top\ \mathbb{I}(\theta)^{-1}\ \nabla g(\theta)\,. \tag{I-4.23}$$

Démonstration. $T(Z)$ étant un estimateur sans biais et régulier, il vient :

$$\begin{aligned}\nabla g(\theta) = \nabla \mathbb{E}_\theta T(Z) &= \int T(z)\ \nabla p_\theta(z)\ \mu(\mathrm{d}z) \\ &= \int T(z)\ \left(\nabla \ell(\theta, z)\right)\ p_\theta(z) \mu(\mathrm{d}z) = \mathbb{E}_\theta[T(Z)\ \nabla \ell(\theta, Z)]\,.\end{aligned}$$

Or $\mathbb{E}_\theta[\nabla \ell(\theta, Z)] = 0$ (Proposition I-4.19), donc $\mathbb{E}_\theta[g(\theta)\ \ (\nabla \ell(\theta, Z))] = 0$ et l'égalité précédente peut s'écrire

$$\nabla g(\theta) = \mathbb{E}_\theta[(T(Z) - g(\theta))\, \nabla \ell(\theta, Z)]\,.$$

On a donc, pour tout $u \in \mathbb{R}^d$,

$$\begin{aligned}\{\nabla g(\theta)^T u\}^2 &= \mathbb{E}_\theta[\{T(Z) - g(\theta)\}\ \nabla \ell(\theta, Z)^T u]^2 \\ &\le \mathbb{E}_\theta[\{T(Z) - g(\theta)\}^2] \mathbb{E}_\theta[\{\nabla \ell(\theta, Z)^T u\}^2]\end{aligned}$$

par l'inégalité de Cauchy-Schwarz. Or, pour tout $a \in \mathbb{R}^d$,

$$\{a^T u\}^2 = u^T a a^T u = u^\top (a a^\top) u\,,$$

donc en prenant respectivement $a = \nabla g(\theta)$ et $a = \nabla \ell(\theta, Z)$, l'inégalité précédente devient :

$$\begin{aligned}u^\top \nabla g(\theta) \nabla g(\theta)^\top u &\le \mathbb{E}_\theta[(T(Z) - g(\theta))^2] \mathbb{E}_\theta[u^\top \nabla \ell(\theta, Z) \nabla \ell(\theta, Z)^\top u] \\ &= \mathrm{Var}_\theta[T(Z)]\, u^\top \mathbb{I}(\theta) u\end{aligned}$$

pour tout $u \in \mathbb{R}^d$. En particulier, en prenant $u = \mathbb{I}(\theta)^{-1}\nabla g(\theta)$, l'inégalité précédente devient

$$\left(\nabla g(\theta)^\top \mathbb{I}(\theta)^{-1}\nabla g(\theta)\right)^2 \leq \mathrm{Var}_\theta[T(Z)] \ \left(\nabla g(\theta)^\top \mathbb{I}(\theta)^{-1}\nabla g(\theta)\right) ,$$

ce qui établit l'inégalité (I-4.23) ($\nabla g(\theta)\mathbb{I}(\theta)^{-1}\nabla g(\theta)$ est positif car $\mathbb{I}(\theta)$ est positive ; si ce terme est nul l'inégalité (I-4.23) est triviale, sinon on simplifie ce terme strictement positif). □

Une conséquence importante du théorème I-4.24 est la suivante :

Corollaire I-4.25. *Soit $(\mathsf{Z}, \mathscr{Z}, \{p_\theta \cdot \mu\,,\ \theta \in \Theta\})$ un modèle régulier, où Θ est un ouvert de $\mathbb{R}^d$, et $T(Z)$ un estimateur sans biais régulier de θ. Alors, pour tout $\theta \in \Theta$,*

$$\mathrm{Var}_\theta(T(Z)) \geq \mathbb{I}(\theta)^{-1} \tag{I-4.24}$$

i.e. la matrice symétrique $\mathrm{Var}_\theta(T(Z)) - \mathbb{I}(\theta)^{-1}$ est positive.

Démonstration. Pour tout $\lambda \in \mathbb{R}^d$, considérons la fonction $g(\theta) = \lambda^T\theta$. Clairement, $\lambda^T T(Z)$ est un estimateur sans biais régulier de $g : \theta \mapsto \lambda^T\theta$. Ainsi, en appliquant le théorème I-4.24 et en notant que $\nabla g(\theta) = \lambda$ et $\mathrm{Var}_\theta(\lambda^T T(Z)) = \lambda^\top \mathrm{Var}_\theta(T(Z))\lambda$, il vient :

$$\lambda^\top \mathrm{Var}_\theta(T(Z))\lambda \geq \lambda^\top \mathbb{I}(\theta)^{-1}\lambda .$$

Cette inégalité étant valide pour tout $\lambda \in \mathbb{R}^d$, il vient $\mathrm{Var}_\theta(T(Z)) \geq \mathbb{I}(\theta)^{-1}$ comme voulu. □

Du théorème I-4.24 de Cramér-Rao et de l'additivité de l'information de Fisher (Proposition I-4.19), on déduit immédiatement le corollaire suivant :

Corollaire I-4.26. *Soient $(X_1, \ldots, X_n)$ un n-échantillon du modèle régulier*

$$(\mathsf{X}, \mathscr{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$$

et $T_n(X_1, \ldots, X_n)$ un estimateur sans biais et régulier de la fonction $g : \Theta \to \mathbb{R}$. En notant

$$\mathbb{I}_1(\theta) = \mathbb{E}_\theta\left[\nabla\ell(\theta, X_1)\ (\nabla\ell(\theta, X_1))^\top\right] , \quad \ell(\theta, x) := \log q_\theta(x) ,$$

l'information de Fisher associée à une seule observation, on a pour tout $\theta \in \Theta$:

$$\mathrm{Var}_\theta(T_n(X_1, \ldots, X_n)) \geq \frac{1}{n}\nabla g(\theta)^\top \mathbb{I}_1(\theta)^{-1}\nabla g(\theta) . \tag{I-4.25}$$

Avant d'étudier quelques exemples, il peut être utile d'interpréter la borne de Cramér-Rao. Cette borne inférieure est d'autant plus petite que l'information de Fisher du modèle en θ est élevée. Intuitivement, l'information de Fisher $\mathbb{I}(\theta_0)$ quantifie la mesure dans laquelle les lois $\mathbb{P}_\theta$ varient autour de θ_0 : plus cette information est grande, plus $\mathbb{P}_\theta$ varie autour de θ_0, et donc plus il est possible d'estimer précisément θ_0 à partir de réalisations de $\mathbb{P}_{\theta_0}$.

De plus, le corollaire I-4.26 montre que, pour un modèle régulier, la variance d'un estimateur sans biais (et donc son risque quadratique) est au moins d'ordre n^{-1}.

Exemple I-4.27. La variance des estimateurs sans biais $\bar{X}_n$ de l'espérance $\mu(\theta) = \mathbb{E}_\theta[Z]$ (exemple I-4.7, en supposant Z de carré intégrable) est d'ordre n^{-1}, tout comme celle de l'estimateur sans biais S_n^2 de la variance $\sigma^2(\theta) = \mathrm{Var}_\theta[Z]$ (exemple I-4.13, en supposant le modèle gaussien, ou plus généralement que Z admet un moment d'ordre 4 sous $\mathbb{P}_\theta$).

Ces résultats, valables pour des modèles très généraux et en particulier pour des modèles réguliers, sont compatibles avec la borne de Cramér-Rao d'ordre n^{-1} (et montrent que cette borne inférieure est optimale dans sa dépendance asymptotique en n).

En revanche, dans le cas de l'estimation du support d'une loi uniforme (exemple I-4.14), l'estimateur sans biais $(n+1)/nX_{n:n}$ de θ a pour variance $\theta^2/\{n(n+2)\}$, qui décroît en n^{-2}. Cela ne contredit pas le théorème de Cramér-Rao, puisque le modèle n'est pas régulier (voir le commentaire qui suit la définition I-4.16). ♢

Nous venons de discuter de la dépendance en la taille n de l'échantillon de la borne de Cramér-Rao. Nous allons voir que la constante $\nabla g(\theta)^\top \mathbb{I}_1(\theta)^{-1} \nabla g(\theta)$ elle-même est optimale, au sens où elle ne peut pas être améliorée en général : il existe des estimateurs sans biais dont la variance atteint la borne inférieure de Cramér-Rao.

Exemple I-4.28 (Estimation du paramètre d'une loi de Bernoulli). Considérons le problème de l'estimation du paramètre $\theta \in]0,1[$ du modèle du n-échantillon de Bernoulli. Nous avons vu dans l'exemple I-4.21 que l'information de Fisher pour ce modèle vaut $\mathbb{I}(\theta) = n/(\theta(1-\theta))$. La borne de Cramér-Rao pour l'estimation sans biais de θ est donc de $\theta(1-\theta)/n$. Cette borne est atteinte par l'estimateur sans biais $\bar{X}_n$ (la fréquence empirique), car

$$\mathrm{Var}_\theta(\bar{X}_n) = \frac{\mathrm{Var}_\theta(X_1)}{n} = \frac{\theta(1-\theta)}{n} \; ; \qquad \diamondsuit$$

cet estimateur est par conséquent E.S.B.V.M.

Exemple I-4.29 (Estimation de la moyenne d'un échantillon gaussien). Soit $(X_1,\dots,X_n)$ un n-échantillon d'un modèle gaussien

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{ \mathrm{N}(\theta, \sigma_0^2) \, , \; \theta \in \mathbb{R} \right\}\right) \, ,$$

où la variance $\sigma_0 > 0$ est connue. Nous avons

$$\ell(\theta, X_i) = -\frac{(X_i-\theta)^2}{2\sigma_0^2} - \log\sqrt{2\pi\sigma_0^2} \, ,$$

donc $\ell''(\theta, X_i) = -1/\sigma_0^2$ et $\mathbb{I}(\theta_0) = n/\sigma_0^2$. Considérons l'estimateur sans biais $\bar{X}_n$ de θ ; sa variance vaut

$$\mathrm{Var}_\theta(\bar{X}_n) = \frac{\sigma_0^2}{n} \, .$$

Par conséquent $\mathrm{Var}_\theta(\bar{X}_n) = \mathbb{I}(\theta)^{-1}$, l'estimateur atteint la borne de Cramér-Rao. ♢

Définition I-4.30 (Efficacité). *Soit $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta \, , \; \theta \in \Theta\})$ un modèle régulier, et $g : \Theta \to \mathbb{R}$ une fonction dérivable. Un estimateur sans biais T de $g(\theta)$ est dit* efficace *si sa variance atteint la borne de Cramér-Rao, i.e. si*

$$\mathrm{Var}_\theta(T(Z)) = (\nabla g(\theta))^\top \; \mathbb{I}(\theta)^{-1} \; \nabla g(\theta) \quad \textit{pour tout } \theta \in \Theta \, . \tag{I-4.26}$$

Notons qu'un estimateur sans biais efficace est E.S.B.V.M. (au sens de la définition I-4.15), bien que la réciproque soit fausse en général.

Chapitre I-5

Optimalité en théorie des tests

Dans ce chapitre, nous nous posons la question de l'existence de tests optimaux, l'optimalité étant définie comme un test *uniformément plus puissant* (U. P. P.). Nous commençons par introduire les *tests randomisés* pour montrer que l'on peut construire un test (non trivial) de taille donnée. Dans le cas d'un test d'hypothèses simples, nous démontrons le théorème de Neyman-Pearson qui construit un test U. P. P. ; dans le cas d'un test sur un paramètre scalaire, nous présentons la méthode du *rapport de vraisemblance monotone* pour étendre ce théorème au cas d'un test avec hypothèses unilatérales. Enfin, nous montrons sur un exemple simple qu'il n'existe pas nécessairement de tests U. P. P. lorsque une hypothèse est bilatérale.

Les notions principales sont celles de tests randomisés, de test Uniformément Plus Puissant (U. P. P.), le théorème de Neyman-Pearson pour la construction d'un test UPP pour des hypothèses simples, et la méthode du rapport de vraisemblance monotone.

I-5.1 Tests uniformément plus puissants

On considère un modèle statistique $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$. On note α et β deux réels de $(0, 1/2)$, et Θ_0 et Θ_1 une partition de Θ. On s'intéresse au test d'hypothèses

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

Dans ce chapitre nous allons nous placer dans la classe des tests de niveau $\alpha \in [0,1]$, c'est-à-dire les tests ϕ tels que

$$\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha\ .$$

Nous cherchons à déterminer (lorsqu'il existe) le test uniformément le plus puissant de niveau α, à savoir le test ϕ de niveau α (i.e. $\sup_{\theta\in\Theta_0} \beta_\phi(\theta) \leq \alpha$) tel que, pour tout test ψ de niveau α,

$$\beta_\phi(\theta) \geq \beta_\psi(\theta)\ , \quad \text{pour tout } \theta \in \Theta_1\ .$$

Nous allons présenter dans cette section deux cadres dans lesquels il est possible de montrer qu'un test uniformément plus puissant existe. Dans les deux cas, une hypothèse clé pour obtenir la propriété d'optimalité voulue est de pouvoir construire un test de *taille* α et pas seulement de niveau α. Ceci n'est pas toujours possible lorsque la variable observée est discrète et il est alors commode d'introduire la notion de tests randomisés pour pallier ce problème.

I-5.1.1 Tests randomisés

Soit $(\mathsf{Z}, \mathscr{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ un modèle statistique et $\phi : \mathsf{Z} \to [0,1]$ une fonction mesurable. Considérons le test d'hypothèses

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

Nous pouvons construire à partir de la fonction ϕ une procédure de test de la façon suivante. Étant donnée une observation $Z \in \mathsf{Z}$, nous simulons une variable de Bernoulli de paramètre $\phi(Z)$. Nous choisissons l'hypothèse nulle H_0 si le résultat de ce tirage est égal à 0 et la rejettons dans le cas contraire. Une telle procédure est appelée *randomisée* car la décision que nous prenons dépend du résultat d'une expérience aléatoire (ici la simulation d'une variable de Bernoulli). La fonction ϕ est appelée la *fonction critique du test*. Alors qu'un test pur réalise une partition de Z en une région de rejet et une région d'acceptation, un test randomisé fait une partition en 3 ensembles :

— L'ensemble $\mathscr{Z}_1 = \{z \in \mathsf{Z} :\ \phi(z) = 1\}$ est la région dans laquelle le test rejette H_0.

— L'ensemble $\mathscr{Z}_0 = \{z \in \mathsf{Z} :\ \phi(z) = 0\}$ est la région dans laquelle le test accepte H_0.

— L'ensemble $\mathscr{Z}_p = \{z \in \mathsf{Z} :\ 0 < \phi(z) < 1\}$ est la région dans laquelle l'issue du test n'est pas entièrement déterminée par les observations. On prend notre décision en simulant une variable de Bernoulli de paramètre $\phi(Z)$.

De toute évidence, les tests randomisés généralisent les tests purs : si la fonction critique ϕ d'un test randomisé est à valeur dans $\{0,1\}$, le test est pur. À l'instar d'un test pur (voir définition I-3.5), nous pouvons spécifier la puissance d'un test randomisé.

Définition I-5.1 (Fonction puissance d'un test randomisé). *La* puissance *d'un test randomisé de fonction critique ϕ est définie par*

$$\beta_\phi : \theta \in \Theta \mapsto \mathbb{E}_\theta[\phi(Z)] \in [0,1]\ .$$

Nous pouvons de même étendre directement les notions de taille et de niveau.

Définition I-5.2 (Taille et niveau d'un test). *La taille d'un test de fonction critique ϕ est donnée*

$$\bar{\alpha}(\phi) = \sup_{\theta \in \Theta_0} \beta_\phi(\theta)\ .$$

Un test de fonction critique ϕ est dit de niveau $\alpha \in [0,1]$ si $\bar{\alpha}(\phi) \leq \alpha$. On note $\mathscr{K}_\alpha(\Theta_0)$ l'ensemble des tests de niveau α de l'hypothèse nulle $H_0 : \theta \in \Theta_0$.

On peut remarquer que ces définitions généralisent de manière naturelle celles données pour les tests purs. Néanmoins, contrairement au cas des tests purs, il est toujours possible de construire un test randomisé de taille α : le test randomisé $\phi_\star$ défini par $\phi_\star(z) = \alpha$ pour toute valeur de $z \in \mathsf{Z}$ est un exemple (ce test accepte ou rejette H_0 indépendamment des observations Z ; il a très peu d'intérêts).

L'exemple I-5.3 explique comment modifier un test dont la taille est strictement inférieure à un niveau α donné, pour en faire un test de taille égale au niveau.

Exemple I-5.3. Soit $Z = (X_1, \ldots, X_n)$ un n-échantillon du modèle de Poisson

$$\left(\mathbb{N}, \mathscr{P}(\mathbb{N}), \left\{p_\theta \cdot \mu\,,\ \theta \in \Theta \subseteq \mathbb{R}_+^\star\right\}\right)\ ,$$

où μ est la mesure de comptage sur $\mathbb{N}$ (voir Appendice A.3.4) et

$$p_\theta(x) = \mathrm{e}^{-\theta}\,\frac{\theta^x}{x!}\ , \quad \theta \in \Theta\ , \quad x \in \mathbb{N}\ .$$

Fixons $\theta_0 < \theta_1$, réels strictement positifs. Nous considérons le test d'hypothèses simples

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1 .$$

Afin de choisir une statistique de test, rappelons quel serait l'estimateur du maximum de vraisemblance pour ce modèle. La log-vraisemblance des observations est donnée par

$$\theta \mapsto \ell_n(\theta; Z) = -n\theta + \log(\theta) S_n - \sum_{i=1}^{n} \log(X_i!) , \qquad S_n := \sum_{i=1}^{n} X_i .$$

L'estimateur du maximum de vraisemblance du paramètre θ est donné par

$$\hat{\theta}_n := n^{-1} S_n .$$

Rappelons que sous $\mathbb{P}_\theta$, puisque les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et distribuées suivant des lois de Poisson de paramètre θ, alors la variable S_n est distribuée suivant une loi de Poisson de paramètre $n\theta$.

• Nous considérons le test pur

$$\phi(Z; c) = \mathbb{1}_{\{S_n > c\}} .$$

Nous obtenons un test de niveau $\alpha \in]0,1[$ en choisissant le seuil critique c_α par

$$c_\alpha := \inf\left\{ c \in \mathbb{N} : \mathbb{P}_{\theta_0}(S_n > c) \leq \alpha \right\} .$$

Comme la variable S_n est à valeurs entière, $\mathbb{P}_{\theta_0}(S_n > c_\alpha) < \alpha$ et $\mathbb{P}_{\theta_0}(S_n > c_\alpha - 1) > \alpha$. La taille du test est strictement inférieure à son niveau. Pour donner un exemple, prenons $\theta_0 = 0.01$ et $n = 100$. La variable aléatoire S_n est donc distribuée suivant une loi de Poisson de paramètre $n\theta_0 = 1$. Nous avons $\mathbb{P}_{0,01}(S_n > 2) \simeq 0.08$, $\mathbb{P}_{0,01}(S_n > 3) \simeq 0.019$, donc, si on veut garantir un niveau inférieur à α, on doit prendre $c_\alpha = 3$. Le niveau de ce test est $\alpha = 0.05$ mais la taille du test est 0.019.

• Nous allons maintenant construire un test randomisé dont la taille est exactement égale à α. Pour $\gamma \in [0,1]$, considérons la fonction critique $\phi^\star_\gamma$ définie de la façon suivante :

$$\phi^\star_\gamma(s) = \begin{cases} 1 & \text{si } s > c_\alpha \\ \gamma & \text{si } s = c_\alpha \\ 0 & s \leq c_\alpha - 1 \end{cases}$$

La taille du test de fonction critique $\phi^\star_\gamma$ est donnée par

$$\tilde{\alpha}(\phi^\star_\gamma) = \mathbb{E}_{\theta_0}[\phi_\gamma(S_n)] = \mathbb{P}_{\theta_0}(S_n > c_\alpha) + \gamma \mathbb{P}_{\theta_0}(S_n = c_\alpha)$$

Si nous choisissons γ de façon à ce que cette taille soit égale à α, nous devons prendre

$$\gamma = \frac{\alpha - \mathbb{P}_{\theta_0}(S_n > c_\alpha)}{\mathbb{P}_{\theta_0}(S_n = c_\alpha)} .$$

Observons que par définition de c_α, nous avons $\alpha < \mathbb{P}_{\theta_0}(S_n > c_\alpha - 1)$ et $\alpha \geq \mathbb{P}_{\theta_0}(S_n > c_\alpha)$ ce qui entraine que

$$0 \leq \gamma < \frac{\mathbb{P}_{\theta_0}(S_n > c_\alpha - 1) - \mathbb{P}_{\theta_0}(S_n > c_\alpha)}{\mathbb{P}_{\theta_0}(S_n = c_\alpha)} \leq 1 .$$

Pour $\alpha = 0.05$, cela revient à choisir $\gamma \approx 0.5$.

• Revenons sur le choix de la statistique de test. Observons que le rapport de vraisemblance $p^{\otimes n}_{\theta_1}(x_1, \cdots, x_n) / p^{\otimes n}_{\theta_0}(x_1, \cdots, x_n)$ de la loi de l'observation sous $\mathbb{P}_{\theta_1}$ et de celle sous $\mathbb{P}_{\theta_0}$ est donné par

$$\left(\frac{\theta_1}{\theta_0}\right)^{\sum_{i=1}^n x_i} \exp\left(-n(\theta_1 - \theta_0)\right)$$

et que pour tout seuil $c > 0$, puisque $\theta_1 > \theta_0$,

$$\{z : p^{\otimes n}_{\theta_1}(z) > c\, p^{\otimes n}_{\theta_0}(z)\} = \{z : \sum_{i=1}^{n} x_i > \frac{\ln c + n(\theta_1 - \theta_0)}{\ln \theta_1 - \ln \theta_0}\} .$$

Ce rapport de vraisemblance est de la forme $\psi(T(z); \theta_1) / \psi(T(z); \theta_0)$ avec $T(z) = \sum_{i=1}^n x_i$; les ensembles de niveau de ce rapport sont les ensembles de niveau de la statistique $T(z)$; et il existe des valeurs de c telles que $\{z : p^{\otimes n}_{\theta_1}(z) = c\, p^{\otimes n}_{\theta_0}(z)\}$ est de probabilité positive sous $\mathbb{P}_{\theta_0}$. Cet exemple illustre des notions que nous allons maintenant détailler : la construction d'un test "uniformément plus puissant" de type test randomisé pour tester deux hypothèses simples, et la méthode du rapport de vraisemblance monotone pour l'obtention de tests "uniformément plus puissants" dans le cas d'un test unilatéral. ♢

I-5.1.2 Le Théorème de Neyman-Pearson

Etant données un test H_0, H_1 et un niveau α, il n'y a pas unicité de la fonction critique ϕ garantissant un niveau α et une taille α. On s'intéresse donc à une optimalité définie ici au sens de "uniformément plus puissant". Le théorème de Neyman-Pearson montre l'existence et exhibe une construction explicite d'un test optimum, dans le cas de deux hypothèses simples. Son extension à d'autres hypothèses est discutée en Section I-5.2.

Définition I-5.4 (Test Uniformément Plus Puissant (U.P.P.)). *Un test de fonction critique ϕ est dit* uniformément plus puissant au niveau α *(*U.P.P.(α)*) s'il est de niveau α, i.e. $\phi \in \mathscr{K}_\alpha(\Theta_0)$ et si sa fonction puissance β_ϕ vérifie :*

$$\beta_\phi(\theta) = \sup_{\psi \in \mathscr{K}_\alpha(\Theta_0)} \beta_\psi(\theta)\,, \qquad \forall \theta \in \Theta_1 \ .$$

Théorème I-5.5 (Neyman-Pearson). *Soit ν une mesure σ-finie sur un espace mesurable $(\mathsf{Z}, \mathscr{Z})$. Considérons le modèle*

$$(\mathsf{Z}, \mathscr{Z}, \{p_\theta \cdot \nu\,,\ \theta \in \Theta := \{\theta_0, \theta_1\}\})$$

où $\theta_0 \neq \theta_1$.

1. *Pour tout $\alpha \in]0,1[$, il existe des constantes $c_\alpha > 0$ et $\gamma_\alpha \in [0,1]$, telles que la fonction critique :*

$$\phi^\star(z) := \begin{cases} 1 & \text{si } p_{\theta_1}(z) > c_\alpha p_{\theta_0}(z)\,, \\ \gamma_\alpha & \text{si } p_{\theta_1}(z) = c_\alpha p_{\theta_0}(z)\,, \\ 0 & \text{si } p_{\theta_1}(z) < c_\alpha p_{\theta_0}(z)\,, \end{cases} \tag{I-5.1}$$

   vérifie

$$\mathbb{E}_{\theta_0}[\phi^\star(Z)] = \int \phi^\star(z) p_{\theta_0}(z) \nu(\mathrm{d}z) = \alpha\,.$$

2. *Le test de fonction critique $\phi^\star$ est* U.P.P.(α) *pour le test d'hypothèses*

$$H_0 :\ \theta = \theta_0, \quad \textit{contre} \quad H_1 :\ \theta = \theta_1\,,$$

   et sa puissance est supérieure ou égale à α :

$$\mathbb{E}_{\theta_1}[\phi^\star(Z)] = \int \phi^\star(z) p_{\theta_1}(z) \nu(\mathrm{d}z) \geq \alpha\,.$$

3. *De plus, si $\phi^{\star\star}$ est* U.P.P.(α) *alors, pour ν-presque tout $z \in \mathsf{Z}$,*

$$\phi^{\star\star}(z) = \begin{cases} 1 & \text{si } p_{\theta_1}(z) > c_\alpha p_{\theta_0}(z)\,, \\ 0 & \text{si } p_{\theta_1}(z) < c_\alpha p_{\theta_0}(z)\,. \end{cases} \tag{I-5.2}$$

Remarquons que si $\mathbb{P}_{\theta_0}(\{z : p_{\theta_1}(z) = c p_{\theta_0}(z)\}) = 0$ pour tout $c \geq 0$, on peut choisir $\gamma = 0$ dans (I-5.3) et donc obtenir un test U.P.P.(α) non randomisé. L'exemple I-5.3 et exemple I-5.9 sont des situations où ce n'est pas le cas $(\gamma > 0)$; l'exemple I-5.6, l'exemple I-5.7 et l'exemple I-5.8 sont des situations où c'est

le cas ($\gamma = 0$). En observant que

$$\begin{aligned}\mathbb{P}_{\theta_0}(p_{\theta_1}(Z) = cp_{\theta_0}(Z)) &= \mathbb{P}_{\theta_0}\left(p_{\theta_1}(Z) = cp_{\theta_0}(Z), p_{\theta_0}(Z) > 0\right) \\ &= \mathbb{P}_{\theta_0}\left(\frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} = c, p_{\theta_0}(Z) > 0\right) = \mathbb{P}_{\theta_0}\left(\frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} = c\right) ,\end{aligned}$$

cette propriété se formule aussi en disant que sous $\mathbb{P}_{\theta_0}$, le *rapport de vraisemblance* n'a pas d'atomes.

Démonstration. Pour tout $c \geq 0$ et $\gamma \in [0,1]$, considérons le test randomisé

$$\phi_{c,\gamma}(z) := \begin{cases} 1 & \text{si } p_{\theta_1}(z) > cp_{\theta_0}(z) , \\ \gamma & \text{si } p_{\theta_1}(z) = cp_{\theta_0}(z) , \\ 0 & \text{si } p_{\theta_1}(z) < cp_{\theta_0}(z) . \end{cases} \tag{I-5.3}$$

Nous allons tout d'abord montrer que nous pouvons toujours choisir les constantes c et γ de telle sorte que ce test soit de taille α :

$$\mathbb{E}_{\theta_0}[\phi_{c,\gamma}(Z)] = \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > cp_{\theta_0}(Z)) + \gamma\mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > cp_{\theta_0}(Z)) = \alpha . \tag{I-5.4}$$

• La fonction $\bar{F} : \mathbb{R} \to [0,1]$,

$$c \mapsto \bar{F}(c) := \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > cp_{\theta_0}(Z)) = 1 - \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) \leq cp_{\theta_0}(Z))$$

est décroissante sur $\mathbb{R}$, continue à droite et admet des limites à gauche, i.e. en tout point $c_0 \in \mathbb{R}$:

$$\begin{aligned}\bar{F}(c_0) &= \lim_{c \downarrow c_0} \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > cp_{\theta_0}(Z)) = \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_0 p_{\theta_0}(Z)) \\ \bar{F}_r(c_0-) &= \lim_{c \uparrow c_0} \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > cp_{\theta_0}(Z)) = \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) \geq c_0 p_{\theta_0}(Z)) .\end{aligned}$$

Notons que la taille du saut de discontinuité de $\bar{F}$ en c_0 est donnée par

$$\bar{F}(c_0-) - \bar{F}(c_0) = \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) = c_0 p_{\theta_0}(Z)) .$$

De plus, $\bar{F}(0-) = 0$ et $\lim_{c\to\infty} \bar{F}(c) = 0$.

• Prouvons (I-5.4). Soit $\alpha \in]0,1[$. Vues les propriétés de $\bar{F}$, il existe $c_\alpha \geq 0$ tel que

$$\bar{F}(c_\alpha) = \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_\alpha p_{\theta_0}(Z)) \leq \alpha \leq \bar{F}(c_\alpha-) = \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) \geq c_\alpha p_{\theta_0}(Z)) . \tag{I-5.5}$$

Nous posons

$$\gamma_\alpha := \begin{cases} 0 & \text{si } \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_\alpha p_{\theta_0}(Z)) = \alpha \\ \frac{\alpha - \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_\alpha p_{\theta_0}(Z))}{\mathbb{P}_{\theta_0}(p_{\theta_1}(Z) = c_\alpha p_{\theta_0}(Z))} & \text{si } \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_\alpha p_{\theta_0}(Z)) < \alpha . \end{cases}$$

Notons que si $\mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_\alpha p_{\theta_0}(Z)) < \alpha$ nous avons par (I-5.5)

$$\begin{aligned}\mathbb{P}_{\theta_0}(p_{\theta_1}(Z) = c_\alpha p_{\theta_0}(Z)) &= \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) \geq c_\alpha p_{\theta_0}(Z)) - \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_\alpha p_{\theta_0}(Z)) \\ &\geq \alpha - \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) > c_\alpha p_{\theta_0}(Z)) > 0 .\end{aligned}$$

γ_α est donc toujours bien défini, et il vérifie $\gamma_\alpha \in [0,1]$. Sur la fig. I-5.1, nous traçons deux exemples de fonction de répartition $F : c \mapsto \mathbb{P}_{\theta_0}(p_{\theta_1}(Z) \leq c\ p_{\theta_0}(Z))$. En remarquant que trouver (c,γ) vérifiant (I-5.4) est équivalent à trouver (c,γ) vérifiant

$$1 - \alpha = (1-\gamma)\mathbb{P}_{\theta_0}(p_{\theta_1}(Z) \leq c\ p_{\theta_0}(Z)) + \gamma\mathbb{P}_{\theta_0}(p_{\theta_1}(Z) < c\ p_{\theta_0}(Z)) = (1-\gamma)F(c) + \gamma F(c^-) ,$$

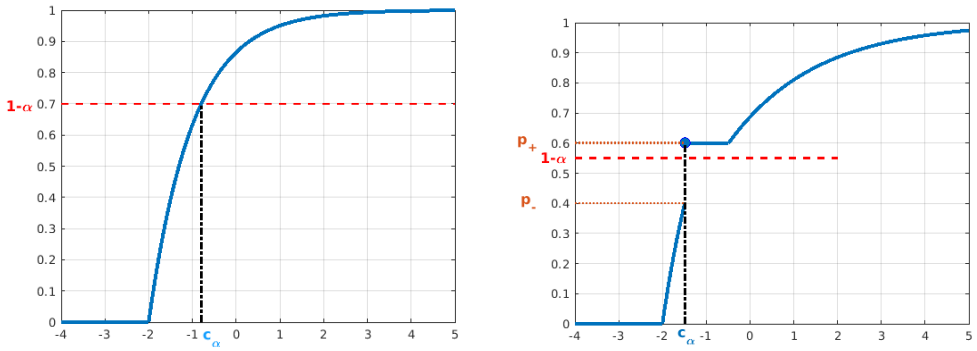

FIGURE I-5.1 – Pour deux exemples de fonction de répartition F (en trait plein, bleu), illustration de l'existence et unicité de la paire (c,γ) vérifiant (I-5.4).

sur le schéma de gauche où l'inverse généralisée F^{-1} est continue en $1-\alpha$, nous voyons qu'il suffit de prendre $c_\alpha = F^{-1}(1-\alpha)$ et $\gamma_\alpha = 0$; sur le schéma de droite, nous prenons $c_\alpha = F^{-1}(1-\alpha)$ et γ_α est l'unique coefficient qui écrit $1-\alpha$ comme barycentre des quantités $p_- := F(c_\alpha^-)$ et $p_+ := F(c_\alpha)$.

• Nous allons maintenant démontrer que $\phi^\star$ est U.P.P.(α). Soit ϕ une fonction de test de niveau α i.e. $\mathbb{E}_{\theta_0}[\phi(Z)] \le \alpha$. Montrons que pour tout $z \in \mathrm{Z}$,

$$\{\phi^\star(z)-\phi(z)\}\{p_{\theta_1}(z)-c_\alpha p_{\theta_0}(z)\} \ge 0\,. \tag{I-5.6}$$

Distinguons deux cas :

— Si $\phi^\star(z)-\phi(z) > 0$, alors $\phi^\star(z) > 0$ (puisque $\phi^\star(z) > \phi(z) \ge 0$) et donc $p_{\theta_1}(z) \ge c_\alpha p_{\theta_0}(z)$.

— Si $\phi^\star(z)-\phi(z) < 0$, alors $\phi^\star(z) < 1$ (puisque $\phi^\star(z) < \phi(z) \le 1$) et donc $p_{\theta_1}(z) \le c_\alpha p_{\theta_0}(z)$.

Ce qui conclut la preuve de (I-5.6). Par conséquent, on a

$$\int \{\phi^\star(z)-\phi(z)\}\{p_{\theta_1}(z)-c_\alpha p_{\theta_0}(z)\}\nu(\mathrm{d}z) \ge 0\,,$$

ce qui entraine

$$\begin{aligned}\int \{\phi^\star(z)-\phi(z)\}p_{\theta_1}(z)\nu(\mathrm{d}z) &= \mathbb{E}_{\theta_1}[\phi^\star(Z)] - \mathbb{E}_{\theta_1}[\phi(Z)] \\ &\ge c_\alpha \int \{\phi^\star(z)-\phi(z)\}p_{\theta_0}(z)\nu(\mathrm{d}z) = c_\alpha\left(\mathbb{E}_{\theta_0}[\phi^\star(Z)] - \mathbb{E}_{\theta_0}[\phi(Z)]\right)\,.\end{aligned}$$

Comme le test $\phi^\star$ est de taille α et que le test ϕ est de niveau α, il vient

$$\mathbb{E}_{\theta_0}[\phi^\star(Z)] - \mathbb{E}_{\theta_0}[\phi(Z)] = \alpha - \mathbb{E}_{\theta_0}[\phi(Z)] \ge 0$$

ce qui prouve que $\phi^\star$ est U.P.P.(α).

•. Enfin, comparons deux tests U.P.P.(α). Considérons le test randomisé de fonction de test constante $\psi(z) = \alpha$. Nous avons

$$\mathbb{E}_{\theta_0}[\psi(Z)] = \alpha\,, \qquad \beta_\psi(\theta_1) = \alpha\,.$$

Comme le test de fonction critique $\phi^\star$ défini ci-dessus est U.P.P.(α), nous avons : $\mathbb{E}_{\theta_1}[\phi^\star(Z)] \ge \alpha$.

Soit $\phi^{\star\star}$ la fonction critique d'un test U.P.P.(α). Comme $\phi^\star$ et $\phi^{\star\star}$ sont tous les deux U.P.P.(α) (donc tous les deux de niveau α et de puissance dominant celle de tout autre test de niveau α et $\theta = \theta_1$), nous avons

$$\mathbb{E}_{\theta_1}[\phi^\star(Z)] - \mathbb{E}_{\theta_1}[\phi^{\star\star}(Z)] = 0\,.$$

De plus, puisque la taille de $\phi^\star$ est α et que $\phi^{\star\star}$ est de niveau α, il vient

$$c_\alpha\{\mathbb{E}_{\theta_0}[\phi^\star(Z)] - \mathbb{E}_{\theta_0}[\phi^{\star\star}(Z)]\} = c_\alpha\{\alpha - \mathbb{E}_{\theta_0}[\phi^{\star\star}(Z)]\} \ge 0\,.$$

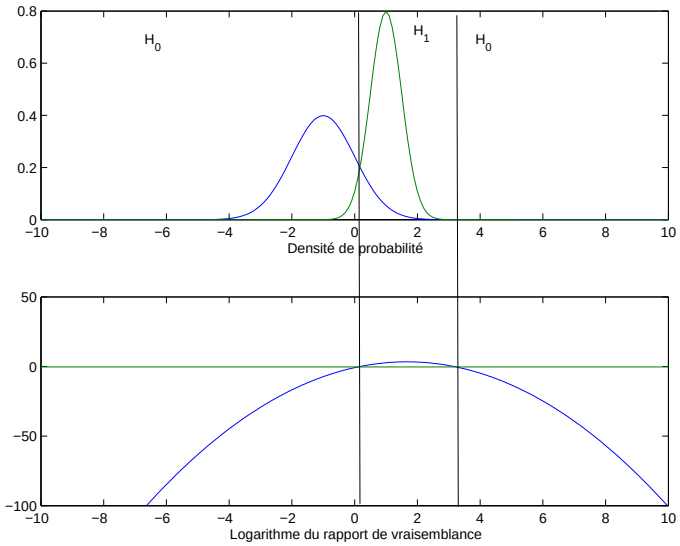

FIGURE I-5.2 – Figure du haut : densité de probabilité de deux variables aléatoires gaussiennes de moyenne et de variance $\theta_0 := (\mu_0, \sigma_0^2) = (-1, 1)$ et $\theta_1 := (\mu_1, \sigma_1^2) = (1, 0.5)$. Figure du bas : logarithme du rapport de vraisemblance $z \mapsto p_{\theta_1}(z)/p_{\theta_0}(z)$.

Ces deux relations entrainent

$$\int \{\phi^\star(z) - \phi^{\star\star}(z)\} p_{\theta_1}(z) \nu(\mathrm{d}z) \leq c_\alpha \int \{\phi^\star(z) - \phi^{\star\star}(z)\} p_{\theta_0}(z) \nu(\mathrm{d}z)$$

et donc

$$\int \{\phi^\star(z) - \phi^{\star\star}(z)\} \{p_{\theta_1}(z) - c_\alpha p_{\theta_0}(z)\} \nu(\mathrm{d}z) \leq 0 \,.$$

Comme par ailleurs (I-5.6) est valide pour $\phi = \phi^{\star\star}$, on obtient que

$$\left\{z \in \mathsf{Z} : \{\phi^\star(z) - \phi^{\star\star}(z)\} \{p_{\theta_1}(z) - c_\alpha p_{\theta_0}(z)\} > 0\right\}$$

est ν-négligeable, ce qui conclut la preuve. □

Exemple I-5.6 (Deux variables gaussiennes scalaires). Considérons un modèle statistique gaussien

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{N(\mu, \sigma^2),\ \theta = (\mu, \sigma^2) \in \Theta := \{\theta_0, \theta_1\}\right\}\right) ,$$

où $\theta_0 \neq \theta_1$; et le test d'hypothèses simples

$$H_0 :\ \theta = \theta_0, \quad \text{contre} \quad H_1 :\ \theta = \theta_1 \,.$$

Notons p_{θ_i} la densité de la loi gaussienne de paramètre $\theta_i = (\mu_i, \sigma_i^2)$,

$$p_{\theta_i}(z) := \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-(z-\mu_i)^2/2\sigma_i^2) \,.$$

La région critique du test U. P. P. défini par le Théorème I-5.5 est donnée par :

$$-\frac{1}{2\sigma_1^2}(z-\mu_1)^2 + \frac{1}{2\sigma_0^2}(z-\mu_0)^2 \geq \log(c_\alpha) + (1/2)\log(\sigma_1^2/\sigma_0^2).$$

Nous avons représenté dans la fig. I-5.2 les deux densités p_{θ_i} ainsi que le logarithme du rapport $r(z) = p_{\theta_1}(z)/p_{\theta_0}(z)$ dans le cas où $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 0.5)$. Les deux traits verticaux délimitent la zone $\{z : p_{\theta_1}(z) > c\, p_{\theta_0}(z)\}$ dans le cas $c = 1$. Dans la fig. I-5.3, nous avons visualisé les régions d'acceptation et de rejet du test lorsque $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 1)$ et le niveau est tel que le seuil c_α vaut 1. La variance est identique sous les

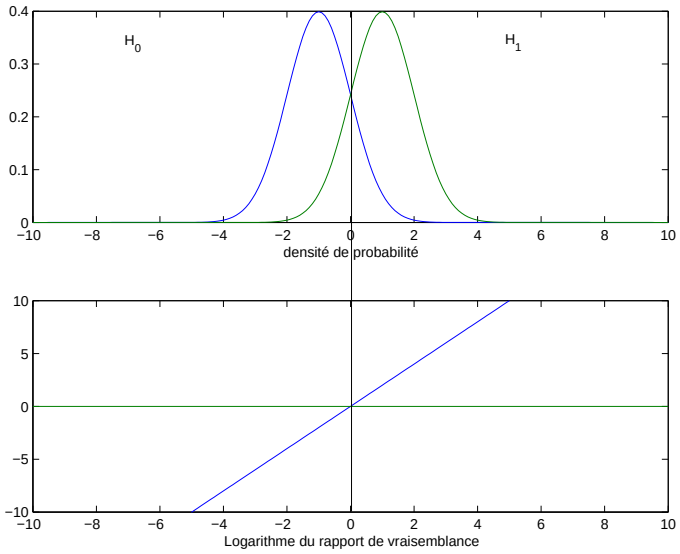

FIGURE I-5.3 – Panneau du haut : densité de probabilité de deux v.a. gaussiennes de moyenne et de variance $(\mu_0,\sigma_0^2)=(-1,1)$ et $(\mu_1,\sigma_1^2)=(1,1)$. Panneau du bas : logarithme du rapport de vraisemblance $z\mapsto p_{\theta_1}(z)/p_{\theta_0}(z)$.

deux hypothèses. On remarque que, dans ce cas particulier, le rapport de vraisemblance est une fonction monotone croissante de z et que la région critique du test est donnée par

$$2z(\mu_1-\mu_0)\geq 2\log(c_\alpha)-(\mu_0^2-\mu_1^2)\,.$$

♢

Exemple I-5.7 (Test de la moyenne de variables aléatoires gaussiennes : variance connue). Soit $(X_1,\ldots,X_n)$ un n-échantillon d'un modèle gaussien

$$(\mathbb{R},\mathscr{B}(\mathbb{R}),\left\{\mathrm{N}(\mu,\sigma^2)\,,\ \mu\in\{0,m\}\right\})$$

où la variance $\sigma^2>0$ est supposée connue et $m\neq 0$. Considérons le test

$$H_0:\ \mu=0,\quad \text{contre}\quad H_1:\ \mu=m$$

où m est une constante connue. Nous cherchons à déterminer un test $\mathrm{U.P.P.}(\alpha)$. Nous formons le rapport de vraisemblance,

$$\begin{aligned} r(x_1,\ldots,x_n) &:= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-m)^2+\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right),\\ &= \exp\left(\frac{nm}{\sigma^2}\bar{x}_n-\frac{nm^2}{2\sigma^2}\right), \end{aligned}$$

où $\bar{x}_n:=n^{-1}\sum_{i=1}^n x_i$ et donc

$$r(x_1,\ldots,x_n)=\tilde{r}(\bar{x}_n)\,,\quad \tilde{r}(t)=\exp\left(\frac{nm}{\sigma^2}t-\frac{nm^2}{2\sigma^2}\right)$$

dépend uniquement de la statistique $\bar{x}_n$. On remarque que la fonction $t\to\tilde{r}(t)$ est une fonction strictement monotone de t, croissante si $m>0$ et décroissante dans le cas contraire. Si $m\geq 0$, la condition $\tilde{r}(\bar{x}_n)>c_\alpha$ est équivalente à $\bar{x}_n>d_\alpha$. Pour déterminer le seuil d_α, nous devons résoudre l'équation suivante, qui correspond à la condition sur la taille du test :

$$\mathbb{P}_0(\bar{X}_n>d_\alpha)=\alpha\,. \tag{I-5.7}$$

Sous $\mathbb{P}_0$, $\sqrt{n}\bar{X}_n/\sigma\sim\mathrm{N}(0,1)$; l'équation (I-5.7) admet comme seule solution $d_\alpha=z_{1-\alpha}\sigma/\sqrt{n}$ où $z_{1-\alpha}$ est le quantile $1-\alpha$ de la loi $\mathrm{N}(0,1)$. Il est intéressant de remarquer que le seuil d_α ne dépend pas de m, la valeur de la moyenne sous l'hypothèse alternative. La puissance du test est alors donnée par :

$$\mathbb{P}_m(\bar{X}_n>z_{1-\alpha}\sigma/\sqrt{n})=1-\Phi(z_{1-\alpha}-\sqrt{n}m/\sigma)\,.$$

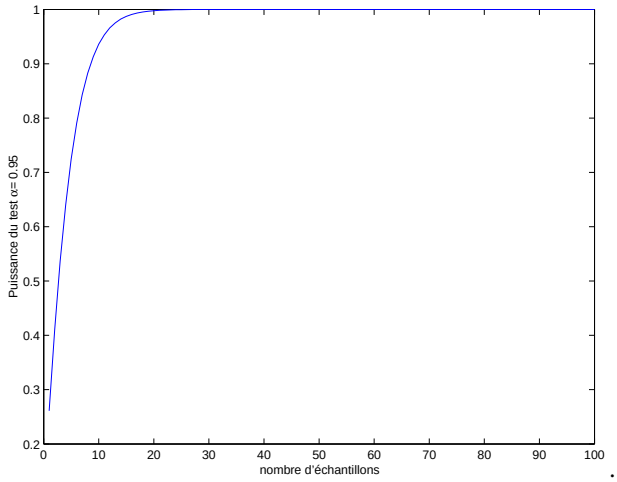

FIGURE I-5.4 – Puissance du test U.P.P.(α) de $H_0 = \{\mu = 0\}$ contre $H_1 = \{\mu = 1\}$ de niveau $\alpha = 0,05$ en fonction de la taille de l'échantillon.

Nous avons visualisé dans la figure I-5.4 la fonction puissance dans le cas particulier où $m = 1$, $\sigma = 1$ et $\alpha = 0.05$ ($z_{1-\alpha} = 1.6449$), pour des tailles d'échantillon variant de 10 à 1000. Ce test se généralise aisément au cas où la moyenne sous la contre-alternative n'est pas constante, i.e. les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes de loi gaussienne de moyenne $m_1, \ldots, m_n$ et de variance unité. Dans ce cas particulier, le rapport de vraisemblance vaut

$$r(x_1, \ldots, x_n) = \exp\left(\frac{1}{\sigma^2} \sum_{i=1}^{n} m_i x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} m_i^2 \right).$$

Le rapport de vraisemblance est cette fois fonction de la statistique $\sum_{i=1}^n m_i X_i$, et le test de rapport de vraisemblance est alors de la forme

$$\sum_{i=1}^{n} m_i X_i \geq d_\alpha \, .$$

En remarquant que sous $\mathbb{P}_0$,

$$\frac{\sum_{i=1}^n m_i X_i}{\sigma \sqrt{\sum_{i=1}^n m_i^2}} \sim \mathrm{N}(0,1) \, ,$$

on obtient un test de niveau α en rejetant l'hypothèse de base si

$$\sum_{i=1}^{n} m_i X_i \geq z_{1-\alpha} \sigma \sqrt{\sum_{i=1}^{n} m_i^2} \, .$$

Ce test est à la base de nombreuses applications en traitement du signal. ♢

Exemple I-5.8 (Variance d'une gaussienne : moyenne connue). Soit $(X_1, \ldots, X_n)$ un n-échantillon de v.a. gaussiennes $\mathrm{N}(0, \theta)$. Nous souhaitons tester l'hypothèse $\theta = \theta_0$ contre $\theta = \theta_1$, où $0 < \theta_0 < \theta_1$. Le rapport de vraisemblance est de la forme :

$$r(x_1, \ldots, x_n) = \left(\frac{\theta_0}{\theta_1} \right)^{n/2} \exp\left(- \left(\frac{1}{2\theta_1} - \frac{1}{2\theta_0} \right) \sum_{i=1}^{n} x_i^2 \right).$$

La condition $r(x_1, \ldots, x_n) > c_\alpha$ est équivalente à $\sum_{i=1}^n x_i^2 > d_\alpha$ pour un d_α convenablement choisi. Pour déterminer le seuil d_α, nous devons donc résoudre l'équation

$$\mathbb{P}_{\theta_0}\left(\sum_{i=1}^{n} X_i^2 \geq d_\alpha \right) = \alpha \, .$$

Comme sous $\mathbb{P}_{\theta_0}$, $\sum_{i=1}^n X_i^2 / \theta_0$ est distribuée suivant une loi du χ^2 centrée à n degrés de liberté, on peut déterminer d_α à partir des quantiles de cette loi. ♢

Exemple I-5.9 (Sondage). Soit $(X_1,\dots,X_n)$ un n-échantillon de Bernoulli

$$(\{0,1\},\mathscr{P}(\{0,1\}),\{\mathrm{Ber}(\theta)\,,\ \theta\in\Theta:=\{\theta_0,\theta_1\}\})\ .$$

Considérons le test

$$H_0:\ \theta=\theta_0,\quad \text{contre}\quad H_1:\ \theta=\theta_1$$

où $0\le\theta_0<\theta_1\le 1$. En posant $S(Z)=\sum_{i=1}^n X_i$ le nombre total de succès, le rapport de vraisemblance s'écrit

$$r(Z)=\frac{\theta_1^{S(Z)}(1-\theta_1)^{n-S(Z)}}{\theta_0^{S(Z)}(1-\theta_0)^{n-S(Z)}}$$

Par conséquent, le rapport de vraisemblance est une fonction de la statistique $S(Z)$, $r(Z)=\tilde r(S(Z))$ avec

$$\tilde r(s)=\left(\frac{\theta_1}{\theta_0}\right)^s\left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-s}\ ,$$

le théorème de Neyman-Pearson (Théorème I-5.5) implique qu'il existe des constantes c_α et γ telles que le test de fonction critique

$$\phi^\star(Z):=\begin{cases}1 & \text{si } \tilde r(S(Z))>c_\alpha\\ \gamma & \text{si } \tilde r(S(Z))=c_\alpha\\ 0 & \text{si } \tilde r(S(Z))<c_\alpha\end{cases}$$

soit U.P.P.(α). La fonction $s\mapsto\tilde r(s)$ est monotone croissante en s (croissante car on a supposé que $\theta_1>\theta_0$) ; ce qui implique que le test précédent peut s'écrire

$$\phi^\star(Z):=\begin{cases}1 & \text{si } S(Z)>m_\alpha\\ \gamma & \text{si } S(Z)=m_\alpha\\ 0 & \text{si } S(Z)<m_\alpha\end{cases}\ .$$

Pour que le test soit de niveau α, les constantes $m_\alpha\in\mathbb{N}$ et γ doivent vérifier l'équation

$$\alpha=\mathbb{E}_{\theta_0}[\phi^\star(Z)]=\mathbb{P}_{\theta_0}(S(Z)>m_\alpha)+\gamma\mathbb{P}_{\theta_0}(S(Z)=m_\alpha)\ .$$

Comme, sous $\mathbb{P}_{\theta_0}$, $S(Z)$ est distribuée suivant une loi binômiale de paramètres (n,θ_0), nous pouvons déterminer m_α et γ en résolvant

$$\alpha=\sum_{j=m_\alpha+1}^n\binom{n}{j}\theta_0^j(1-\theta_0)^{n-j}+\gamma\binom{n}{m_\alpha}\theta_0^{m_\alpha}(1-\theta_0)^{n-m_\alpha}\ .$$

A l'exception des valeurs de α vérifiant

$$\alpha=\sum_{j=m_\alpha+1}^n\binom{n}{j}\theta_0^j(1-\theta_0)^{n-j}\ ,$$

pour un entier m_α (auquel cas nous pouvons poser $\gamma=0$), le test U.P.P. est un test randomisé. ♢

I-5.2 Rapport de vraisemblance monotone

Nous avons construit des tests U.P.P. dans le cas de deux hypothèses simples. La situation la plus simple, quand on cherche à généraliser les tests au delà des hypothèses simples est de supposer que le paramètre inconnu est scalaire et que l'on considère un test d'hypothèses *unilatérales*.

Plus précisément, considérons un modèle statistique $(\mathsf{Z},\mathscr{Z},\{\mathbb{P}_\theta\,,\ \theta\in\Theta\})$ où $\Theta\subseteq\mathbb{R}$. Une hypothèse est dite *unilatérale* lorsqu'elle est de la forme $\theta\le\theta_0$ ou $\theta>\theta_0$. Dans cette section, nous considérons des tests avec hypothèses unilatérales. Le test le plus puissant pour une hypothèse unilatérale contre une alternative simple $\theta=\theta_1$ peut dépendre de la valeur de θ_1, et dans ce cas, on ne sait pas construire de test uniformément plus puissant pour une hypothèse alternative unilatérale. Nous allons voir toutefois qu'il existe des tests U.P.P. pour un test d'hypothèses unilatérales lorsque l'on impose une hypothèse supplémentaire sur la structure statistique du modèle.

***H* I-5.10 (Famille à rapport de vraisemblance monotone).** Soient μ une mesure σ-finie sur $(\mathsf{Z},\mathscr{Z})$ et $(\mathsf{Z},\mathscr{Z},\{\mathbb{P}_\theta = p_\theta\cdot\mu, \theta\in\Theta\})$ un modèle statistique où $\Theta\subseteq\mathbb{R}$. La famille est à *rapport de vraisemblance monotone* si

1. il existe $\psi:\mathbb{R}^2\to\mathbb{R}_+$, $T:\mathsf{Z}\to\mathbb{R}$ et $h:\mathsf{Z}\to\mathbb{R}^+$ mesurables telles que
$$p_\theta(z) = h(z)\ \psi(T(z);\theta)\ . \tag{I-5.8}$$
2. pour tout $\theta,\theta'\in\Theta$ tels que $\theta>\theta'$, la fonction
$$t\mapsto\frac{\psi(t;\theta)}{\psi(t;\theta')}$$
est une fonction strictement croissante de t. ♢

Le fait de supposer que pour $\theta>\theta'$ la fonction $t\mapsto\psi(t;\theta)/\psi(t;\theta')$ est croissante ne fait pas perdre de généralité. En effet, si le rapport est monotone décroissant, nous posons $\widetilde{T}(z):=-T(z)$ et $\widetilde{\psi}(t;\theta):=\psi(-t;\theta)$ et écrivons
$$p_\theta(z) = h(z)\widetilde{\psi}(\widetilde{T}(z);\theta)\ .$$
La fonction $t\mapsto\widetilde{\psi}(t;\theta)/\widetilde{\psi}(t;\theta')$ est monotone croissante, et l'on se retrouve dans le cadre de la définition.

Exemple I-5.11 (Loi gaussienne $\mathrm{N}(\theta,1)$*).* Soit $Z=(X_1,\dots,X_n)$ un n-échantillon gaussien
$$(\mathbb{R},\mathscr{B}(\mathbb{R}),\{\mathrm{N}(\theta,1)\ ,\ \theta\in\mathbb{R}\})\ .$$
La vraisemblance est donnée, en notant $z=(x_1,\dots,x_n)$,
$$p_\theta(z) := \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)\exp\left(\theta nT(z)-(n/2)\theta^2\right)$$
où nous avons posé $T(z):=n^{-1}\sum_{i=1}^n x_i$. La densité $p_\theta(z)$ satisfait la condition **H** I-5.10 avec
$$h(z) := \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)\ ,\quad \psi(t;\theta):=\exp(n\theta t-(n/2)\theta^2)\ .$$
Pour tout $\theta>\theta'$, la fonction
$$t\mapsto\frac{\psi(t;\theta)}{\psi(t;\theta')} = \exp\left((\theta-\theta')nt-(n/2)(\theta^2-\theta'^2)\right)$$
est strictement croissante. ♢

Exemple I-5.12 (Loi gaussienne $\mathrm{N}(0,\theta^{-2})$*).* Soit $(X_1,\dots,X_n)$ un n-échantillon de gaussien
$$\left(\mathbb{R},\mathscr{B}(\mathbb{R}),\left\{\mathrm{N}(0,\theta^{-2})\ ,\ \theta\in\Theta:=\mathbb{R}_+^*\right\}\right)\ .$$
Le paramètre θ^{-2} est l'*inverse* de la variance ; cette quantité est souvent appelée la *précision*. Cette famille est dominée par la mesure de Lebesgue sur $\mathbb{R}^n$ et en posant $z=(x_1,\dots,x_n)\in\mathbb{R}^n$, nous avons
$$p_\theta(z) := \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{\theta^2}{2}\sum_{i=1}^n x_i^2+n\log(\theta)\right)\ .$$
On peut donc prendre ici $T(z):=-\sum_{i=1}^n x_i^2$ et
$$h(z) := \frac{1}{(2\pi)^{n/2}}\ ,\quad \psi(t;\theta):=\exp\left\{\frac{\theta^2}{2}t+n\log(\theta)\right\}\ .$$
Pour tout $\theta>\theta'$, la fonction
$$t\mapsto\frac{\psi(t;\theta)}{\psi(t;\theta')} = \exp\left\{\frac{1}{2}(\theta^2-\theta'^2)t+n\{\log(\theta)-\log(\theta')\}\right\}$$
est strictement croissante. ♢

Exemple I-5.13 (Loi binômiale). Soient $(X_1,\dots,X_n)$ un n-échantillon de Bernoulli

$$(\{0,1\},\mathscr{P}(\{0,1\}),\{\mathrm{Ber}(\theta)\,,\ \theta\in\Theta:=[0,1]\})\,.$$

Le modèle est dominé par rapport à la mesure de comptage sur $\{0,1\}^n$ et la vraisemblance est donnée, en $z=(x_1,\dots,x_n)\in\{0,1\}^n$, par

$$p_\theta(z):=(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i}\,.$$

La fonction $p_\theta(z)$ satisfait **H** I-5.10 avec $T(z):=\sum_{i=1}^n x_i$ et

$$h(z):=1\,,\quad \psi(t;\theta):=(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^t$$

Pour tout $\theta>\theta'$, la fonction

$$t\mapsto\frac{\psi(t;\theta)}{\psi(t;\theta')}=\frac{(1-\theta)^n}{(1-\theta')^n}\left(\frac{\theta(1-\theta')}{\theta'(1-\theta)}\right)^t$$

est strictement croissante : en effet, la fonction $\theta\mapsto\theta/(1-\theta)$ est strictement croissante, par conséquent pour $\theta'<\theta$,

$$\frac{\theta(1-\theta')}{\theta'(1-\theta)}>1\,.$$

♢

De façon plus générale, supposons que l'observation $Z=(X_1,\dots,X_n)$ est un n-échantillon d'une famille exponentielle (Chapitre IV-4) associée à la paire (h,T), où T est une statistique scalaire, c'est-à-dire que $\mathbb{P}_\theta=p_\theta\cdot\mu$ s'écrit

$$p_\theta(x)=h(x)\exp(\phi(\theta)T(x)-\psi(\theta))\,.$$

Nous avons

$$\frac{p_\theta^{\otimes n}(x_1,\cdots,n)}{p_{\theta'}^{\otimes n}(x_1,\cdots,x_n)}=\exp\left(\left(\phi(\theta)-\phi(\theta')\right)\sum_{i=1}^n T(x_i)-n\psi(\theta)+n\psi(\theta')\right)\,.$$

Un tel modèle est à rapports de vraisemblance monotones en $\sum_{i=1}^n T(X_i)$ si et seulement si la fonction $\theta\mapsto\phi(\theta)$ est monotone. Par exemple, si la fonction $\theta\mapsto\phi(\theta)$ est strictement croissante, alors le modèle est à rapports de vraisemblance croissants en $\sum_{i=1}^n T(X_i)$. Si la fonction $\theta\mapsto\phi(\theta)$ est strictement décroissante, le modèle est à rapports de vraisemblance croissants en $-\sum_{i=1}^n T(X_i)$

Remarquons que **H** I-5.10 implique que, pour tout $\theta>\theta'$ et tout c, la condition $p_\theta^{\otimes n}(z)/p_{\theta'}^{\otimes n}(z)\geq c$ s'écrit de manière équivalente $\sum_{i=1}^n x_i\geq d(\theta,\theta',c)$.

Le lemme suivant est utile pour étudier la propriété U.P.P. des tests sur les familles à rapports de vraisemblance monotones.

Lemme I-5.14. *Supposons **H** I-5.10. Soit $\varphi:\mathbb{R}\to\mathbb{R}$ une fonction monotone croissante (au sens large) telle que, pour tout $\theta\in\Theta$, $\mathbb{E}_\theta[|\varphi\circ T(Z)|]<\infty$. Alors la fonction*

$$\theta\mapsto\mathbb{E}_\theta[\varphi\circ T(Z)]$$

est une fonction croissante (au sens large)

Démonstration. Soit $\theta'<\theta$. Nous écrivons

$$\begin{aligned}\mathbb{E}_\theta[\varphi\circ T(Z)]-\mathbb{E}_{\theta'}[\varphi\circ T(Z)]&=\int\varphi\circ T(z)\left\{p_\theta(z)-p_{\theta'}(z)\right\}\mu(\mathrm{d}z)\\&=\int_A\varphi\circ T(z)\left\{p_\theta(z)-p_{\theta'}(z)\right\}\mu(\mathrm{d}z)+\int_{A^c}\varphi\circ T(z)\left\{p_\theta(z)-p_{\theta'}(z)\right\}\mu(\mathrm{d}z)\end{aligned}$$

où l'on pose

$$A:=\{z\in\mathsf{Z}:\ p_{\theta'}(z)>p_\theta(z)\}\,.$$

Sous **H** I-5.10, nous avons

$$\frac{p_\theta(z)}{p_{\theta'}(z)}=\frac{\psi(T(z);\theta)}{\psi(T(z);\theta')}\,,$$

où $t \mapsto \psi(t;\theta)/\psi(t;\theta')$ est une fonction monotone croissante. Par suite,

$$A = \{z \in \mathsf{Z} : \, T(z) < d\}$$

où $d := \sup\{t \in \mathbb{R} : \, \psi(t;\theta)/\psi(t;\theta') < 1\}$. Par hypothèse, la fonction φ est monotone croissante ; par conséquent,

$$b := \inf_{z \in A^c} \varphi \circ T(z) \geq a := \sup_{z \in A} \varphi \circ T(z) \, .$$

Nous avons donc

$$\begin{aligned} \mathbb{E}_\theta[\varphi \circ T(Z)] - \mathbb{E}_{\theta'}[\varphi \circ T(Z)] &\geq a \int_A \{p_\theta(z) - p_{\theta'}(z)\} \, \mu(\mathrm{d}z) + b \int_{A^c} \{p_\theta(z) - p_{\theta'}(z)\} \, \mu(\mathrm{d}z) \\ &= (b-a) \int_{A^c} \{p_\theta(z) - p_{\theta'}(z)\} \, \mu(\mathrm{d}z) \geq 0 \, ; \end{aligned}$$

dans la dernière égalité, nous avons utilisé que

$$\int_A \{p_\theta(z) - p_{\theta'}(z)\} \, \mu(\mathrm{d}z) + \int_{A^c} \{p_\theta(z) - p_{\theta'}(z)\} \, \mu(\mathrm{d}z) = 0 \, . \qquad \square$$

Théorème I-5.15. *Supposons* ***H*** *I-5.10. Soient* $\theta_0 \in \Theta$ *et* $\alpha \in (0,1)$. *Alors :*

(i) Il existe un test $\mathrm{U.P.P.}(\alpha)$ *de*

$$H_0 : \, \theta \leq \theta_0, \quad \textit{contre} \quad H_1 : \, \theta > \theta_0 \, .$$

La fonction critique de ce test est donnée par :

$$\phi(z) := \begin{cases} 1 & \textit{si } T(z) > c, \\ \gamma & \textit{si } T(z) = c, \\ 0 & \textit{si } T(z) < c, \end{cases} \tag{I-5.9}$$

où les constantes c et γ *sont solutions de l'équation :*

$$\mathbb{P}_{\theta_0}(T(Z) > c) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c) = \alpha \, .$$

(ii) La puissance du test

$$\theta \mapsto \beta_\phi(\theta) := \mathbb{E}_\theta[\phi(Z)]$$

est une fonction strictement croissante sur l'ensemble $\big\{\theta \in \Theta : \, \beta_\phi(\theta) < 1\big\}$.

Si nous souhaitons tester

$$H_0 : \, \theta \geq \theta_0, \quad \text{contre} \quad H_1 : \, \theta < \theta_0$$

le théorème I-5.15 reste vrai en changeant le sens des inégalités dans la définition de la fonction critique du test

$$\phi(z) := \begin{cases} 1 & \text{si } T(z) < c, \\ \gamma & \text{si } T(z) = c, \\ 0 & \text{si } T(z) > c, \end{cases} \tag{I-5.10}$$

où les constantes c et γ sont solutions de l'équation :

$$\mathbb{P}_{\theta_0}(T(Z) < c) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c) = \alpha \, .$$

Démonstration. Considérons tout d'abord pour $\theta_1 > \theta_0$ le test d'hypothèses simple

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1 , \tag{I-5.11}$$

Soit $\alpha \in]0,1[$. Le théorème de Neyman-Pearson (théorème I-5.5) montre qu'il existe des constantes d_{θ_0,θ_1} et $\gamma_{\theta_0,\theta_1}$ telles que

$$\mathbb{P}_{\theta_0}\left(p_{\theta_1}(Z) > d_{\theta_0,\theta_1} p_{\theta_0}(Z)\right) + \gamma_{\theta_0,\theta_1}\, \mathbb{P}_{\theta_0}\left(p_{\theta_1}(Z) = d_{\theta_0,\theta_1} p_{\theta_0}(Z)\right) = \alpha ; \tag{I-5.12}$$

et que le test de fonction critique ϕ_{θ_0,θ_1}

$$\phi_{\theta_0,\theta_1}(z) := \begin{cases} 1 & \text{si } p_{\theta_1}(Z) > d_{\theta_0,\theta_1} p_{\theta_0}(Z), \\ \gamma_{\theta_0,\theta_1} & \text{si } p_{\theta_1}(Z) = d_{\theta_0,\theta_1} p_{\theta_0}(Z), \\ 0 & \text{si } p_{\theta_1}(Z) < d_{\theta_0,\theta_1} p_{\theta_0}(Z) \end{cases}$$

est U.P.P.(α) pour le test d'hypothèses simple (I-5.11).

• Nous allons maintenant exploiter la monotonie du rapport de vraisemblance pour montrer que le test ϕ_{θ_0,θ_1} est indépendant de θ_1. Comme la fonction $t \mapsto \psi(t;\theta_1)/\psi(t;\theta_0)$ est strictement croissante en t, pour tout $d > 0$, il existe $c > 0$ tel que

$$\{z : p_{\theta_1}(z) > d p_{\theta_0}(z)\} = \{z : T(z) > c\} , \qquad \{z : p_{\theta_1}(z) = d p_{\theta_0}(z)\} = \{z : T(z) = c\} .$$

Par suite, le test ϕ_{θ_0,θ_1} est le test

$$\phi_{\theta_0,\theta_1}(z) = \begin{cases} 1 & \text{si } T(Z) > c_{\theta_0,\theta_1}, \\ \gamma_{\theta_0,\theta_1} & \text{si } T(Z) = c_{\theta_0,\theta_1}, \\ 0 & \text{si } T(Z) < c_{\theta_0,\theta_1} \end{cases}$$

où $c_{\theta_0,\theta_1}, \gamma_{\theta_0,\theta_1}$ sont solutions de

$$\mathbb{P}_{\theta_0}(T(Z) > c) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c) = \alpha . \tag{I-5.13}$$

Cette équation ne dépend pas de θ_1 et par conséquent, les constantes $c_{\theta_0,\theta_1}, \gamma_{\theta_0,\theta_1}$ et le test ϕ_{θ_0,θ_1} sont indépendants de θ_1. Dans la suite, nous posons $\phi_{\theta_0,\theta_1} := \phi_{\theta_0}$, $c_{\theta_0,\theta_1} := c_{\theta_0}$ et $\gamma_{\theta_0,\theta_1} := \gamma_{\theta_0}$:

$$\phi_{\theta_0}(z) = \begin{cases} 1 & \text{si } T(z) > c_{\theta_0}, \\ \gamma_{\theta_0} & \text{si } T(z) = c_{\theta_0}, \\ 0 & \text{si } T(z) < c_{\theta_0}. \end{cases} \tag{I-5.14}$$

• Nous montrons maintenant que le test ϕ_{θ_0} est U.P.P.(α) pour le test

$$H_0' : \theta = \theta_0 , \qquad H_1' : \theta > \theta_0 . \tag{I-5.15}$$

Par construction (voir (I-5.13)), $\beta_{\phi_{\theta_0}}(\theta_0) = \mathbb{E}_{\theta_0}[\phi_{\theta_0}(Z)] = \alpha$ donc ϕ_{θ_0} est de taille et de niveau α pour le test (I-5.15). De plus, pour tout $\theta_1 > \theta_0$, le test de fonction critique ϕ_{θ_0} est U.P.P.(α) pour le test d'hypothèses simple

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1 .$$

Par conséquent, ceci étant vrai pour tout $\theta_1 > \theta_0$, le test de fonction critique ϕ_{θ_0} est aussi U.P.P.(α) pour le test d'hypothèse (I-5.15).

• Enfin, nous montrons que le test ϕ_{θ_0} reste U.P.P.(α) pour le test qui nous intéresse :

$$H_0^\star : \theta \le \theta_0 , \qquad H_1^\star : \theta > \theta_0 . \tag{I-5.16}$$

La fonction critique ϕ_{θ_0} définie par (I-5.14) peut s'écrire $\phi_{\theta_0}(z) = \varphi(T(z))$ avec $\varphi(u) = 0$ si $u < c_{\theta_0}$, $\varphi(u) = \gamma_{\theta_0}$ si $u = c_{\theta_0}$ et $\varphi(u) = 1$ si $u > c_{\theta_0}$. La fonction φ est croissante. Le lemme I-5.14 montre que la puissance du test de fonction critique ϕ_{θ_0} donnée par

$$\theta \mapsto \beta_{\phi_{\theta_0}}(\theta)$$

est une fonction croissante du paramètre θ. En particulier, pour tout $\theta \leq \theta_0$,

$$\beta_{\phi_{\theta_0}}(\theta) \leq \beta_{\phi_{\theta_0}}(\theta_0) \, .$$

Par conséquent, le test randomisé de fonction critique ϕ_{θ_0} est un test de niveau α de l'hypothèse (I-5.16) ; ce que l'on écrit $\phi_{\theta_0} \in \mathscr{K}_\alpha(\{\theta \leq \theta_0\})$. Comme

$$\mathscr{K}_\alpha(\{\theta \leq \theta_0\}) \subset \mathscr{K}_\alpha(\{\theta = \theta_0\})$$

et comme ϕ_{θ_0} est U.P.P.(α) dans la classe $\mathscr{K}_\alpha(\{\theta = \theta_0\})$ contre les alternatives de la forme $H_1 = \{\theta > \theta_0\}$, ϕ_{θ_0} est U.P.P.(α) dans $\mathscr{K}_\alpha(\{\theta \leq \theta_0\})$ contre $H_1^\star = \{\theta > \theta_0\}$. Ceci démontre le point (i) du théorème. Le point (ii) a été établi lorsque nous avons invoqué le lemme I-5.14. □

Exemple I-5.16 (Modèle binomial). Soit $Z = (X_1, \ldots, X_n)$ un n-échantillon de Bernoulli

$$(\{0,1\}, \mathscr{P}(\{0,1\}), \{\mathrm{Ber}(\theta) \, , \, \theta \in \Theta := [0,1]\}) \, .$$

Le modèle est dominé par rapport à la mesure de comptage sur $\{0,1\}^n$ et la vraisemblance est donnée, en $z = (x_1, \ldots, x_n) \in \{0,1\}^n$, par

$$p_\theta(z) := (1-\theta)^n \left(\frac{\theta}{1-\theta} \right)^{\sum_{i=1}^n x_i} \, .$$

Ce modèle satisfait **H** I-5.10 (voir exemple I-5.13) avec $T(z) = \sum_{i=1}^n x_i$. Pour $\theta_0 \in (0,1)$, nous effectuons tout d'abord le test

$$H_0 : \, \theta \leq \theta_0, \quad \text{contre} \quad H_1 : \, \theta > \theta_0 \, .$$

Pour tout $\alpha \in]0,1[$ considérons la fonction

$$\phi_{\alpha,\theta_0}(z) := \begin{cases} 1 & \text{si } T(z) > c_{\alpha,\theta_0}, \\ \gamma_{\alpha,\theta_0} & \text{si } T(z) = c_{\alpha,\theta_0}, \\ 0 & \text{si } T(z) < c_{\alpha,\theta_0}, \end{cases} \tag{I-5.17}$$

où les constantes c_{α,θ_0} et γ_{α,θ_0} sont solutions de l'équation :

$$\mathbb{E}_{\theta_0}[\phi_{\alpha,\theta_0}(Z)] = \mathbb{P}_{\theta_0}(T(Z) > c_{\alpha,\theta_0}) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c_{\alpha,\theta_0}) = \alpha \, .$$

Le théorème I-5.15 montre que le test de fonction critique ϕ_{α,θ_0} est U.P.P.(α). Sous $\mathbb{P}_{\theta_0}$ la loi de la statistique de test est binomiale de paramètre n et θ_0. Il faut donc déterminer tout d'abord la constante c_{α,θ_0}

$$c_{\alpha,\theta_0} = \inf\left\{ c \in \{0, \ldots, n\} \, : \, \mathbb{P}_{\theta_0}(T(Z) > c) \leq \alpha \right\} \, .$$

puis calculer

$$\gamma_{\alpha,\theta_0} = \frac{\alpha - \mathbb{P}_{\theta_0}(T(Z) > c_\alpha)}{\mathbb{P}_{\theta_0}(T(Z) = c_\alpha)} \, .$$

♢

A titre d'exemple, si $\theta_0 = 1/2$, $n = 1000$, et $\alpha = 0.05$, nous avons $c_{0.05,1/2} = 526$ et $\gamma_{0.05,1/2} = 0.4832$.

Exemple I-5.17 (Variance d'une loi gaussienne (suite)). Soit $(X_1, \cdot, X_n)$ un n-échantillon gaussien

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{\mathrm{N}(0,\theta) \, , \, \theta \in \Theta := \mathbb{R}_+^*\right\}\right) \, .$$

Considérons l'hypothèse de base $H_0 = \{\theta \geq \theta_0\}$ et l'hypothèse alternative $H_1 = \{\theta < \theta_0\}$. Le rapport de vraisemblance est strictement croissant par rapport à la statistique $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i^2$. Le test U.P.P.$(\alpha)$ rejette H_0 lorsque $T(X_1, \ldots, X_n) \leq d_\alpha$, où d_α est solution de l'équation :

$$\mathbb{P}_{\theta_0}(T \leq d_\alpha) = \alpha.$$

Sous $\mathbb{P}_{\theta_0}$, $T(X_1, \ldots, X_n)/\theta_0 = \sum_{i=1}^n X_i^2/\theta_0$ est un χ^2 centré à n degrés de liberté, la constante critique du test est $\theta_0 \chi^2_\alpha(n)$ où $\chi^2_\alpha(n)$ est le quantile d'ordre α d'un χ^2 à n degrés de liberté. ♢

Exemple I-5.18 (Loi de Poisson). Soit $Z = (X_1, \dots, X_n)$ un n-échantillon de la loi de Poisson

$$(\mathbb{N}, \mathscr{P}(\mathbb{N}), \{\mathrm{Poi}(\theta)\,,\ \theta \in \mathbb{R}_+^*\})\,.$$

La vraisemblance de l'observation (par rapport à la mesure de comptage) est donnée, en posant $z = (x_1, \dots, x_n) \in \mathbb{N}^n$,

$$p_\theta(z) = \frac{1}{\prod_{i=1}^n x_i!} \mathrm{e}^{-n\theta}\, \theta^{T(z)} \quad T(z) := \sum_{i=1}^n x_i\,.$$

C'est une famille à rapports de vraisemblance monotones. Nous effectuons le test avec hypothèses unilatérales

$$H_0 :\ \theta \le \theta_0, \quad \text{contre} \quad H_1 :\ \theta > \theta_0\,. \tag{I-5.18}$$

Notons que sous $\mathbb{P}_\theta$, $T(Z) := \sum_{i=1}^n X_i$ suit une loi de Poisson de paramètre $n\theta$. Soit $\alpha \in]0,1[$. Considérons la fonction

$$\phi_{\alpha,\theta_0}(z) = \begin{cases} 1 & T(z) > c_{\alpha,\theta_0} \\ \gamma_{\alpha,\theta_0} & T(z) = c_{\alpha,\theta_0} \\ 0 & T(z) < c_{\alpha,\theta_0} \end{cases} \tag{I-5.19}$$

où c_{α,θ_0} et γ_{α,θ_0} sont solutions de l'équation

$$\alpha = \mathbb{P}_{\theta_0}(T(Z) > c) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c) = \sum_{j=c+1}^{\infty} \frac{\mathrm{e}^{-n\theta_0}(n\theta_0)^j}{j!} + \gamma \frac{\mathrm{e}^{-n\theta_0}(n\theta_0)^c}{c!}\,.$$

Le test de fonction critique (I-5.19) est U.P.P.(α) pour le test unilatéral (I-5.18). ♢

Exemple I-5.19 (Loi Uniforme). Soit $(X_1, \dots, X_n)$ un n-échantillon d'une loi uniforme

$$\left(\mathbb{R}_+, \mathscr{B}(\mathbb{R}^+), \{\mathrm{Unif}(0,\theta)\,,\ \theta \in \Theta := \mathbb{R}_+^*\}\right)\,.$$

Pour $\theta_0 \in \mathbb{R}_+^*$, considérons le test

$$H_0 :\ \theta \le \theta_0, \quad \text{contre} \quad H_1 :\ \theta > \theta_0\,.$$

La densité de probabilité de $(X_1, \dots, X_n)$ est donnée par $p_\theta(x_1, \dots, x_n) = \theta^{-n} \mathbb{1}_{[0,\theta]}(x_{n:n})$ où $x_{n:n} := \max(x_1, \dots, x_n)$. Pour $\theta_1 < \theta_2$,

$$\frac{p_{\theta_2}(x_1, \dots, x_n)}{p_{\theta_1}(x_1, \dots, x_n)} := \frac{\theta_1^n}{\theta_2^n} \frac{\mathbb{1}_{[0,\theta_2]}(x_{n:n})}{\mathbb{1}_{[0,\theta_1]}(x_{n:n})}\,,$$

qui est croissante en $x_{n:n}$ pour tout $(x_1, \dots, x_n)$ tels que $p_{\theta_1}(x_1, \dots, x_n) > 0$ ou $p_{\theta_2}(x_1, \dots, x_n) > 0$, i.e. $x_{n:n} \le \theta_2$. Donc cette famille vérifie **H** I-5.10.

Sous $\mathbb{P}_{\theta_0}$, la distribution de $X_{n:n}$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par $q_{\theta_0}(x) := n\theta_0^{-n} x^{n-1} \mathbb{1}_{[0,\theta_0]}(x)$ (on obtient ce résultat en écrivant que $\mathbb{P}_{\theta_0}(X_{n:n} \le u) = \left(\mathbb{P}_{\theta_0}(X_1 \le u)\right)^n$). En utilisant cette loi sous $\mathbb{P}_{\theta_0}$, nous en déduisons que le test (pur) donné par (I-5.9) est U.P.P. de niveau α en choisissant c_α de telle sorte que

$$\alpha = \frac{n}{\theta_0^n} \int_{c_\alpha}^{\theta_0} x^{n-1} \mathrm{d}x = 1 - \frac{c_\alpha^n}{\theta_0^n}\,.$$

Dans ce problème, le test U.P.P. n'est pas unique. On peut montrer que le test (randomisé) de fonction critique

$$\phi(X_1, \dots, X_n) = \begin{cases} 1 & \text{si } X_{n:n} > \theta_0 \\ \alpha & \text{si } X_{n:n} \le \theta_0\,, \end{cases}$$

est aussi U.P.P. de niveau α. ♢

Pour aller plus loin : Cas d'hypothèses bilatérales

La construction ci-dessus ne s'étend pas directement au cas d'hypothèses bilatérales. On parle d'hypothèse *bilatérale* lorsqu'elle est de la forme $\theta \neq \theta_0$ ou $\theta \notin]\theta_0, \theta_1[$

Illustrons cela sur le test d'hypothèses

$$H_0 :\ \theta = \theta_0, \quad \text{contre} \quad H_1 :\ \theta \neq \theta_0\,. \tag{I-5.20}$$

Considérons $Z = (X_1, \ldots, X_n)$ un n-échantillon d'une famille exponentielle de densité (par rapport à une mesure de domination μ)

$$p_\theta(x) = h(x)\ \exp(\phi(\theta)T(x) - \psi(\theta))\ , \tag{I-5.21}$$

où les fonctions h, ϕ, T, ψ sont à valeur scalaire et $\theta \mapsto \phi(\theta)$ est une fonction croissante de θ. Supposons aussi que $\mathbb{P}_\theta\left(\sum_{i=1}^n T(X_i) = c\right) = 0$ pour tout $\theta \in \Theta$ et pour tout c.

En vertu du théorème de Neyman-Pearson, le test U.P.P. pour l'hypothèse de base $H_0' = \{\theta = \theta_0\}$ contre l'hypothèse $H_1' = \{\theta = \theta_1\}$ est de la forme (cas $\theta_1 > \theta_0$)

$$Z \mapsto \begin{cases} 1 & \text{si } \sum_{i=1}^n T(X_i) > c \\ 0 & \text{si} \sum_{i=1}^n T(X_i) < c\ , \end{cases}$$

ou (cas $\theta_1 < \theta_0$)

$$Z \mapsto \begin{cases} 1 & \text{si } \sum_{i=1}^n T(X_i) < c \\ 0 & \text{si} \sum_{i=1}^n T(X_i) > c\ . \end{cases}$$

Notons que ce sont des tests non randomisés puisque l'on a supposé que $\{Z : \sum_{i=1}^n T(X_i) = c\}$ est de probabilité nulle sous $\mathbb{P}_\theta$, pour tout θ et tout c.

On voit que la structure des tests U.P.P. est différente suivant que l'on considère des alternatives $\theta_1 > \theta_0$ et $\theta_1 < \theta_0$. Or, un test U.P.P.(α) pour le test (I-5.20) est un test U.P.P.(α) pour le test d'hypothèses H_0', H_1' et ce, quelle que soit la valeur de $\theta_1 \neq \theta_0$. C'est pourquoi il n'existe pas de test U.P.P. dans ce cadre.

Le cas d'un modèle gaussien

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathrm{N}(\theta, 1)\ ,\ \theta \in \mathbb{R}\})\ ,$$

pour lequel on veut tester la moyenne lorsque la variance est connue, illustre le contexte précédent. La mesure de domination est la mesure de Lebesgue sur $\mathbb{R}$; nous avons

$$p_\theta(z) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z-\theta)^2\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\theta^2/2 - z^2/2 + z\theta\right)\ ,$$

de sorte que p_θ est de la forme (I-5.21) en ayant posé

$$h(z) := \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\ , \quad T(z) := z\ , \quad \phi(\theta) := \theta\ , \quad \psi(\theta) := \theta^2/2\ .$$

La fonction ϕ est croissante. Enfin, $\{z : \sum_{i=1}^n x_i = c\}$ est un ensemble de mesure nulle pour la mesure de Lebesgue sur $\mathbb{R}^n$ donc pour tout $\theta \in \mathbb{R}$ et pour tout $c \in \mathbb{R}$, nous avons $\mathbb{P}_\theta(\sum_{i=1}^n X_i = c) = 0$.
Etant donné $\theta_1 \neq \theta_0$, le test de Neyman-Pearson pour le test d'hypothèses

$$H_0' : \theta = \theta_0 \qquad H_1' : \theta = \theta_1 \tag{I-5.22}$$

est donné par ϕ_+ (resp. ϕ_-) dans le cas $\theta_1 > \theta_0$ (resp. $\theta_1 < \theta_0$)

$$\phi_+(Z) := \mathbb{1}_{\{\sum_{i=1}^n X_i > n\theta_0 + \sqrt{n} z_{1-\alpha}\}} \qquad \phi_-(Z) := \mathbb{1}_{\{\sum_{i=1}^n X_i < n\theta_0 + \sqrt{n} z_\alpha\}} = \mathbb{1}_{\{\sum_{i=1}^n X_i < n\theta_0 - \sqrt{n} z_{1-\alpha}\}}$$

où z_u désigne le quantile d'ordre u d'une loi $\mathrm{N}(0,1)$; voir l'exemple I-5.7.
Supposons qu'il existe un test $\phi^\star$ U.P.P.(α) pour le test bilatéral (I-5.20). Alors $\mathbb{E}_{\theta_0}[\phi^\star(Z)] \leq \alpha$ (il est de niveau α) et pour tout autre test ϕ de niveau α du test d'hypothèses (I-5.20), $\beta_{\phi^\star}(\theta) \geq \beta_\phi(\theta)$ pour tout $\theta \neq \theta_0$. On en déduit que $\phi^\star$ est aussi U.P.P.(α) pour le test d'hypothèses (I-5.22), et ce, quelle que soit la valeur de $\theta_1 \neq \theta_0$. D'après le théorème I-5.5 item-3 appliqué dans le cas $\theta_1 > \theta_0$, nous avons donc $\phi^\star(z) = \phi_+(z)$; et appliqué dans le cas $\theta_1 < \theta_0$, nous avons $\phi^\star(z) = \phi_-(z)$. Cela est impossible car sur les ensembles $\{z : \sum_{i=1}^n x_i < n\theta_0 - \sqrt{n} z_{1-\alpha}\}$ et $\{z : \sum_{i=1}^n x_i > n\theta_0 + \sqrt{n} z_{1-\alpha}\}$, les deux fonctions ϕ_+ et ϕ_- diffèrent.

Il ressort de cette discussion que la notion de tests U.P.P. doit être affaiblie si on veut traiter le problème de comparaison de tests de façon un peu plus générique. Une solution est d'introduire les vitesses de séparation (voir par exemple le livre "Testing Statistical Hypotheses" de E.H. Lehmann et J.P. Romano, Chapitre 12).

Deuxième partie

Statistiques asymptotiques

Chapitre II-1

Introduction aux statistiques asymptotiques

Pour comparer deux estimateurs, construire un intervalle de confiance ou un test d'hypothèses, il est indispensable de connaître la distribution des statistiques sous-jacentes. Pour comparer deux estimateurs, nous devons par exemple calculer le risque des estimateurs, qui est égal à l'espérance de la perte sous la distribution des deux estimateurs. Pour construire un test d'hypothèses, nous devons connaître la distribution de la statistique de test pour pouvoir calculer les valeurs critiques. Dans certains cas, ces distributions sont connues de façon explicite (ce sont les exemples que nous avons traités dans les chapitres précédents). Mais dans de nombreuses situations, ces distributions sont impossibles à calculer explicitement. Nous ne disposons que d'approximations de ces lois, qui sont valables lorsque la taille de l'échantillon est grande.

Les notions principales sont celles de suite d'expériences statistiques, de suite d'estimateurs consistants, de normalité asymptotique, de variance asymptotique d'une suite d'estimateurs, de fonctions pivotales asymptotiques, de suite de tests consistante, et la construction de régions de confiance asymptotiques et de tests asymptotiques.

Commençons par un exemple de construction de tests d'hypothèses. Soit $Z = (X_1, \dots, X_n)$ un n-échantillon d'une famille paramétrique

$$\left(\mathbb{R}, \mathscr{B}(\mathbb{R}), \left\{\mathbb{P}_\theta\,,\ \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\right\}\right)\ .$$

Supposons que pour tout $\theta \in \mathbb{R} \times \mathbb{R}_+^*$, les composantes (μ, σ^2) collectent l'espérance et la variance de la loi $\mathbb{P}_\theta$ i.e.

$$\int x_1\ \mathbb{P}_\theta(\mathrm{d}x_1) = \mu \quad \text{et} \quad \int x_1^2\ \mathbb{P}_\theta(\mathrm{d}x_1) = \mu^2 + \sigma^2\ .$$

Nous effectuons un test sur l'espérance :

$$H_0 :\ \mu = \mu_0, \quad \text{contre} \quad H_1 :\ \mu \neq \mu_0\ ,$$

où $\mu_0 \in \mathbb{R}$ est une valeur donnée.

Nous avons déjà étudié ce test dans l'Exemple I-3.23, dans le cas particulier du modèle gaussien $\mathbb{P}_\theta \equiv \mathrm{N}(\mu, \sigma^2)$; nous avons proposé à cette occasion la statistique de test donnée par

$$T_n(Z) := \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

où $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ et $S_n^2 := (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ sont la moyenne empirique et un estimateur de la variance. Nous avons aussi établi dans cet exemple que sous $\mathbb{P}_{\mu_0,\sigma^2}$, la statistique de test $T_n(Z)$ suit une loi

n	Normal	Exponentielle
5	0.122	0.19
10	0.082	0.14
15	0.070	0.11
20	0.065	0.10
25	0.062	0.09
50	0.056	0.07
100	0.053	0.06

TABLE II-1.1 – Taille du test de région de rejet $\{z : |T_n(z)| > 1.96\}$ pour des observations Z indépendantes de distribution gaussienne ou exponentielle.

de Student à $(n-1)$ degrés de liberté (voir Section IV-3.5 pour la définition et les propriétés de cette loi) – cette propriété utilise le fait que sous $\mathbb{P}_\theta$, les v.a. X_i sont gaussiennes (en plus d'être indépendantes).
Considérer un modèle gaussien est très restrictif ; l'avantage est que l'on est capable de déterminer la loi exacte de $T_n(Z)$ sous $\mathbb{P}_\theta$. Sans cette hypothèse de gaussianité (mais en gardant celle d'un n-échantillon d'un modèle statistique), lorsque le nombre d'observations n est "suffisamment" grand, nous pouvons montrer que pour tout $\theta \in \mathbb{R}\times\mathbb{R}_+^*$, la loi de la statistique T_n sous $\mathbb{P}_\theta$ est "approximativement" distribuée suivant une loi normale centrée et réduite. Il est bien entendu nécessaire de préciser le terme "approximativement" : nous utiliserons à cet effet les différentes notions de convergences qui sont rappelées dans le Chapitre IV-5. Plus précisément, nous pouvons établir que, pour tout $\theta \in \mathbb{R}\times\mathbb{R}_+^*$, T_n converge en loi vers une loi normale centrée réduite (voir Définition IV-5.24 et Exemple II-1.21), i.e. pour toute fonction h continue et bornée et $\theta \in \mathbb{R}\times\mathbb{R}_+^*$, nous avons

$$\int\cdots\int h(T_n(x_1,\ldots,x_n))\mathbb{P}_\theta^{\otimes n}(\mathrm{d}x_1,\cdots,\mathrm{d}x_n) \to_{n\to\infty} \int h(z)\frac{1}{\sqrt{2\pi}}\exp(-0.5z^2)\mathrm{d}z\,.$$

Nous pouvons exploiter la connaissance de la loi *asymptotique* de $T_n(Z)$ pour construire une fonction de test, et ce, sans spécifier la loi exacte des X_i sous $\mathbb{P}_\theta$.
Comme la loi limite admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, la convergence en loi implique aussi que, pour tout $a>0$ et $\theta \in \mathbb{R}\times\mathbb{R}_+^*$,

$$\lim_{n\to\infty}\mathbb{P}_{n,\theta}\left(T_n(Z)\in[-a,a]\right) = \Phi(a)-\Phi(-a)\,,$$

où Φ désigne ici la fonction de répartition d'une loi $\mathrm{N}(0,1)$. Pour $\alpha\in]0,1[$, nous pouvons choisir a_α de telle sorte que $\Phi(a_\alpha)-\Phi(-a_\alpha)=1-\alpha$ ce qui conduit à $a_\alpha = z_{1-\alpha/2}$ – le quantile d'ordre $1-\alpha/2$ d'une loi $\mathrm{N}(0,1)$. Ainsi, pour tout $\sigma^2\in\mathbb{R}_+^*$, nous avons

$$\lim_{n\to\infty}\mathbb{P}_{n,\theta}\left(|T_n(Z)|\geq a_\alpha\right) = \alpha\,.$$

Nous sommes ainsi capables de construire un test de niveau *asymptotiquement* égal à α, sous des hypothèses très générales, en utilisant la notion de convergence en loi pour justifier la définition du seuil critique du test à partir de la loi asymptotique de $T_n(Z)$ à défaut de savoir calculer sa loi exacte sous $\mathbb{P}_{n,\theta}$.

La table II-1 donne la taille du test $\phi_n(Z)=\mathbb{1}\{|T_n(Z)|>z_{1-\alpha/2}\}$ pour $\alpha=0.05$ (dans ce cas $z_{1-\alpha/2}\approx 1.96$) lorsque la loi $\{\mathbb{P}_\theta\,,\ \theta\in\mathbb{R}\times\mathbb{R}_+\}$ est soit gaussienne, soit exponentielle. Nous voyons que pour $n\geq 20$, l'approximation est déjà très satisfaisante dans le cas gaussien ; lorsque la distribution des observations est exponentielle, l'approximation n'est satisfaisante que pour $n\geq 50$.

Notations : suite d'expériences statistiques

Dans cette partie II, nous allons déterminer des propriétés des estimateurs et des tests lorsque le nombre d'observations n tend vers l'infini. Pour ce faire, nous introduisons le formalisme suivant : nous définissons une *suite d'expériences statistiques* indexée par le nombre d'échantillons $n\in\mathbb{N}$: $(\mathsf{Z}_n,\mathscr{Z}_n,\{\mathbb{P}_{n,\theta},\theta\in\Theta\})$.

Dans la plupart des cas traités dans ce cours, nous étudions des suites d'expériences statistiques produits : pour tout $n \in \mathbb{N}$,

$$\mathsf{Z}_n := \mathsf{X}^n \,, \quad \mathscr{Z}_n := \mathscr{X}^{\otimes n} \,, \quad \mathbb{P}_{n,\theta} := \mathbb{Q}_\theta^{\otimes n} \,,$$

où $\{\mathbb{Q}_\theta, \theta \in \Theta\}$ est une famille de lois paramétrique sur $(\mathsf{X}, \mathscr{X})$. On définit sur chaque espace les statistiques canoniques $X_i : \mathsf{X}^n \to \mathsf{X}$ par $X_i(x_1, \cdots, x_n) = x_i$ – pour tout $i \geq 1$; dans ces expériences "produits", ces statistiques sont indépendantes et de loi $\mathbb{Q}_\theta$ sous $\mathbb{P}_{n,\theta}$ (c'est le sens de la loi produit $\mathbb{P}_{n,\theta} = \mathbb{Q}_\theta^{\otimes n}$).

II-1.1 Consistance d'une suite d'estimateurs

Etant donné un estimateur $T_n(Z_n)$ de la quantité inconnue $g(\theta)$, construit à partir de l'expérience Z_n, la propriété de *consistance* exprime que l'estimateur est d'autant meilleur que n est grand. La qualité de l'approximation est ici définie par la convergence en loi vers zero de l'erreur $\|T_n(Z_n) - g(\theta)\|$ sous $\mathbb{P}_{n,\theta}$; pour un vecteur $x \in \mathbb{R}^\ell$, $\|x\|$ est la norme euclidienne.

Définition II-1.1 (Suite d'estimateurs consistant). *Considérons une suite d'expériences statistiques paramétriques,*

$$(\mathsf{Z}_n, \mathscr{Z}_n, \{\mathbb{P}_{n,\theta} \,:\, \theta \in \Theta\}) \,, \quad \textit{où } \Theta \subseteq \mathbb{R}^d.$$

Soit $g : \Theta \to \mathbb{R}^\ell$ une fonction.

Une suite d'estimateurs $\{T_n,\ n \in \mathbb{N}^*\}$ où pour tout $n \in \mathbb{N}^*$

$$T_n : (\mathsf{Z}_n, \mathscr{Z}_n) \to (\mathbb{R}^\ell, \mathscr{B}(\mathbb{R}^\ell))$$

est dite consistante *pour g si pour tout $\theta \in \Theta$ et $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta} \left(\|T_n(Z_n) - g(\theta)\| \geq \varepsilon \right) = 0 \,.$$

La consistance n'est pas une propriété très forte, car elle ne permet pas de quantifier la vitesse à laquelle la suite d'estimateurs converge vers sa limite.

C'est une propriété satisfaite par de très nombreux estimateurs. Dès lors que T_n est une moyenne empirique $T_n(Z_n) = n^{-1} \sum_{i=1}^n T(X_i)$, la loi des grands nombres joue directement ou indirectement, un rôle important pour établir la propriété de consistance. En effet, elle montre que (voir Théorème IV-5.18) si pour tout $\theta \in \Theta$, $\mathbb{E}_{n,\theta}[|T(X_1)|] := \int |x_1| \mathbb{P}_\theta(\mathrm{d}x_1) < \infty$ et que sous $\mathbb{P}_{n,\theta}$, les v.a. $Z_n = (X_1, \cdots, X_n)$ sont i.i.d. de loi $\mathbb{P}_\theta$, alors

$$T_n(Z_n) := n^{-1} \sum_{i=1}^n T(X_i) \,, \tag{II-1.1}$$

est une une suite consistance d'estimateurs de $g : \theta \mapsto g(\theta) := \mathbb{E}_\theta[T(X_1)] = \int T(x_1) \mathbb{P}_\theta(\mathrm{d}x_1)$.

L'exemple II-1.2 établit la consistance d'un estimateur en utilisant la loi des grands nombres. L'exemple II-1.3 considère un exemple où l'estimateur $\hat{\theta}_n$ n'a pas de forme additive ; la consistance est établie par un calcul exact de la probabilité de l'événement $\{|\hat{\theta}_n - g(\theta)| \geq \varepsilon\}$.

Exemple II-1.2 (Paramètre d'une loi exponentielle). Soit la suite de modèles statistiques

$$\left(\mathbb{R}_+^n, \mathscr{B}(\mathbb{R}_+^n), \{\mathrm{Expo}(\theta)^{\otimes n},\ \theta \in \Theta := \mathbb{R}_+^*\}\right) \,.$$

Rappellons que pour $\theta \in \mathbb{R}_+^*$, la loi exponentielle $\mathrm{Expo}(\theta)$ admet une densité par rapport à la mesure de Lebesgue donnée par

$$p_\theta(x) = \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x) \,.$$

Pour tout $n \in \mathbb{N}^*$, l'estimateur du maximum de vraisemblance du paramètre θ est donné par

$$\hat{\theta}_n := \bar{X}_n^{-1} \quad \text{où} \quad \bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \,.$$

Notons que $\mathbb{E}_{n,\theta}[X_1] = \theta^{-1}$ et $\mathrm{Var}_{n,\theta}(X_1) = \theta^{-2}$. Le corollaire IV-5.17 montre que

$$\bar{X}_n \overset{\mathbb{P}_{n,\theta}-\text{prob}}{\longrightarrow} \theta^{-1} .$$

Considérons la fonction $g : \theta \mapsto \theta^{-1}$. Cette fonction est continue sur $\mathbb{R}_+^*$. Le théorème IV-5.6 montre que $g(\bar{X}_n) = \bar{X}_n^{-1} \overset{\mathbb{P}_{n,\theta}-\text{prob}}{\longrightarrow} \theta$. Par conséquence, $\{\bar{X}_n^{-1},\, n \in \mathbb{N}^*\}$ est une suite consistante pour θ. ♢

Exemple II-1.3. Soit une suite d'expériences statistiques produits du modèle uniforme

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{\mathrm{Unif}^{\otimes n}([0,\theta])\,,\ \theta \in \Theta := \mathbb{R}_+^\star\right\}\right).$$

Pour tout $n \in \mathbb{N}^*$, nous considérons l'estimateur défini par

$$\hat{\theta}_n := X_{n:n} = \max(X_1, \ldots, X_n) .$$

Par définition, sous $\mathbb{P}_{n,\theta}$, $X_{n:n} \leq \theta : \mathbb{P}_{n,\theta}(\hat{\theta}_n \leq \theta) = 1$. Par suite, pour tout $0 < \varepsilon < \theta$, nous avons

$$\begin{aligned}\mathbb{P}_{n,\theta}(|\hat{\theta}_n - \theta| > \varepsilon) &= \mathbb{P}_{n,\theta}(\hat{\theta}_n < \theta - \varepsilon) \\ &= \mathbb{P}_{n,\theta}(X_1 < \theta - \varepsilon,\ \ldots,\ X_n < \theta - \varepsilon) = \prod_{i=1}^n \mathbb{P}_{n,\theta}(X_i < \theta - \varepsilon) = (1 - \varepsilon/\theta)^n .\end{aligned}$$

Par conséquent, pour tout $\theta > 0$ and $0 < \varepsilon < \theta$,

$$\lim_n \mathbb{P}_{n,\theta}(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

ce qui établit que la suite $\{\hat{\theta}_n,\ n \in \mathbb{N}\}$ est une suite consistante d'estimateurs de θ. ♢

II-1.2 Normalité Asymptotique

La consistance d'une suite d'estimateurs ne permet pas de quantifier la vitesse à laquelle une suite d'estimateurs converge vers sa limite. En pratique, il est important de savoir quantifier l'erreur que l'on commet en approchant $g(\theta)$ par $T_n(Z_n)$. Idéalement, nous voudrions disposer de la distribution de $T_n(Z_n) - g(\theta)$ sous $\mathbb{P}_{n,\theta}$, pour pouvoir calculer la probabilité $\mathbb{P}_{n,\theta}(\|T_n(Z_n) - g(\theta)\| \geq \delta)$ par exemple. Cet objectif est en général trop ambitieux. En revanche, dans de nombreuses situations, la suite de variables $n^{1/2}\{T_n(Z_n) - g(\theta)\}$ converge vers une distribution non dégénérée, le plus souvent gaussienne, c'est pourquoi nous introduisons la définition suivante. Dans l'exemple II-1.18, nous verrons un cas où la loi limite n'est pas gaussienne.

Définition II-1.4 (Normalité asymptotique). *Considérons une suite d'expériences statistiques paramétriques,*

$$(\mathrm{Z}_n, \mathscr{Z}_n, \left\{\mathbb{P}_{n,\theta} :\ \theta \in \Theta\right\}) , \quad \textit{où } \Theta \subseteq \mathbb{R}^d.$$

Soit $g : \Theta \to \mathbb{R}^\ell$ une fonction. La suite d'estimateurs $\{T_n(Z_n), n \geq\}$ est asymptotiquement normale pour $g : \theta \to g(\theta)$ si, pour tout $\theta \in \Theta$,

$$\sqrt{n}\{T_n(Z_n) - g(\theta)\} \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \Gamma(\theta)) ,$$

où pour tout $\theta \in \Theta$, $\Gamma(\theta)$ est une matrice symétrique et positive dite matrice de variance-covariance asymptotique *de l'estimateur $T_n(Z_n)$ (ou simplement* variance asymptotique *dans le cas scalaire).*

Puisque $1/\sqrt{n} \to 0$, le lemme de Slutsky (Lemme IV-5.33) entraine que pour tout $\theta \in \Theta$ et pour tout $\varepsilon > 0$,

$$\lim_n \mathbb{P}_{n,\theta}\left(\|T_n(Z_n) - g(\theta)\| \geq \varepsilon\right) = 0 .$$

Par conséquent, une suite d'estimateurs asymptotiquement normale est nécessairement consistante.

II-1.2.1 Exemples

Exemple II-1.5. Considérons une suite d'expériences statistiques produit

$$(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_{n,\theta} = \mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\}) , \quad \text{où } \Theta \subseteq \mathbb{R}^d.$$

Supposons que pour tout $\theta \in \Theta$,

$$\mathbb{E}_\theta[\|X_1\|^2] = \int \|x_1\|^2 \, \mathbb{P}_\theta(\mathrm{d}x_1) < \infty,$$

et que la matrice de covariance existe :

$$\Sigma(\theta) := \mathrm{Cov}_\theta(X_1) = \int \{x_1 - \mu(\theta)\}\{x_1 - \mu(\theta)\}^T \mathbb{P}_\theta(\mathrm{d}x_1) \quad \text{où} \quad \mu(\theta) := \int x_1 \mathbb{P}_\theta(\mathrm{d}x_1) .$$

Pour tout $n \in \mathbb{N}^*$, posons

$$T_n(X_1, \ldots, X_n) := n^{-1} \sum_{i=1}^n X_i .$$

La normalité asymptotique de la suite $\{T_n(Z_n), n \geq 0\}$ découle du Théorème de la Limite Centrale ou T.L.C. (voir Théorème IV-5.39). Nous avons en effet, pour tout $\theta \in \Theta$,

$$\sqrt{n}\{\bar{X}_n - \mu(\theta)\} \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \Sigma(\theta)) .$$

Par conséquent, la suite d'estimateurs $\{T_n(Z_n), n \geq 0\}$ est asymptotiquement normale, avec pour moyenne $\mu : \theta \to \mu(\theta) := \mathbb{E}_\theta[X_1]$. La convergence vers la loi normale est illustrée dans les figures II-1.1, II-1.2 pour une distribution exponentielle et dans les figures II-1.3 II-1.4 pour une distribution de Student à 3 degrés de liberté (voir Section IV-3.5). Chacune de ces figures montrent quatre analyses : en haut à gauche, on représente un histogramme de n observations X_i – cet histogramme est donc une approximation de la loi exponentielle ou de la loi de Student ; en bas à gauche, on représente l'histogramme de N réalisations indépendantes de $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ avec $\mu = \sigma = 1$ dans le cas exponentiel et $\mu = 0, \sigma = \sqrt{3}$ dans le cas Student – cet histogramme est une approximation de la densité d'une loi $\mathrm{N}(0,1)$ d'après le TLC ; en haut à droite, on représente la fonction de répartition des N réalisations indépendantes ; en bas à droite, on présenter une dernière visualisation de la proximité à la loi $\mathrm{N}(0,1)$ en traçant une analyse quantile-quantile. Ces analyses montrent que la proximité à la loi gaussienne centrée réduite est d'autant meilleure que N est grand. ♢

La plupart des estimateurs que nous rencontrerons dans la suite de cet exposé seront asymptotiquement normaux : bien entendu, dans tous les cas, le T.L.C. joue un rôle essentiel, parfois de façon indirecte comme nous le verrons dans le Chapitre II-2.

Exemple II-1.6 (Paramètre d'une loi exponentielle (suite de Exemple II-1.2)). Pour tout $n \in \mathbb{N}^*$, l'estimateur du maximum de vraisemblance est donné par

$$T_n(X_1, \ldots, X_n) := \bar{X}_n^{-1} \quad \text{où} \quad \bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i .$$

En utilisant (IV-3.5) avec $p = 0$ et $r = 1, 2$, on montre que

$$\mathbb{E}_\theta[X_1] = \frac{1}{\theta} \quad \mathrm{Var}_\theta(X_1) = \frac{1}{\theta^2} .$$

Le théorème de la limite centrale (Théorème IV-5.39) montre que, pour tout $\theta \in \Theta$,

$$n^{-1/2} \sum_{i=1}^n (X_i - 1/\theta) = n^{1/2}(\bar{X}_n - 1/\theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, 1/\theta^2) .$$

Nous avons vu qu'une suite d'estimateurs consistante de θ était donnée par $\{h(\bar{X}_n), n \geq 0\}$ où $h(u) = 1/u$. Un résultat, connu sous le nom de *δ-méthode*, (voir Théorème IV-5.47 et Exemple IV-5.48) donne des conditions pour passer d'un T.L.C. pour des estimateurs $T_n(Z_n)$ à un T.L.C. pour des fonctions de cette suite $h(T_n(Z_n))$. Applicable dans le contexte de cet exemple, il permet d'affirmer que pour tout $\theta \in \Theta$,

$$n^{1/2} \{h(\bar{X}_n) - h(1/\theta)\} \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, [h'(1/\theta)]^2 \, \mathrm{Var}_\theta(X_1)) ,$$

soit :

$$n^{1/2}(\hat{\theta}_n - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \theta^2) . \qquad \diamondsuit$$

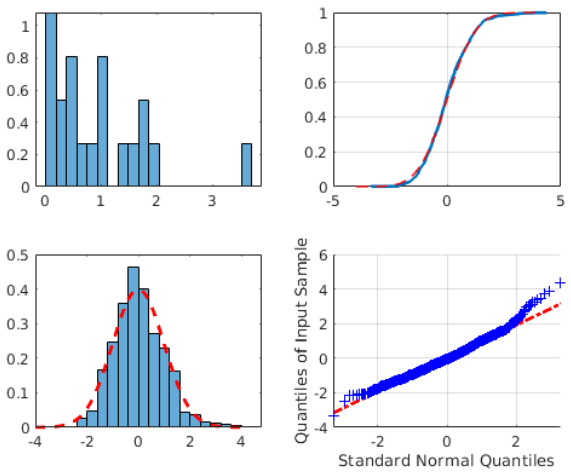

FIGURE II-1.1 – Loi exponentielle de paramètre 1 et $n = 20$. Figure en haut à gauche : histogramme de n échantillons X_i. Figure en bas à gauche : histogramme de $N = 1e3$ réalisations indépendantes de $\sqrt{n}(\bar{X}_n - 1)$. Figure en haut à droite : fonction de répartition empirique de $\sqrt{n}(\bar{X}_n - 1)$, obtenue avec N réalisations indépendantes. Figue en bas à droite : diagramme quantile-quantile de $\sqrt{n}(\bar{X}_n - 1)$ obtenu avec N réalisations indépendantes. En rouge, on superpose la densité de la loi $\mathrm{N}(0,1)$ en bas à gauche et sa fonction de répartition (en haut à gauche) ; en bas à droite, la droite d'équation $y = x$ est en rouge.

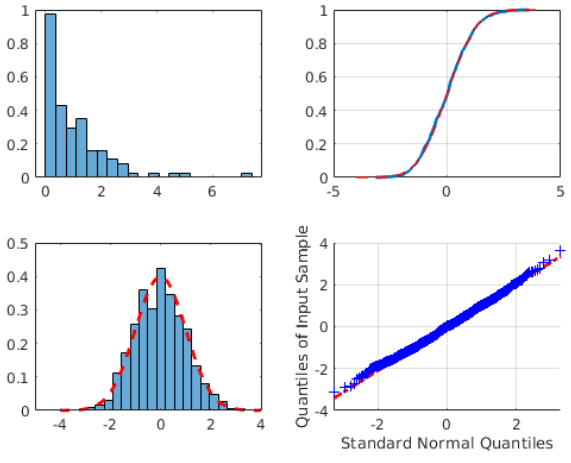

FIGURE II-1.2 – Loi exponentielle de paramètre 1, cas $n = 100$. Mêmes analyses que dans la fig. II-1.1

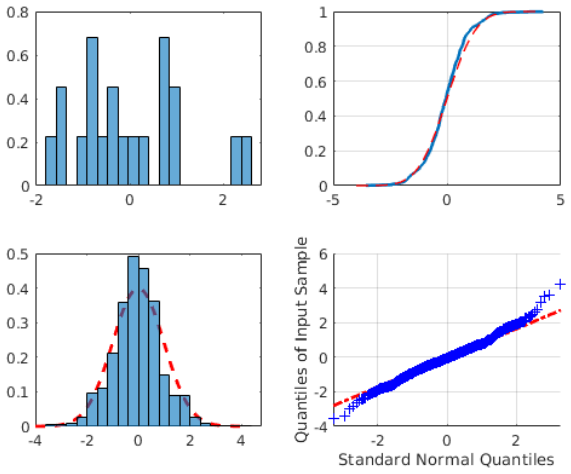

FIGURE II-1.3 – Loi de Student à 3 degrés de liberté, $n = 20$. Figure en haut à gauche : histogramme de n échantillons X_i. Figure en bas à gauche : histogramme de $N = 1e3$ réalisations indépendantes de $\sqrt{n}(\bar{X}_n)/\sqrt{3}$. Figure en haut à droite : fonction de répartition empirique de $\sqrt{n}(\bar{X}_n)/\sqrt{3}$, obtenue avec N réalisations indépendantes. Figue en bas à droite : diagramme quantile-quantile de $\sqrt{n}(\bar{X}_n)/\sqrt{3}$ obtenu avec N réalisations indépendantes. En rouge, on superpose la densité de la loi $\mathrm{N}(0,1)$ en bas à gauche et sa fonction de répartition (en haut à gauche) ; en bas à droite, la droite d'équation $y = x$ est en rouge.

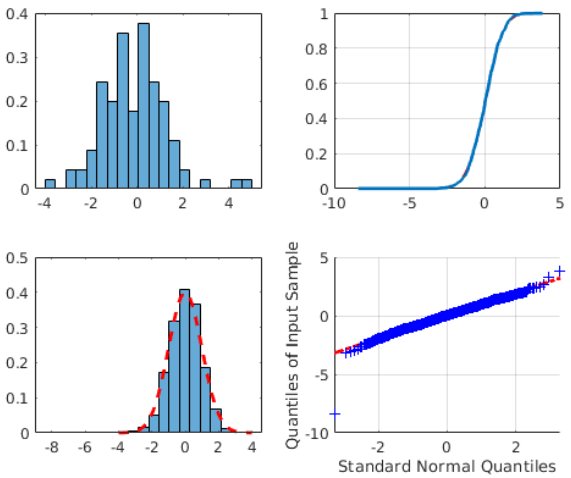

FIGURE II-1.4 – Loi de Student à 3 degrés de liberté, $n = 100$. Mêmes analyses qu'en fig. II-1.3

Exemple II-1.7 (Distribution limite pour le coefficient de régression). Dans ce chapitre introductif, nous nous intéressons exclusivement aux suites d'expériences statistiques produits (ce qui entraine que sous $\mathbb{P}_{n,\theta}$, les observations $(X_1,\cdots,X_n)$ sont indépendantes et identiquement distribuées). Toutefois, les différentes notions que nous avons introduites (consistance, normalité asymptotique) se généralisent de façon naturelle à des expériences statistiques plus générales. Nous allons étudier le modèle de régression linéaire à un facteur (il s'agit du modèle présenté dans l'exemple I-2.20 lorsque $\phi(x)=x$ et $p=1$; nous considérons ici un cas un peu plus général dans lequel les résidus ne sont pas nécessairement gaussiens mais distribués selon une loi $h\cdot \mathrm{d}\lambda_{\mathrm{Leb}}$). Soit h une densité par rapport à la mesure de Lebesgue sur $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. Supposons que

$$\int x^2 h(x)\lambda_{\mathrm{Leb}}(\mathrm{d}x) = 1 \quad \int x h(x)\lambda_{\mathrm{Leb}}(\mathrm{d}x) = 0\,.$$

Considérons pour tout $n\in\mathbb{N}$, le modèle statistique

$$\left(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n),\left\{p_{n,\theta}\cdot\lambda_{\mathrm{Leb}}^{\otimes n},\ \theta=(\beta_0,\beta_1,\sigma^2)\in\mathbb{R}\times\mathbb{R}\times\mathbb{R}_+^*\right\}\right),$$

où, pour tout $\theta\in\Theta$, la densité $p_{n,\theta}$ sur $\mathbb{R}^n$ est donnée par

$$p_{n,\theta}(y_1,\ldots,y_n) = \sigma^{-n}\prod_{i=1}^n h\left(\sigma^{-1}\left\{y_i-\beta_0-\beta_1 x_i\right\}\right)\,.$$

Notons $(Y_1,\ldots,Y_n)$ les observations. L'expression de la densité jointe $p_{n,\theta}$ montre que les observations sont indépendantes (forme produit de la densité) mais ne sont pas identiquement distribuées : la loi de Y_i sous $\mathbb{P}_{n,\theta}$ est $y\mapsto h(\sigma^{-1}(y-\beta_0-\beta_1 x_i))$ et elle dépend de l'indice i via la *covariable* x_i. Plus précisément, sous $\mathbb{P}_{n,\theta}$, les v.a.

$$\varepsilon_i(\theta) := \sigma^{-1}\left(Y_i-\beta_0-\beta_1 x_i\right)$$

sont i.i.d. de loi $h(u)\cdot\lambda_{\mathrm{Leb}}(\mathrm{d}u)$.

L'estimateur des moindres carrés des paramètres (β_0,β_1) pour les n observations $(Y_1,\ldots,Y_n)$ est obtenu en minimisant la somme des carrés des erreurs

$$(\hat\beta_{n,0},\hat\beta_{n,1}) := \operatorname*{arg\,min}_{(\beta_0,\beta_1)\in\mathbb{R}\times\mathbb{R}}\sum_{i=1}^n (Y_i-\beta_0-\beta_1 x_i)^2\,.$$

Ce problème d'optimisation admet une solution explicite. Ainsi, l'estimateur $\hat\beta_{n,1}$ est donné par

$$\hat\beta_{n,1} := \frac{\sum_{i=1}^n (Y_i-\overline{Y}_n)(x_i-\bar{x}_n)}{\sum_{i=1}^n (x_i-\bar{x}_n)^2} \qquad \bar{x}_n := n^{-1}\sum_{i=1}^n x_i\,,\quad \overline{Y}_n := n^{-1}\sum_{i=1}^n Y_i\,.$$

Nous allons montrer que pour que la suite $\{\hat\beta_{n,1}, n\in\mathbb{N}\}$ est consistante et asympotiquement normale. Pour ce faire, nous écrivons que sous $\mathbb{P}_{n,\theta}$,

$$Y_i = \beta_0+\beta_1 x_i+\sigma\varepsilon_i(\theta)\,,\qquad \overline{Y}_n = \beta_0+\beta_1\bar{x}_n+\sigma n^{-1}\sum_{j=1}^n \varepsilon_j(\theta)$$

ce qui entraine que

$$Y_i-\overline{Y}_n-\beta_1(x_i-\bar{x}_n) = \sigma\varepsilon_i(\theta)-\sigma n^{-1}\sum_{j=1}^n\varepsilon_j(\theta)\,;$$

par suite, sous $\mathbb{P}_{n,\theta}$, nous pouvons écrire

$$\sigma^{-1}\{\hat\beta_{n,1}-\beta_1\} = \frac{\sigma^{-1}\sum_{i=1}^n\left(Y_i-\overline{Y}_n-\beta_1(x_i-\bar{x}_n)\right)c_{n,i}}{\sum_{i=1}^n c_{n,i}^2} = \frac{\sum_{i=1}^n\varepsilon_i(\theta)c_{n,i}}{\sum_{i=1}^n c_{n,i}^2}$$

où nous avons posé $c_{n,i}:=(x_i-\bar{x}_n)$ et utilisé que $\sum_{i=1}^n c_{n,i}=0$. Les hypothèses faites sur la fonction h et la forme produit de la loi $\mathbb{P}_{n,\theta}$ entrainent que sous $\mathbb{P}_{n,\theta}$, les v.a. $\{\varepsilon_i(\theta), i=1,\cdots,n\}$ sont indépendantes, centrées et de variance 1. Par conséquent, en utilisant la condition de Hajek-Sidak (Exemple IV-5.43), nous obtenons

$$\sqrt{\sum_{i=1}^n (x_i-\bar{x}_n)^2}\,\frac{\hat\beta_{n,1}-\beta_1}{\sigma} = \frac{\sum_{i=1}^n\varepsilon_i(\theta)c_{ni}}{\sqrt{\sum_{i=1}^n c_{ni}^2}} \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1)\,,$$

à condition que

$$\lim_{n\to\infty} \max_{1\le i\le n} \frac{(x_i-\overline{x}_n)^2}{\sum_{j=1}^n (x_j-\overline{x}_n)^2} = 0\,.$$

Cette condition est satisfaite par de très nombreux plans d'échantillonnage, par exemple lorsque $x_i = i/n$ pour tout $i \in \{1,\dots,n\}$. ♢

Nous allons maintenant donner un exemple d'estimateur consistant dont la loi limite (après normalisation), n'est pas asymptotiquement normale.

Exemple II-1.8 (estimation du support d'une loi uniforme). Soit une suite d'expériences statistiques produit

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{\mathrm{Unif}^{\otimes n}([0,\theta])\,,\ \theta\in\Theta := \mathbb{R}_+^*\right\}\right)\,.$$

Pour tout $n\in\mathbb{N}^*$, considérons l'estimateur $\hat{\theta}_n := X_{n:n} = \max(X_1,\dots,X_n)$. Nous avons montré que la suite $\{\hat{\theta}_n,\ n\in\mathbb{N}^*\}$ est une suite consistante d'estimateurs de θ (voir exemple II-1.3). Démontrons une convergence en loi de l'estimateur correctement normalisée ; la convergence en loi est ici établie en utilisant la convergence des fonctions de répartition (voir théorème IV-5.26). Pour tout $x\ge 0$ et $\theta\in\Theta$, nous avons

$$\begin{aligned}\mathbb{P}_{n,\theta}(n(\theta-\hat{\theta}_n)\ge x) &= \mathbb{P}_{n,\theta}(\hat{\theta}_n \le \theta - x/n) = \mathbb{P}_{n,\theta}\left(X_1\le\theta-x/n,\dots,X_n-x/n\right)\\ &= \prod_{i=1}^n \mathbb{P}_{n,\theta}(X_i\le\theta-x/n) = (1-x/(n\theta))^n\end{aligned}$$

ce qui montre que

$$\lim_{n\to\infty}\mathbb{P}_{n,\theta}(n(\theta-\hat{\theta}_n)\le x) = 1-\mathrm{e}^{-x/\theta}\,.$$

On reconnaît la fonction de répartition d'une loi exponentielle de paramètre θ. Nous avons donc établi que, pour tout $\theta\in\Theta$, la suite de variables aléatoires $n(\theta-\hat{\theta}_n)$ converge en loi vers une loi exponentielle de paramètre θ. Il faut faire attention toutefois qu'il est assez rare d'être en mesure d'identifier explicitement la limite de la fonction de répartition. En ce sens, cet exemple est assez atypique. ♢

II-1.2.2 Etude asymptotique des estimateurs de moments

Nous avons introduit les estimateurs des moments dans la section I-2.1. Nous allons dans cette section établir leur normalité asymptotique. Rappelons tout d'abord brièvement les éléments principaux de cette construction.

Soit une suite d'expériences statistiques produit

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta\in\Theta\subseteq\mathbb{R}^d\right\}\right)\,.$$

La méthode des moments consiste à choisir d statistiques $\{T_j, j=1,\cdots,d\}$ à valeur réelle et telles que $\mathbb{E}_\theta[|T_j(X_1)|]<\infty$; et à estimer le paramètre θ en résolvant le système d'équations de d équations à d inconnues :

$$\theta\in\Theta \qquad \text{tel que} \quad \frac{1}{n}\sum_{i=1}^n T_j(X_i) = \mathbb{E}_\theta[T_j(X_1)]\,, \qquad j=1,\ \dots,\ d,$$

Supposons que ce système d'équations admette pour tout $n\in\mathbb{N}^*$ une solution unique notée $\hat{\theta}_n$, appelée *estimateur des moments*. Sauf dans des cas particuliers, les estimateurs des moments ne sont pas les "meilleurs" estimateurs, mais par contre, sous des hypothèses assez générales, ils sont consistants et sont asymptotiquement normaux. Ces propriétés découlent de la loi des grands nombres (Théorème IV-5.21), du T.L.C. (Théorème IV-5.39) et de la méthode-delta (Théorème IV-5.47).

Pour simplifier la présentation des résultats, notons $\mathbf{T} := (T_1,\ \dots,\ T_d) : \mathsf{X}^n \to \mathbb{R}^d$ et $\mathbf{e} : \Theta \mapsto \mathbb{R}^d$ la fonction à valeurs vectorielles

$$\mathbf{e}(\theta) := \mathbb{E}_\theta[\mathbf{T}(X_1)] = \begin{bmatrix}\mathbb{E}_\theta[T_1(X_1)]\\ \cdots \\ \mathbb{E}_\theta[T_d(X_1)]\end{bmatrix}\,.$$

Avec ces notations, l'estimateur des moments $\hat{\theta}_n$ est la solution du système d'équations

$$\mathbf{T}_n(Z_n) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}(X_i) = \mathbf{e}(\theta) \; ;$$

$Z_n = (X_1, \cdots, X_n)$ est l'observation. Une condition nécessaire pour que l'estimateur des moments soit défini est que l'estimateur empirique des moments $\mathbf{T}_n(Z_n)$ soit un élément de l'image de Θ par $\mathbf{e}$, notée $\mathbf{e}(\Theta)$. Si $\mathbf{e} : \Theta \mapsto \mathbf{e}(\Theta)$ est une bijection de Θ sur $\mathbf{e}(\Theta)$, alors l'estimateur des moments est défini de façon unique par

$$\hat{\theta}_n = \mathbf{e}^{-1}\left(\mathbf{T}_n(Z_n)\right) \; . \tag{II-1.2}$$

Par suite, pour tout $\theta_0 \in \Theta$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}\{\mathbf{e}^{-1}\left(\mathbf{T}_n(Z_n)\right) - \theta_0\} \; .$$

Si l'estimateur des moments empiriques est asymptotiquement normal, alors

$$\sqrt{n}\{\mathbf{T}_n(Z_n) - \mathbf{e}(\theta_0)\} \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}(0, \Gamma(\theta_0)) \; ;$$

on en déduit alors l'estimateur des moments est asymptotiquement normal par application de la méthode-delta (Théorème IV-5.47), dès que la fonction $\mathbf{e}^{-1}$ est différentiable au point $\mathbf{e}(\theta_0)$. En utilisant ce résultat nous pouvons maintenant donner un énoncé général pour l'existence, la consistance et la normalité asymptotique d'un estimateur des moments.

Théorème II-1.9. *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n}, \; \theta \in \Theta\right\}\right) \; ,$$

où Θ est un ouvert de $\mathbb{R}^d$. Etant données d statistiques $T_j : \mathsf{X} \to \mathbb{R}$, pour $j = 1, \cdots, d$, posons $\mathbf{T} := (T_1, \ldots, T_d)^\top : \mathsf{X}^n \to \mathbb{R}^d$. Supposons que

(i) pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[\|\mathbf{T}(X_1)\|^2] < \infty$.

(ii) $\mathbf{e} : \Theta \to \mathbb{R}^d$ définie par $\mathbf{e}(\theta) := \mathbb{E}_\theta[\mathbf{T}(X_1)]$ est un difféomorphisme.

Pour tout $\theta_0 \in \Theta$,

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}\left(n^{-1}\sum_{i=1}^{n} \mathbf{T}(X_i) \in \mathbf{e}(\Theta)\right) = 1 \; .$$

De plus, l'estimateur des moments $\hat{\theta}_n$ solution de $n^{-1}\sum_{i=1}^n \mathbf{T}(X_i) = \mathbf{e}(\theta)$ est asymptotiquement normal :

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}(0, [\mathrm{J}_\mathbf{e}(\theta_0)]^{-1}\,\mathrm{Cov}_{\theta_0}(\mathbf{T}(X_1))[\mathrm{J}_\mathbf{e}(\theta_0)]^{-\top}) \; . \tag{II-1.3}$$

Comme démontré en section II-1.2, la normalité asymptotique de la suite d'estimateurs entraine sa consistance.

Démonstration. Soit $\theta_0 \in \Theta$. Par la loi des grands nombres,

$$n^{-1}\sum_{i=1}^{n} \mathbf{T}(X_i) \stackrel{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \mathbb{E}_{\theta_0}[\mathbf{T}(X_1)] \tag{II-1.4}$$

ce qui implique que

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}\left(n^{-1}\sum_{i=1}^{n} \mathbf{T}(X_i) \in \mathbf{e}(\Theta)\right) = 1 \; .$$

Si $n^{-1}\sum_{i=1}^n \mathbf{T}(X_i) \in \mathbf{e}(\Theta)$ alors $\hat{\theta}_n$ est défini de façon unique et est donné par $\hat{\theta}_n = \mathbf{e}^{-1}(n^{-1}\sum_{i=1}^n \mathbf{T}(X_i))$.

On peut prolonger de façon arbitraire la fonction $\mathbf{e}^{-1}$ à l'extérieur de $\mathbf{e}(\Theta)$ (en préservant bien entendu la mesurabilité !) Le résultat découle de l'application du théorème IV-5.6 et de la δ-méthode (théorème IV-5.47). □

Exemple II-1.10 (Loi exponentielle). Nous reprenons l'exemple I-2.2. Nous cherchons à comparer les estimateurs de moments associés aux statistiques $T(x) = x$ et $\tilde{T}(x) = x^2$. Ces estimateurs sont les solutions des équations

$$e(\theta) = \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \tilde{e}(\theta) = \frac{2}{\theta^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \,,$$

où les fonctions e et $\tilde{e}$ sont définies par :

$$e(\theta) = \mathbb{E}_\theta\left[T(X_1)\right] = \int_0^{+\infty} x\theta\exp(-\theta x)\mathrm{d}x = \frac{1}{\theta}$$
$$\tilde{e}(\theta) = \mathbb{E}_\theta\left[\tilde{T}(X_1)\right] = \int_0^{+\infty} x^2\theta\exp(-\theta x)\mathrm{d}x = \frac{2}{\theta^2}\,.$$

Dans les deux cas considérés, ces équations ont des solutions uniques, qui définissent donc les deux estimateurs de moments suivants :

$$\hat{\theta}_{n,1} := \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1}, \quad \text{et} \quad \hat{\theta}_{n,2} := \sqrt{2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{-1/2}. \tag{II-1.5}$$

Pour comparer ces estimateurs, nous appliquons le théorème II-1.9. Des calculs élémentaires montrent que

$$\mathrm{Var}_\theta\left(X_1\right) = \frac{1}{\theta^2} \qquad \mathrm{Var}_\theta\left(X_1^2\right) = \frac{20}{\theta^4}$$
$$e'(\theta) = -\frac{1}{\theta^2} \qquad \tilde{e}'(\theta) = -\frac{4}{\theta^3}$$

Par conséquent, les estimateurs $\{\hat{\theta}_{n,i}, n \in \mathbb{N}\}$, $i = 1,2$ sont consistants et asymptotiquement normaux. Les variances asymptotiques (Définition II-3.8) sont données par :

$$v(\theta) = e'(\theta)^{-2}\,\mathrm{Var}_\theta\left(X_1\right) = \theta^2$$

et

$$\tilde{v}(\theta) = \tilde{e}'(\theta)^{-2}\,\mathrm{Var}_\theta\left(X_1^2\right) = \frac{5}{4}\theta^2$$

respectivement. La variance asymptotique de l'estimateur $\hat{\theta}_{n,1}$ est inférieure à celle de $\hat{\theta}_{n,2}$ et de ce point de vue, $\hat{\theta}_{n,1}$ semble " préférable" à $\hat{\theta}_{n,2}$ (voir la règle à la définition II-3.8). ♢

Exemple II-1.11 (Loi de Cauchy). Reprenons l'exemple I-2.3. Soit une suite d'expériences statistiques produit

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{\mathrm{Cauchy}^{\otimes n}(\theta)\,,\ \theta \in \Theta = \mathbb{R}\right\}\right)\,.$$

On note $Y_i = \mathrm{signe}(X_i)$. On a

$$\mathbb{E}_\theta\left[\mathrm{signe}(X_1)\right] = \int \mathrm{signe}(x)p_\theta(x)\mathrm{d}x = 1 - 2F(-\theta)\,,$$

où q_θ est la densité d'une loi de Cauchy de paramètre de position (ou translation) θ et d'échelle 1, et F la fonction de répartition de même loi de Cauchy de paramètre de position 0 et d'échelle 1 :

$$F(t) = \frac{1}{\pi}\int_{-\infty}^{t} \frac{\mathrm{d}t}{1+t^2} = \frac{1}{\pi}\arctan(t) + \frac{1}{2}.$$

L'estimateur des moments associé à signe est donné par la valeur θ solution de

$$\frac{2}{\pi}\arctan(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i\,,$$

soit

$$\hat{\theta}_n := \tan\left(\frac{\pi}{2n}\sum_{i=1}^{n} Y_i\right)\,.$$

Par la loi des grands nombres, nous avons, pour tout $\theta_0 \in \Theta$,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \overset{\mathbb{P}_{n,\theta_0}-\mathrm{prob}}{\longrightarrow} 1 - 2F(-\theta_0) = \frac{2}{\pi}\arctan(\theta_0)\,.$$

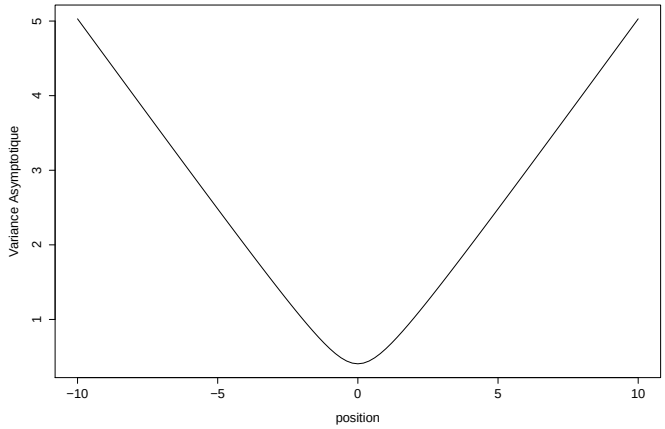

FIGURE II-1.5 – Variance asymptotique de l'estimateur des moments du paramètre de position d'une loi de Cauchy d'échelle 1

Le théorème IV-5.6 implique que $\hat{\theta}_n \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \theta_0$ pour tout $\theta_0 \in \Theta$. Comme

$$\begin{aligned}\mathrm{Var}_{\theta_0}(Y_1) &= \mathbb{E}_{\theta_0}[Y_1^2] - (\mathbb{E}_{\theta_0}[Y_1])^2 = 1 - \{1-2F(-\theta_0)\}^2 \\ &= 4F(-\theta_0)\{1-F(-\theta_0)\} = \{1+(2/\pi)\arctan(\theta_0)\}\{1-(2/\pi)\arctan(\theta_0)\}\,,\end{aligned}$$

le T.L.C. (Théorème IV-5.39) montre que

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - \mathbb{E}_{\theta_0}[Y_1]\} \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}(0, \{1+(2/\pi)\arctan(\theta_0)\}\{1-(2/\pi)\arctan(\theta_0)\})\,.$$

Comme nous avons

$$e'(\theta) = \frac{\pi}{2}\frac{1}{1+\theta^2},$$

le théorème II-1.9 montre que, pour tout $\theta_0 \in \mathbb{R}$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}\left(0, (2/\pi)^2(1+{\theta_0}^2)^2\{1+(2/\pi)\arctan(\theta_0)\}\{1-(2/\pi)\arctan(\theta_0)\}\right)\,. \qquad \diamondsuit$$

Nous avons représenté la variance asymptotique de l'estimateur $\hat{\theta}_n$ dans la fig. II-1.5.

Exemple II-1.12 (Modèle de translation et d'échelle). Soit h une densité sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda_{\mathrm{Leb}})$ vérifiant

$$\int x\,h(x)\mathrm{d}x = 0\,, \qquad m_2 := \int x^2 h(x)\mathrm{d}x > 0\,, \qquad m_4 := \int x^4 h(x)\mathrm{d}x < \infty\,.$$

Considérons la densité de probabilité

$$q_\theta(x) := \frac{1}{\sigma}h\left(\frac{x-\mu}{\sigma}\right)\,, \quad \theta := (\mu,\sigma^2) \in \Theta := \mathbb{R}\times\mathbb{R}_+^*\,. \tag{II-1.6}$$

Soit une suite d'expériences statistiques produit

$$(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n)\{q_\theta \cdot \mathrm{d}\lambda_{\mathrm{Leb}}\,,\ \theta\in\Theta\})\,.$$

Nous considèrons l'estimateur donné pour tout $n \in \mathbb{N}^*$ par $\hat{\theta}_n := (\hat{\mu}_n, \hat{\sigma}_n^2)$ où

$$\begin{cases}\hat{\mu}_n &:= \frac{1}{n}\sum_{i=1}^n X_i \\ \hat{\sigma}_n^2 &:= \frac{1}{nm_2}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2\,.\end{cases} \tag{II-1.7}$$

Nous allons utiliser le théorème II-1.9 pour établir la normalité asymptotique de cette suite d'estimateurs. Notons tout d'abord qu'en appliquant le T.L.C. (Théorème IV-5.39) nous avons, pour tout $\theta_0 = (\mu_0, \sigma_0^2) \in \Theta$,

$$\sqrt{n}\left(\begin{bmatrix} n^{-1}\sum_{i=1}^n X_i \\ n^{-1}\sum_{i=1}^n X_i^2 \end{bmatrix} - \begin{bmatrix} \mu_0 \\ \mu_0^2 + m_2\sigma_0^2 \end{bmatrix} \right) \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}\left(\mathbf{0}_{2\times 1}, \Sigma_{\theta_0}\right)$$
$$\Sigma_{\theta_0} := \begin{bmatrix} \alpha_2(\theta_0) - \alpha_1^2(\theta_0) & \alpha_3(\theta_0) - \alpha_1(\theta_0)\alpha_2(\theta_0) \\ \alpha_3(\theta_0) - \alpha_1(\theta_0)\alpha_2(\theta_0) & \alpha_4(\theta_0) - \alpha_2^2(\theta_0) \end{bmatrix}$$

où nous avons posé, pour $r = 1, \ldots, 4$,

$$\alpha_r(\theta_0) := \mathbb{E}_{\theta_0}[X_1^r] = \sum_{k=0}^{r} \binom{r}{k} \mu_0^k \sigma_0^{r-k} m_{r-k} \,, \quad m_j := \int x^j \, h(x) \mathrm{d}x \,. \qquad \diamondsuit$$

Une application directe du théorème II-1.9 montre que, pour tout $\theta_0 \in \Theta$, nous avons

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}(\mathbf{0}_{2\times 1}, [\mathrm{J}_{\mathbf{e}}(\theta_0)]^{-1} \Sigma_{\theta_0} [\mathrm{J}_{\mathbf{e}}(\theta_0)]^{-T}) \,.$$

II-1.3 Régions de confiance asymptotiques

Définition II-1.13 (Régions de confiance asymptotiques). *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n}, \ \theta \in \Theta\right\}\right) \,,$$

une fonction mesurable $g : \Theta \to \mathbb{R}^p$ et $\alpha \in]0,1[$. Une suite d'ensembles aléatoires

$$\{\mathscr{C}_n(X_1, \ldots, X_n), n \in \mathbb{N}\} \in \mathscr{B}(\mathbb{R}^p)$$

définit une suite de régions de confiance au niveau asymptotique $1-\alpha$ *pour la fonction g si*

$$\inf_{\theta\in\Theta} \liminf_{n\to\infty} \mathbb{P}_{n,\theta}\left(g(\theta) \in \mathscr{C}_n(X_1, \ldots, X_n)\right) \geq 1 - \alpha \,.$$

On prendra garde à l'ordre dans lequel on évalue la limite et l'infimum dans la définition précédente. Tout d'abord, pour $\theta \in \Theta$, on passe à la limite dans le nombre d'échantillons puis on calcule l'infimum sur l'ensemble des paramètres.

Les *fonctions asymptotiquement pivotales* sont utiles notamment la construction d'intervalles de confiance asymptotiques.

Définition II-1.14 (Fonctions asymptotiquement pivotales ou pivots asymptotiques). *Soit une suite de fonctions mesurables $\{G_n,\ n \in \mathbb{N}\}$ où pour tout $n \in \mathbb{N}$, $G_n : \mathsf{X}^n \times \Theta \to \mathbb{R}^p$. Cette suite est* asymptotiquement pivotale *si et seulement si, pour tout $\theta, \vartheta \in \Theta$ et $A \in \mathscr{B}(\mathbb{R}^p)$, on a*

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta}(G_n(X_1, \ldots, X_n, \theta) \in A) = \lim_{n\to\infty} \mathbb{P}_{n,\vartheta}(G_n(X_1, \ldots, X_n, \vartheta) \in A) \,.$$

Etant donnée une telle suite de fonctions, construisons des intervalles de confiance asymptotiques. Pour simplifier, nous travaillons dans le cas $p = 1$, et nous définissons la probabilité sur $(\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p))$: $\mu(A) := \lim_{n\to\infty} \mathbb{P}_{n,\theta}(G_n(X_1, \ldots, X_n, \theta) \in A)$ – qui ne dépend pas de θ. On choisit p_α et q_α de telle sorte que $\mu([p_\alpha, q_\alpha]) = 1 - \alpha$. Les ensembles

$$\mathscr{J}_n(\alpha, X_1, \ldots, X_n) := \{\theta \in \Theta : \ p_\alpha \leq G_n(X_1, \ldots, X_n; \theta) \leq q_\alpha\}$$

vérifient, pour tout $\theta \in \Theta$,

$$\lim_n \mathbb{P}_{n,\theta}\left(\theta \in \mathscr{J}_n(\alpha, X_1, \ldots, X_n)\right) = \lim_n \mathbb{P}_{n,\theta}\left(p_\alpha \leq G_n(X_1, \ldots, X_n; \theta) \leq q_\alpha\right) = 1 - \alpha \, .$$

C'est une suite de régions de confiance de *niveau asymptotique* $1-\alpha$.

II-1.3.1 Exemples

Exemple II-1.15 (Intervalle de confiance pour le paramètre d'une loi de Bernoulli). Soit une suite d'expériences statistiques produit d'un modèle de Bernoulli

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{\mathrm{Ber}^{\otimes n}(\theta),\ \theta \in \Theta :=]0,1[\right\}\right) \, ,$$

La suite d'estimateurs définie pour tout $n \geq 1$ par $\hat{\theta}_n := \bar{X}_n$ où $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ est asymptotiquement normale : pour tout $\theta \in]0,1[$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \theta(1-\theta)) \, .$$

Par suite, la suite de fonctions définies par :

$$G_n(X_1, \ldots, X_n; \theta) := \sqrt{n}\frac{(\bar{X}_n - \theta)}{\sqrt{\theta(1-\theta)}}$$

est une suite de fonctions asymptotiquement pivotales puisque pour tout borélien A, la convergence en loi entraine

$$\forall \theta \in \Theta, \qquad \mathbb{P}_{n,\theta}(G_n(X_1, \ldots, X_n; \theta) \in A) = \frac{1}{\sqrt{2\pi}} \int_A \exp(-0.5x^2)\, \mathrm{d}x \, .$$

On vérifie facilement que la suite de régions de confiance :

$$\left\{ \theta \in \Theta :\ -z_{1-\alpha/2} \leq \sqrt{n}\frac{(\bar{X}_n - \theta)}{\sqrt{\theta(1-\theta)}} \leq z_{1-\alpha/2} \right\}$$

construite à partir de ce pivot asymptotique G_n est une suite d'intervalles dont les limites sont données par :

$$\frac{\bar{X}_n + \frac{z^2_{1-\alpha/2}}{2n} \pm z^2_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + z^2_{1-\alpha/2}/n} \, . \tag{II-1.8}$$

Ce dernier intervalle de confiance est appelé *intervalle de Wilson*.

En utilisant le lemme de Slutsky (Lemme IV-5.33), il est facile de voir que la suite de fonctions définie par

$$\bar{G}_n(X_1, \ldots, X_n; \theta) := \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}$$

est aussi une suite de fonctions asymptotiquement pivotales. Les ensembles de confiance de probabilité de couverture asymptotiques $1-\alpha$ associés à cette fonction pivotale sont des intervalles dont les bornes sont cette fois-ci données par :

$$\bar{X}_n \pm z_{1-\alpha/2}\frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}. \tag{II-1.9}$$

♢

Cet intervalle de confiance a été proposé par Wald. L'intervalle donné par (II-1.8) est a priori plus satisfaisant que celui donné par (II-1.9) . Sur la fig. II-1.6 à la fig. II-1.8), nous représentons 50 intervalles de confiance de niveau asymptotique $1-\alpha = 0.95$ et obtenus selon les formules précédentes, à partir de 50 vecteurs d'observations $(X_1, \cdots, X_n)$ indépendants obtenus comme la réalisation de v.a. de Bernoulli de paramètre θ_0. On a pris $(n, \theta_0) = (100, 0.025)$; $(n, \theta_0) = (100, 0.25)$; et $(n, \theta_0) = (1000, 0.25)$.

La validité des intervalles de confiance dans les exemples précédents est déduite des lois limites de $\sqrt{n}(\bar{X}_n - \theta)$ et dépend donc de la qualité d'approximation de ces lois. Dans l'exemple des variables aléatoires de Bernoulli, $n\bar{X}_n$ suit une loi binomiale et la qualité d'approximation dépend du taux de succès θ, qui est inconnu. Lorsque θ est proche de 0 ou de 1, l'erreur d'approximation peut être grande et les procédures développées ci-dessus -et tout particulièrement l'intervalle de confiance donné par (II-1.9)- peuvent conduire à des résultats erronés, bien que l'intervalle donné par (II-1.8) soit en général satisfaisant.

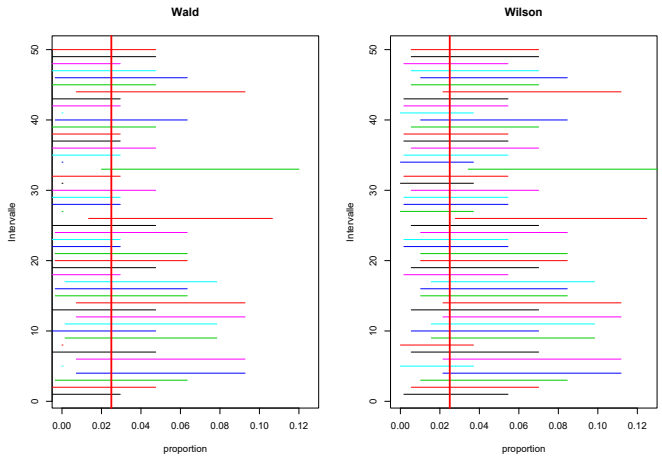

FIGURE II-1.6 – 50 intervalles de confiances de Wald et de Wilson de niveau de couverture de $1-\alpha = 0.95$ pour une proportion $\theta = 0.025$ et $n = 100$

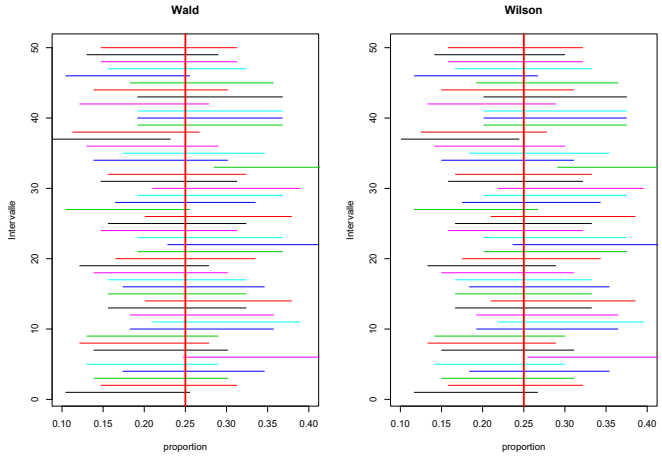

FIGURE II-1.7 – 50 intervalles de confiances de Wald et de Wilson de niveau de couverture de $1-\alpha = 0.95$ pour une proportion $\theta = 0.25$ et $n = 100$

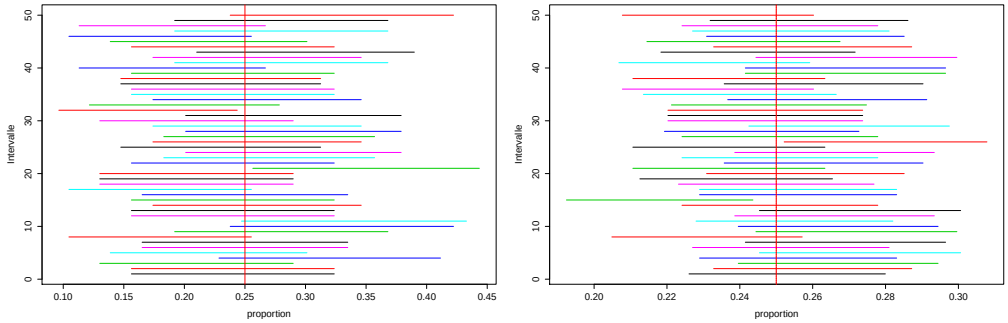

FIGURE II-1.8 – 50 intervalles de confiances de Wald de niveau de couverture de $1-\alpha = 0.95$ pour une proportion $\theta = 0.25$ et $n = 100$ (gauche) ou $n = 1000$ (droite). Observer l'échelle sur l'axe des x.

II-1.3.2 Transformations de stabilisation de la variance

Une des utilisations principales de la δ-méthode est la construction de transformations de stabilisation de la variance, une méthode que l'on utilise par exemple pour construire des intervalles de confiance asymptotiques. L'idée générale est la suivante.

Supposons que nous cherchions à construire un intervalle de confiance asymptotique pour le paramètre $\theta \in \mathbb{R}$ à l'aide de l'observation $Z_n = (X_1, \cdots, X_n)$; soit $T_n(Z_n)$ un estimateur de θ. Supposons que, pour tout $\theta \in \Theta$, $\sqrt{n}(T_n(Z_n) - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \sigma^2(\theta))$ et que $\theta \to \sigma(\theta)$ est continue et strictement positive.

• Une première approche consiste à combiner la propriété de normalité asymptotique de la suite d'estimateurs $\{T_n(Z_n), n \geq 1\}$ et le lemme de Slutsky.
Pour $\alpha \in]0,1[$ et $\theta \in \Theta$, nous avons donc

$$\mathbb{P}_{n,\theta}\left(T_n(Z_n) - z_{1-\frac{\alpha}{2}}\frac{\sigma(\theta)}{\sqrt{n}} \leq \theta \leq T_n(Z_n) + z_{1-\frac{\alpha}{2}}\frac{\sigma(\theta)}{\sqrt{n}}\right) \to_{n\to\infty} 1-\alpha,$$

où $z_{1-\alpha/2} := \Phi^{-1}(1-\alpha/2)$ désigne le quantile d'ordre $1-\alpha/2$ d'une gaussienne centrée réduite. Comme $T_n(Z_n) \overset{\mathbb{P}_{n,\theta}-\text{prob}}{\longrightarrow} \theta$ (rappelons en effet que la normalité asymptotique entraine la consistance) et que la fonction $\theta \mapsto \sigma(\theta)$ est continue, le théorème de continuité (Théorème IV-5.30) montre que $\sigma(T_n(Z_n)) \overset{\mathbb{P}_{n,\theta}-\text{prob}}{\longrightarrow} \sigma(\theta)$ pour tout $\theta \in \Theta$; le lemme de Slutsky (Lemme IV-5.33) montre que

$$\sqrt{n}\left(\frac{T_n(Z_n) - \theta}{\sigma(T_n(Z_n))}\right) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1)\,.$$

Par conséquent,

$$\lim_n \mathbb{P}_{n,\theta}\left(T_n(Z_n) - z_{1-\frac{\alpha}{2}}\frac{\sigma(T_n(Z_n))}{\sqrt{n}} \leq \theta \leq T_n(Z_n) + z_{1-\frac{\alpha}{2}}\frac{\sigma(T_n(Z_n))}{\sqrt{n}}\right) = 1-\alpha,$$

ce qui montre que

$$\left[T_n(Z_n) - z_{1-\frac{\alpha}{2}}\frac{\sigma(T_n(Z_n))}{\sqrt{n}}, T_n(Z_n) + z_{1-\frac{\alpha}{2}}\frac{\sigma(T_n(Z_n))}{\sqrt{n}}\right], \tag{II-1.10}$$

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

• La δ-méthode fournit une autre méthode de construction, qui évite d'estimer la variance asymptotique.
Si la fonction g est différentiable au point $\theta \in \Theta$ et $g'(\theta) \neq 0$, alors,

$$\sqrt{n}\{g(T_n(Z_n)) - g(\theta)\} \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,\ [g'(\theta)]^2\sigma^2(\theta))\,.$$

Par conséquent, si nous choisissons la fonction g de telle sorte que, pour tout $\theta \in \Theta$,

$$[g'(\theta)]^2\sigma^2(\theta) = k^2 \tag{II-1.11}$$

où $k > 0$ est une constante, alors $\sqrt{n}(g(T_n(Z_n)) - g(\theta)) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,\ k^2)$. Ceci nous permet alors de construire l'intervalle de confiance suivant pour $g(\theta)$, de niveau asymptotique $1-\alpha$,

$$\left[g(T_n(Z_n)) - z_{1-\frac{\alpha}{2}}\frac{k}{\sqrt{n}}, g(T_n(Z_n)) + z_{1-\frac{\alpha}{2}}\frac{k}{\sqrt{n}}\right].$$

En supposant que la fonction g est un difféomorphisme et que son inverse est monotone, nous en déduisons un intervalle de confiance pour θ :

$$\left[g^{-1}\left(g(T_n(Z_n)) \pm z_{1-\frac{\alpha}{2}}\frac{k}{\sqrt{n}}\right), g^{-1}\left(g(T_n(Z_n)) \pm z_{1-\frac{\alpha}{2}}\frac{k}{\sqrt{n}}\right)\right].$$

Cet intervalle de confiance est souvent préféré à (II-1.10) car il n'est pas nécessaire d'utiliser un estimateur de substitution de la variance asymptotique (Définition II-3.8). Il découle de (II-1.11) que

$$g(\theta) = k \int_{\theta_0}^{\theta} \frac{1}{\sigma(\theta)} \mathrm{d}\theta \; . \tag{II-1.12}$$

Bien que la δ-méthode soit applicable dans $\mathbb{R}^d$ pour $d > 1$, il est délicat de généraliser le concept de stabilisation de la variance en dimension $d > 1$. Sauf dans des cas très particuliers, il est impossible de construire une fonction g qui "stabilise" la covariance asymptotique en dimension $d > 1$.

Exemple II-1.16 (Loi Bernoulli). Soit une suite d'expériences statistiques produit

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{ \mathrm{Ber}^{\otimes n}(\theta) \, , \; \theta \in \Theta :=]0,1[\right\}\right) \; .$$

Pour tout $\theta \in \Theta$, le théorème de la limite centrale montre que

$$\sqrt{n}(\bar{X}_n - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \; \theta(1-\theta)) \; , \qquad \bar{X}_n := n^{-1} \sum_{i=1}^{n} X_i \; .$$

Posons $\sigma(\theta) := \sqrt{\theta(1-\theta)}$, qui définit une fonction σ continue sur Θ ; et prenons $k = 1/2$. En utilisant (II-1.12), une transformation de stabilisation de la variance est donnée par

$$g(\theta) = \int \frac{1/2}{\sqrt{\theta(1-\theta)}} \mathrm{d}\theta = \arcsin(\sqrt{\theta}) \; .$$

Nous avons, pour tout $\theta \in \Theta$,

$$\sqrt{n} \left(\arcsin\left(\sqrt{\bar{X}_n / n} \right) - \arcsin(\sqrt{\theta}) \right) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \; 1/4) \; .$$

Donc, pour tout $\alpha \in]0,1[$, si $\bar{X}_n \notin \{0,1\}$,

$$\mathscr{I}_{n,\alpha} := \left[\sin^2(\arcsin(\sqrt{\bar{X}_n/n}) - z_{1-\alpha/2}/2\sqrt{n}), \sin^2(\arcsin(\sqrt{\bar{X}_n/n}) + z_{1-\alpha/2}/2\sqrt{n}) \right] \; ,$$

définit un intervalle de confiance de probabilité de couverture asymptotique $1-\alpha$. ♢

Exemple II-1.17 (Transformation de Fisher). Soit une suite d'expériences statistiques produit d'un modèle gaussien sur $\mathbb{R}^2$

$$\left(\mathbb{R}^{2n}, \mathscr{B}(\mathbb{R}^{2n}), \left\{ \mathrm{N}_2^{\otimes n} \left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right) , \; \theta := (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) \in \Theta := \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^* \times \mathbb{R}_+^* \times]-1,1[\right\} \right) \; ;$$

ρ est le coefficient de corrélation (voir Exemple IV-5.50). Notons (X_i, Y_i) l'application canonique $\#i$. Un estimateur de ρ est donné par le *coefficient de corrélation empirique*

$$\hat{\rho}_n := \frac{n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_{n,X} S_{n,Y}} \; , \qquad S_{n,U} := \left(n^{-1} \sum_{i=1}^{n} (U_i - n^{-1} \sum_{j=1}^{n} U_j)^2 \right)^{1/2} \; .$$

♢

On peut montrer (voir Exemple IV-5.50) que

$$\sqrt{n}(\hat{\rho}_n - \rho) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \; (1-\rho^2)^2) \; .$$

Posons

$$g(\rho) := \int \frac{1}{(1-\rho)^2} \mathrm{d}\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho} = \mathrm{arctanh}(\rho) \; .$$

g est une transformation de stabilisation de la variance pour le coefficient de corrélation. Cette transformation particulière est souvent appelée *Fisher z*. Nous avons

$$\sqrt{n}(\mathrm{arctanh}(\hat{\rho}_n) - \mathrm{arctanh}(\rho)) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1) \; .$$

L'intervalle de confiance de probabilité de couverture asymptotique $1-\alpha$ pour le coefficient de corrélation ρ associé à la transformation Fisher-z est

$$\left[\tanh\left(\mathrm{arctanh}(\hat{\rho}_n) - \frac{z_{1-\alpha/2}}{\sqrt{n}} \right) , \tanh\left(\mathrm{arctanh}(\hat{\rho}_n) + \frac{z_{1-\alpha/2}}{\sqrt{n}} \right) \right] \; .$$

Exemple II-1.18. L'idée d'appliquer une transformation g à une suite d'estimateurs pour obtenir une limite ne dépendant pas du paramètre (et ainsi construire des intervalles de confiance asymptotiques) peut se généraliser au cas où la loi limite est non-gaussienne. Il n'est pas possible ici de donner une théorie générale, mais nous allons illustrer ce type de construction à travers un exemple. Soit une suite d'expériences statistiques produit

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{\mathrm{Unif}^{\otimes n}([0,\theta])\,,\ \theta \in \Theta := \mathbb{R}_+^*\right\}\right)\ .$$

Considérons la suite d'estimateurs définie pour tout $n \in \mathbb{N}$ par

$$\hat{\theta}_n := \max(X_1, X_2, \ldots, X_n)\ .$$

Nous avons montré dans l'exemple II-1.18 que, pour tout $\theta \in \Theta$, $n(\theta - \hat{\theta}_n)$ converge en loi vers une loi exponentielle de paramètre θ,

$$n(\theta - \hat{\theta}_n) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{Exp}(\theta)\ .$$

Or, pour tout $\theta \in \Theta$, la loi exponentielle de paramètre θ a pour variance $\sigma^2(\theta) = 1/\theta^2$. Considérons donc la transformation

$$g(\theta) = \int \frac{1}{\theta} \mathrm{d}\theta = \log(\theta)\ .$$

Notons que, pour tout $\theta \in \Theta$, si Y est distribuée suivant la loi $\mathrm{Exp}(\theta)$, alors Y/θ est distribuée suivant une loi exponentielle de paramètre 1. En utilisant le théorème IV-5.47 avec $r_n = n$ et $g(\theta) = \log(\theta)$, nous obtenons, pour tout $\theta \in \Theta$,

$$n(\log(\theta) - \log(\hat{\theta}_n)) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{Exp}(1)\ .$$

Soit $\alpha \in]0,1[$. Puisque $\int_{-\log(1-\alpha)}^{\infty} \exp(-u)\mathrm{d}u = 1-\alpha$, la relation précédente implique que, pour tout $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta}\left(n(\log(\theta) - \log(\hat{\theta}_n)) \leq -\log(1-\alpha)\right) = 1-\alpha\ .$$

La suite d'intervalles définis, pour tout $n \in \mathbb{N}^*$ par

$$\left[\hat{\theta}_n, \hat{\theta}_n \exp\left(-n^{-1}\log(1-\alpha)\right)\right]$$

est une suite d'intervalles de confiance de probabilité de couverture asymptotique $1-\alpha$ du paramètre θ. ♢

II-1.4 Tests asymptotiques

Nous avons construit dans la section I-3.4 des tests en utilisant des fonctions pivotales. Comme pour les intervalles de confiance, il est possible aussi d'utiliser des fonctions pivotales asymptotiques.

Soit une suite d'expériences statistiques produit

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n}\,,\ \theta \in \Theta \subseteq \mathbb{R}\right\}\right)\ .$$

Supposons que pour tout $\theta \in \Theta$,

$$\int x^2 \mathbb{P}_\theta(\mathrm{d}x) < \infty\,, \quad \text{et} \quad \theta) \int x\, \mathbb{P}_\theta(\mathrm{d}x)\ .$$

Posons

$$\sigma^2(\theta) := \int (x_1 - \theta)^2\, \mathbb{P}_\theta(\mathrm{d}x)\ .$$

Le théorème de la limite centrale (théorème IV-5.39) assure alors que, pour tout $\theta \in \Theta$,

$$\sqrt{n}\,\frac{\overline{X}_n - \theta}{\sigma(\theta)} \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1)\ .$$

Si de plus l'application $\theta \mapsto \sigma(\theta)$ est continue, le lemme de Slutsky (lemme IV-5.33) garantit que

$$G_n(X_1, \ldots, X_n, \theta) := \sqrt{n}\,\frac{\overline{X}_n - \theta}{\sigma(\overline{X}_n)} \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1)\ .$$

Considérons le test des hypothèses

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

et la statistique de test

$$\phi(X_1, \ldots, X_n) := \mathbb{1}_{\{\overline{X}_n > \sigma(\overline{X}_n) z_{1-\alpha}/\sqrt{n}\}} \ .$$

Cette statistique de test vérifie, pour tout $\theta \leq 0$,

$$\begin{aligned} \mathbb{P}_{n,\theta}\left(\phi(X_1, \ldots, X_n) = 1\right) &= \mathbb{P}_{n,\theta}\left(\overline{X}_n > \sigma(\overline{X}_n)\frac{z_{1-\alpha}}{\sqrt{n}}\right) \\ &\leq \mathbb{P}_{n,\theta}\left(G_n(X_1, \ldots, X_n, \theta) > z_{1-\alpha}\right) \xrightarrow[n\to\infty]{} \alpha \ . \end{aligned}$$

Le test est de niveau asymptotique α.

Définition II-1.19 (Taille et niveau asymptotique). *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n}, \ \theta \in \Theta\right\}\right) \ .$$

Soit $\{\phi_n, \ n \in \mathbb{N}^*\}$ *une suite de tests (possiblement randomisés) : (pour chaque* $n \in \mathbb{N}$*,* $\phi_n : \mathsf{X}^n \to [0,1]$ *est une fonction de test randomisé.*

Considérons le test d'hypothèses

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

La taille asymptotique *de la suite de tests* $\{\phi_n, \ n \in \mathbb{N}\}$ *est définie par*

$$\sup_{\theta \in \Theta_0} \limsup_{n\to\infty} \beta_{\phi_n}(\theta) \ .$$

La suite de tests $\{\phi_n, \ n \in \mathbb{N}^*\}$ *est de* niveau asymptotique α *si sa taille asymptotique est inférieure à* α*, c'est à dire si*

$$\textit{pour tout } \theta \in \Theta_0, \qquad \limsup_{n\to\infty} \beta_{\phi_n}(\theta) \leq \alpha \ .$$

La propriété de *consistance* d'une suite de tests permet de contrôler la fonction puissance : elle converge vers 1 lorsque $n \to \infty$ pour tout θ vérifiant l'hypothèse alternative.

Définition II-1.20 (Consistance). *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n}, \ \theta \in \Theta\right\}\right) \ .$$

Soit $\{\phi_n, \ n \in \mathbb{N}\}$ *une suite de test des hypothèses*

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

La suite $\{\phi_n, \ n \in \mathbb{N}\}$ *est* consistante *ou* convergente *si*

$$\textit{pour tout } \theta \in \Theta_1, \qquad \liminf_{n\to\infty} \beta_{\phi_n}(\theta) = 1 \ .$$

II-1.4.1 Exemples

L'exemple II-1.21 étudie un test sur la moyenne lorsque la variance est inconnue et que l'on dispose d'un n-échantillon d'un modèle statistique. A la différence de ce que nous avons vu dans l'Exemple I-3.19, le modèle n'est pas supposé gaussien ici.

Exemple II-1.21 (Distribution asymptotique de la statistique de Student). Soit une suite d'expériences statistiques produit

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta \in \Theta\right\}\right) .$$

On suppose que pour tout $\theta \in \Theta$, les quantités suivantes existent

$$\mu(\theta) := \mathbb{E}_\theta[X_1] , \qquad \sigma^2(\theta) := \mathbb{E}_\theta[X_1^2] .$$

Soit $\mu_0 \in \mathbb{R}$. Pour tester les hypothèses

$$H_0 :\ \mu(\theta) = \mu_0, \quad \text{contre} \quad H_1 :\ \mu(\theta) \neq \mu_0 ,$$

nous considérons la statistique de test

$$T_n := \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

où $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ est la moyenne empirique et $S_n^2 := (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ est un estimateur de la variance. Nous avons pour tout $\theta \in \Theta_0$,

$$T_n \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1) .$$

Considérons le test de fonction

$$\phi(X_1, \ldots, X_n) := \mathbb{1}_{\{|T_n| > z_{1-\alpha/2}\}} ,$$

où z_u est le quantile d'ordre u de la loi $\mathrm{N}(0,1)$. On déduit de la normalité asymptotique que pour tout $\theta \in \Theta_0$, lorsque $n \to \infty$,

$$\mathbb{P}_{n,\theta}\left(T_n \in \mathscr{R}\right) = \mathbb{P}_{n,\theta}\left(|T_n| > z_{1-\alpha/2}\right) \to \mathbb{P}(|U| > z_{1-\alpha/2})$$

où $U \sim \mathrm{N}(0,1)$. Le terme de droite vaut α par définition du seuil $z_{1-\alpha/2}$. Ainsi, le test est de taille (et de niveau) asymptotique α. Regardons sa puissance. Nous écrivons

$$T_n = \frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu(\theta)\right) + \frac{\sqrt{n}}{S_n}\left(\mu(\theta) - \mu_0\right)$$

de sorte que la région d'acceptation s'écrit

$$\begin{aligned}
&\left\{Z_n = (X_1, \cdots, X_n)\ : |T_n| \leq z_{1-\alpha/2}\right\} \\
&= \left\{Z_n = (X_1, \cdots, X_n)\ : -z_{1-\alpha/2} - \frac{\sqrt{n}}{S_n}\left(\mu(\theta) - \mu_0\right) \leq \frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu(\theta)\right) \leq z_{1-\alpha/2} - \frac{\sqrt{n}}{S_n}\left(\mu(\theta) - \mu_0\right)\right\}
\end{aligned}$$

Soit $\theta \in \Theta_1$ tel que $\mu(\theta) - \mu_0 > 0$. Nous avons pour tout $M > 0$

$$\lim_n \mathbb{P}_{n,\theta}\left(\sqrt{n}(\mu(\theta) - \mu_0)/S_n \geq M\right) = 1$$

qui résulte du fait que $S_n \overset{\mathbb{P}_{n,\theta}-\text{prob}}{\longrightarrow} \sigma^2(\theta)$ et que $\sqrt{n}(\mu(\theta) - \mu_0) \to +\infty$. Par suite

$$\lim_n \mathbb{P}_{n,\theta}\left(|T_n| \leq z_{1-\alpha/2}\right) \leq \lim_n \mathbb{P}_{n,\theta}\left(\frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu(\theta)\right) \leq z_{1-\alpha/2} - M\right) = \mathbb{P}\left(U \leq z_{1-\alpha/2} - M\right)$$

où nous avons utilisé que pour $\theta \in \Theta_1$,

$$\frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu(\theta)\right) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1) .$$

Cette relation étant vraie pour tout M, lorsque $M \to +\infty$ il vient

$$\lim_n \mathbb{P}_{n,\theta}\left(|T_n| \leq z_{1-\alpha/2}\right) = 0 .$$

De même, on montre que pour $\theta \in \Theta_1$ tel que $\mu(\theta) - \mu_0 < 0$, $\lim_n \mathbb{P}_{n,\theta}\left(|T_n| \leq z_{1-\alpha/2}\right) = 0$. Cela entraine que pour tout $\theta \in \Theta_1$

$$\lim_n \mathbb{P}_{n,\theta}\left(|T_n| > z_{1-\alpha/2}\right) = 1 .$$

Donc, $\phi(X_1, \ldots, X_n) = \mathbb{1}_{\{|T_n| > z_{1-\alpha/2}\}}$ est un test consistant de H_0 contre H_1. ♢

Exemple II-1.22 (Test dans un modèle de Bernoulli). Reprenons l'exemple de Bernoulli, voir exemple I-1.1, exemple I-3.26 et exemple I-3.27 ; il s'agit de tester $H_0 : \theta \leq 1/2$ contre $H_1 : \theta > 1/2$.

L'estimateur $\overline{X}_n := n^{-1}\sum_{i=1}^n X_i$ est asymptotiquement normal par le T.L.C. (Théorème IV-5.39) et $\overline{X}_n(1-\overline{X}_n)$ est un estimateur consistant de $\theta(1-\theta)$ par les théorèmes IV-5.6 et IV-5.18.

Par conséquent, le test de région de rejet

$$\mathscr{R}_n = \left\{ z = (x_1, \ldots, x_n) \in \{0,1\}^n : \bar{x}_n > \frac{1}{2} + z_{1-\alpha}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \right\} ,$$

est de taille asymptotique égale à α. On peut vérifier avec des arguments similaires qu' il est consistant.

On peut aussi comparer sa puissance pour n fixé à celle du test obtenu par l'inégalité d'Hoeffding. Pour cela, on remarque d'abord qu'on a toujours $\bar{x}_n(1-\bar{x}_n) \leq 1/4$, donc il suffit de comparer $\ln(1/\alpha)$ et $z_{1-\alpha}^2/2$. Or, si $N \sim \mathrm{N}(0,1)$, on a pour tout $x > 0$,

$$\mathbb{P}(N > x) \leq \frac{1}{2}\mathrm{e}^{-x^2/2} .$$

Donc $z_{1-\alpha}^2 \leq 2\ln\left(\frac{1}{2\alpha}\right) \leq 2\ln(1/\alpha)$ et le test asymptotique est donc toujours plus puissant que le test basé sur l'inégalité d'Hoeffding.

Cette situation se produit très généralement et on est donc tenté de préférer les tests asymptotiques de manière systématique. Il convient toutefois de rester prudent car le niveau de ces tests n'est garanti qu'asymptotiquement. ♢

II-1.4.2 Tests de Wald

Soit une suite d'expériences statistiques produit

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta \in \Theta\right\}\right) .$$

Soit $T : \mathsf{X} \to \mathbb{R}$ une fonction mesurable. Supposons que pour tout $\theta \in \Theta$,

$$\int \{T(x)\}^2 \mathbb{P}_\theta(\mathrm{d}x) < \infty .$$

Soit $\mu : \Theta \to \mathbb{R}$ la fonction définie par

$$\mu : \theta \mapsto \mu(\theta) := \int T(x)\mathbb{P}_\theta(\mathrm{d}x)$$

On s'intéresse aux test des hypothèses

$$H_0 :\ \mu(\theta) = f_0, \quad \text{contre} \quad H_1 :\ \mu(\theta) \neq f_0 ,$$

$$H_0 :\ \mu(\theta) \leq f_0, \quad \text{contre} \quad H_1 :\ \mu(\theta) > f_0 .$$

Le test de Wald se base sur la suite de fonctions pivotales définies par

$$G_n(X_1, \ldots, X_n, \mu(\theta)) = \sqrt{n}\frac{n^{-1}\sum_{i=1}^n T(X_i) - \mu(\theta)}{\sqrt{n^{-1}\sum_{i=1}^n t^2(X_i) - \left(n^{-1}\sum_{i=1}^n T(X_i)\right)^2}} . \tag{II-1.13}$$

Par le T.L.C. (Théorème IV-5.39), la loi faible des grands nombres (Théorème IV-5.18), le théorème de l'application continue pour la convergence en probabilité (Théorème IV-5.6) et le lemme de Slutsky (Lemme IV-5.33),

$$G_n(X_1, \ldots, X_n, \mu(\theta)) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1) . \tag{II-1.14}$$

Pour le premier test, comme $\frac{1}{n}\sum_{i=1}^n T(X_i)$ est un estimateur de $\mu(\theta)$, une idée naturelle est de chercher la région de rejet sous la forme

$$\mathscr{R}_n = \left\{ (x_1, \ldots, x_n) \in \mathsf{X}^n : \left| n^{-1}\sum_{i=1}^n T(x_i) - f_0 \right| > t_{\mathrm{crit}} \right\} ,$$

où, de manière équivalente, sous la forme

$$\mathscr{R}_n = \{(x_1,\dots,x_n) \in \mathsf{X}^n \,:\, |G_n(x_1,\dots,x_n,f_0)| > t_{\text{crit}}\} \ ,$$

pour une certaine valeur t_{crit}. On choisit alors t_{crit} de façon à assurer au test un niveau asymptotiquement inférieur à α. Pour cela, d'après (II-1.14), on choisit t_{crit} solution de

$$\mathbb{P}(|\mathrm{N}(0,1)| > t_{\text{crit}}) = 1-\alpha \ ,$$

c'est à dire $t_{\text{crit}} = z_{1-\alpha/2}$ le $(1-\alpha/2)$-quantile de la loi $\mathrm{N}(0,1)$. On a montré le résultat suivant.

Proposition II-1.23. *Le test de Wald de*

$$H_0 :\, \mu(\theta) = f_0, \quad \textit{contre} \quad H_1 :\, \mu(\theta) \neq f_0 \ ,$$

de région de rejet

$$\mathscr{R}_n = \left\{x_1,\dots,x_n \in \mathsf{X}^n \,:\, |G_n(x_1,\dots,x_n,f_0)| > z_{1-\alpha/2}\right\} \ ,$$

où le pivot G_n est défini en (II-1.13) *est de taille asymptotique α.*

Pour le second test, on choisit la région de rejet

$$\mathscr{R}_n = \{x_1,\dots,x_n \in \mathsf{X}^n \,:\, G_n(x_1,\dots,x_n,f_0) > z_{1-\alpha}\} \ , \tag{II-1.15}$$

où $z_{1-\alpha}$ est le $(1-\alpha)$-quantile de la loi $\mathrm{N}(0,1)$. Soit θ tel que $\mu(\theta) \leq f_0$, on a alors, d'après (II-1.14),

$$\begin{aligned}\mathbb{P}_{n,\theta}\left(G_n(X_1,\dots,X_n,f_0) > z_{1-\alpha}\right) &\leq \mathbb{P}_{n,\theta}\left(G_n(X_1,\dots,X_n,\mu(\theta)) > z_{1-\alpha}\right) \\ &\xrightarrow[n\to\infty]{} \mathbb{P}\left(\mathrm{N}(0,1) > z_{1-\alpha}\right) = 1-\alpha \ .\end{aligned}$$

On en déduit le résultat suivant :

Proposition II-1.24. *Le test de Wald, de*

$$H_0 :\, \mu(\theta) \leq f_0, \quad \textit{contre} \quad H_1 :\, \mu(\theta) > f_0 \ ,$$

de région de rejet $\mathscr{R}_n$ définie en (II-1.15) *est de niveau asymptotique α.*

Chapitre II-2

Théorie asymptotique des (M,Z)-estimateurs

Nous avons introduit les M-estimateurs et les Z-estimateurs resp. en section I-2.4 et section I-2.2. L'objet de ce chapitre est d'établir les propriétés asymptotiques des M et Z-estimateurs : nous donnerons des conditions suffisantes pour la consistance et la normalité asymptotique de ces estimateurs.
Nous présentons ces conditions dans le cas où la définition des M et Z-estimateurs est plus générale que celle introduite au chapitre I-2.

Il est important de comprendre ces conditions suffisantes : pour les section II-2.1.1 et section II-2.1.2 au moins dans le cas de la définition des M et Z-estimateurs donnée au chapitre I-2 (les processus aléatoires M_n et Ψ_n sont des moyennes empiriques ; les estimateurs $\hat{\theta}_n$ sont définis resp. comme le maximum et le zero de ces processus empiriques) ; pour la section II-2.2.1, au moins dans le cas scalaire.
Les démonstrations des théorèmes peuvent être omises dans une première lecture, notamment celles de la section II-2.1.3.

Dans toute la suite de ce chapitre, nous supposerons disposer d'une suite d'expériences statistiques produit

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta \in \Theta \subseteq \mathbb{R}^d\right\}\right) .$$

Nous noterons $X_i, i \geq 1$, les observations canoniques.

Nous prenons ici une définition des M-estimateurs un peu plus générale que celle donnée en section I-2.4, notion qui permet de couvrir aussi les situations où l'estimateur $\hat{\theta}_n$ est une solution approchée du problème d'optimisation. Cette définition est donnée en **H** II-2.1. En pratique, le problème d'optimisation est souvent résolu à l'aide d'une méthode numérique : la solution obtenue numériquement approche le maximum avec une certaine précision, qu'il est possible de contrôler (on est capable de donner une borne supérieure de l'erreur d'approximation). Il est beaucoup plus rare que nous disposions d'une solution exacte.

Comme pour les M-estimateurs, nous prendrons ici une définition des Z-estimateurs plus générale. Cette définition est donnée en **H** II-2.5.

II-2.1 Consistance des Z- et des M-estimateurs

Dans cette section, nous donnons des critères simples sur la famille paramétrique $\left\{\mathbb{P}_\theta, \theta \in \Theta\right\}$ et la fonction m (resp. ψ) pour les M-estimateurs (resp. les Z-estimateurs) qui garantissent la consistance de

l'estimateur correspondant. Les conditions que nous présentons sont classiques mais sous-optimales. La recherche de conditions minimales est un problème délicat qui dépasse le cadre de ce cours.

Pour des raisons techniques, nous commençons par traiter la consistance des M-estimateurs, dont nous déduirons celle des Z-estimateurs.

II-2.1.1 Consistance des M-estimateurs

Soit $m : \Theta \times \mathsf{X} \to \mathbb{R}$ une application mesurable telle que *(i)* pour tout $\theta, \theta_0 \in \Theta$, $\mathbb{E}_{\theta_0}\left[|m(\theta, X_1)|\right] < \infty$ et *(ii)* pour tout $\theta_0 \in \Theta$,

$$\theta \mapsto M_{\theta_0}(\theta) := \mathbb{E}_{\theta_0}[m(\theta, X_1)] \tag{II-2.1}$$

atteint son maximum en θ_0. On pose

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} m(\theta, X_i) \; .$$

La consistance des M-estimateurs est liée au comportement asymptotique de la suite de *processus aléatoires* $\{\theta \mapsto M_n(\theta), n \geq 1\}$. La loi des grands nombres (Théorème IV-5.18) montre, que pour tout $\theta \in \Theta$,

$$M_n(\theta) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} M_{\theta_0}(\theta) \; .$$

Il semble raisonnable de penser que la suite $\{\hat{\theta}_n, n \in \mathbb{N}\}$ converge vers la valeur du paramètre qui maximise la fonction $\theta \mapsto M_{\theta_0}(\theta)$. Or, la fonction m est choisie de telle sorte que $\theta \mapsto M_{\theta_0}(\theta)$ atteigne son maximum en θ_0. Par suite, nous pouvons donc espérer que $\hat{\theta}_n \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \theta_0$, i.e. que $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est un estimateur consistant pour θ_0.

Toutefois, sans grande surprise, la "convergence simple" en probabilité (on fixe la valeur de $\theta \in \Theta$ et on prend la limite quand $n \to \infty$) de la suite de fonctions $\theta \mapsto M_n(\theta)$ vers sa limite $\theta \mapsto M_{\theta_0}(\theta)$ est une propriété trop faible pour espérer conclure à la convergence du maximum de l'une vers le maximum de l'autre. Nous devons donc établir la "convergence uniforme" de cette suite de fonctions.

Nous devons aussi supposer que le maximum de la fonction limite $\theta \to M_{\theta_0}(\theta)$ est isolé, i.e. que seules des valeurs dans un voisinage de θ_0 donnent des valeurs proches de $M_{\theta_0}(\theta_0)$.

Nous allons maintenant donner une formulation mathématique de ces intuitions. Nous le faisons dans un cadre assez général où le processus $\theta \mapsto M_n(\theta)$ n'est pas nécessairement celui donné par $n^{-1}\sum_{i=1}^{n} m(\theta, X_i)$.

***H** II-2.1 (Hypothèses pour la consistance des M-estimateurs généraux).* Soit $M_n : \Theta \to \mathbb{R}$, $\theta \mapsto M_n(\theta)$ une suite de processus aléatoires, et pour tout $\theta_0 \in \Theta$, soit $\theta \mapsto M_{\theta_0}(\theta)$ une fonction mesurable.

(i) Pour tout $\theta_0 \in \Theta$ et $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}\left(\sup_{\theta\in\Theta} \left|M_n(\theta) - M_{\theta_0}(\theta)\right| \geq \varepsilon\right) = 0 \; .$$

(ii) Pour tout $\theta_0 \in \Theta$ et tout $\varepsilon > 0$,

$$\sup_{\|\theta-\theta_0\|\geq\varepsilon} M_{\theta_0}(\theta) < M_{\theta_0}(\theta_0) \; .$$

(iii) Il existe une suite de variables aléatoires positives $\{\rho_n,\ n \in \mathbb{N}\}$ et une suite de variables aléatoires $\{\hat{\theta}_n, n \in \mathbb{N}\} \subset \Theta$ telles que

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}\left(\rho_n \geq \varepsilon\right) = 0 \; , \quad \text{pour tout } \varepsilon > 0 \tag{II-2.2}$$

$$\liminf_{n\to\infty} \mathbb{P}_{n,\theta_0}\left(M_n(\hat{\theta}_n) \geq M_n(\theta_0) - \rho_n\right) = 1 \; . \tag{II-2.3}$$

♢

Théorème II-2.2 (Convergence des *M*-estimateurs généraux). *Supposons* ***H*** *II-2.1. Alors le M-estimateur* $\{\hat\theta_n, n\in\mathbb{N}\}$ *est consistant : pour tout* $\theta_0\in\Theta$ *et* $\varepsilon>0$*,*

$$\lim_{n\to\infty}\mathbb{P}_{n,\theta_0}\left(\left\|\hat\theta_n-\theta_0\right\|\geq\varepsilon\right)=0\,.$$

Démonstration. Soit $\theta_0\in\Theta$. Comme θ_0 est le maximum de la fonction $\theta\mapsto M_{\theta_0}(\theta)$, nous avons

$$\begin{aligned}0\leq M_{\theta_0}(\theta_0)-M_{\theta_0}(\hat\theta_n)&=M_{\theta_0}(\theta_0)-M_n(\theta_0)+M_n(\theta_0)-M_n(\hat\theta_n)+M_n(\hat\theta_n)-M_{\theta_0}(\hat\theta_n),\\&\leq 2\sup_{\theta\in\Theta}|M_n(\theta)-M_{\theta_0}(\theta)|+\rho_n+\{M_n(\theta_0)-M_n(\hat\theta_n)-\rho_n\}\mathbb{1}_{\{M_n(\theta_0)-\rho_n>M_n(\hat\theta_n)\}}\,.\end{aligned}$$

Les hypothèses (i) et (iii) entrainent que pour tout $\eta>0$, nous avons

$$\mathbb{P}_{n,\theta_0}\left(M_{\theta_0}(\theta_0)-M_{\theta_0}(\hat\theta_n)\geq\eta\right)=0\,.$$

Soit $\varepsilon>0$. D'après la condition (ii), il existe $\eta>0$ tel que $M_{\theta_0}(\theta)\leq M_{\theta_0}(\theta_0)-\eta$ pour tout $\theta\in\Theta$ tels que $\|\theta-\theta_0\|\geq\varepsilon$, ce qui implique

$$\left\{\left\|\hat\theta_n-\theta_0\right\|\geq\varepsilon\right\}\subset\left\{M_{\theta_0}(\hat\theta_n)\leq M_{\theta_0}(\theta)-\eta\right\}\,. \tag{II-2.4}$$

Par conséquent, nous avons

$$\begin{aligned}\mathbb{P}_{n,\theta_0}\left(|\hat\theta_n-\theta_0|\geq\varepsilon\right)&\leq\mathbb{P}_{n,\theta_0}\left(M_{\theta_0}(\hat\theta_n)<M_{\theta_0}(\theta_0)-\eta\right)\\&=\mathbb{P}_{n,\theta_0}\left(M_{\theta_0}(\theta_0)-M_{\theta_0}(\hat\theta_n)>\eta\right)\xrightarrow[n\to\infty]{}0\,.\end{aligned}$$

Ce qui établit la convergence en probabilité et donc la consistance. □

Le résultat suivant donne des conditions suffisantes pour vérifier l'hypothèse **H** II-2.1-(ii).

Lemme II-2.3. *Supposons que* Θ *soit un compact de* $\mathbb{R}^d$*. Supposons de plus que*

(i) pour tout $\theta_0\in\Theta$*, la fonction* $\theta\mapsto M_{\theta_0}(\theta)$ *est continue,*

(ii) pour tout $\theta\neq\theta_0\in\Theta$*,* $M_{\theta_0}(\theta)<M_{\theta_0}(\theta_0)$*.*

Alors ***H*** *II-2.1-(ii) est vérifiée.*

Démonstration. Toute fonction continue atteint son maximum sur un ensemble compact. Donc la fonction $\theta\mapsto M_{\theta_0}(\theta)$ atteint son maximum sur $\Theta\setminus\mathrm{B}(\theta_0,\varepsilon)$ qui est un sous-ensemble compact. Notons θ_ε un point où le maximum est atteint. Nous avons $M_{\theta_0}(\theta_\varepsilon)<M_{\theta_0}(\theta_0)$ ce qui montre le résultat désiré. □

Exemple II-2.4 (Estimateur des moindres carrés, modèle de translation). Soit q une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ qui satisfait

$$\sigma^2:=\int x^2\,q(x)\mathrm{d}x<\infty,\qquad\int x\,q(x)\mathrm{d}x=0\,.$$

Soit Θ un sous ensemble compact de $\mathbb{R}$ (par exemple un segment). Pour tout $\theta\in\Theta$, définissons

$$q_\theta(x):=q(x-\theta)\,.$$

Soit une suite d'expériences statistiques produits

$$\left(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n),\left\{p_{n,\theta}:=\mathbb{Q}_\theta^{\otimes n},\,\theta\in\Theta\right\}\right)\,,\qquad\mathbb{Q}_\theta:=q_\theta\cdot\mathrm{d}\lambda_{\mathrm{Leb}}\,.$$

Considérons un M-estimateur de θ basé sur la fonction

$$m(\theta,x):=-(x-\theta)^2\,.$$

• Calculons la fonction M_{θ_0} pour $\theta_0 \in \Theta$: pour tout $\theta \in \mathbb{R}$,

$$M_{\theta_0}(\theta) = -\mathbb{E}_{\theta_0}[(X_1-\theta)^2] = -\int (x-\theta)^2 q(x-\theta_0)\mathrm{d}x = -\int (x+\theta_0-\theta)^2 q(x)\mathrm{d}x$$
$$= -\sigma^2 - (\theta_0-\theta)^2 \,;$$

dans la dernière égalité, nous avons utilisé le fait que l'espérance sous la loi $q\cdot\lambda_{\mathrm{Leb}}$ est nulle.

Le maximum de la fonction $\theta \mapsto M_{\theta_0}(\theta)$ est atteint en θ_0 et la condition **H** II-2.1-(ii) est satisfaite en appliquant le Lemme II-2.3.

• Déterminons le M-estimateur associé à ce constraste. La fonction

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

admet un maximum unique, donné ici par $\hat\theta_n := \bar X_n = n^{-1}\sum_{i=1}^n X_i$.

• Prouvons sa consistance en vérifiant le jeu de conditions suffisantes **H** II-2.1. Par définition de $\hat\theta_n$, $M_n(\hat\theta_n) \geq \sup_{\theta\in\Theta} M_n(\theta)$ et la condition **H** II-2.1-(iii) est automatiquement satisfaite.

Il reste à vérifier **H** II-2.1-(i), qui est en général la condition la plus délicate. Dans ce cas précis, la situation est simple car

$$M_n(\theta) - M_{\theta_0}(\theta) = -\frac{1}{n}\sum_{i=1}^n X_i^2 + 2\theta\frac{1}{n}\sum_{i=1}^n X_i - \theta^2 + \sigma^2 + (\theta-\theta_0)^2$$
$$= -\left\{\frac{1}{n}\sum_{i=1}^n X_i^2 - \sigma^2 - {\theta_0}^2\right\} - 2\theta\left\{\frac{1}{n}\sum_{i=1}^n X_i - \theta_0\right\}$$

Par conséquent, nous avons

$$\sup_{\theta\in\Theta}|M_n(\theta)-M_{\theta_0}(\theta)| \leq \left|\frac{1}{n}\sum_{i=1}^n X_i^2 - \sigma^2 - {\theta_0}^2\right| + 2\mathrm{diam}(\Theta)\left|\frac{1}{n}\sum_{i=1}^n X_i - \theta_0\right| \,,$$

où $\mathrm{diam}(\Theta)$ est le diamètre de Θ. Par la loi des grands nombres, pour tout $\theta_0 \in \Theta$, nous avons

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \overset{\mathbb{P}_{n,\theta_0}-\mathrm{prob}}{\longrightarrow} \sigma^2 + {\theta_0}^2 \,,$$
$$\frac{1}{n}\sum_{i=1}^n X_i \overset{\mathbb{P}_{n,\theta_0}-\mathrm{prob}}{\longrightarrow} \theta_0 \,,$$

ce qui montre que, pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty}\mathbb{P}_{n,\theta_0}\left(\sup_{\theta\in\Theta}|M_n(\theta)-M_{\theta_0}(\theta)| \geq \varepsilon\right) = 0 \,.$$

Il est important de remarquer ici qu'il est essentiel, pour obtenir le résultat de convergence uniforme, de supposer que l'espace des paramètres est un intervalle compact.

• Utiliser ici le résultat général de consistance des M-estimateurs est ici assez maladroit, car l'estimateur $\hat\theta_n$ a une forme explicite et la consistance de la suite d'estimateurs $\{\hat\theta_n, n\in\mathbb{N}\}$ est élémentaire.

Le théorème II-2.2 prend tout son sens lorsque les M-estimateurs n'admettent pas d'expressions explicites, ce qui se produit dans la grande majorité des cas.

• Sur la fig. II-2.1(gauche), on considère le cas d'un modèle gaussien ($q \equiv \mathrm{N}(0,1)$). On trace la fonction $\theta \mapsto M_{\theta_0}(\theta)$ dans le cas $\theta_0 = 2$; ainsi qu'une réalisation du processus aléatoire $\theta \mapsto M_n(\theta)$. Trois processus sont considérés, correspondant à $n=2$, $n=10$ et $n=20$. Au centre et à droite, on représente 10 réalisations du processus M_n dans le cas $n=2$ puis dans le cas $n=50$.
Sur la fig. II-2.2, on trace les fonctions $\Psi_{\theta_0} = \nabla M_{\theta_0}$ et une réalisation du processus empirique $\Psi_n = \nabla_\theta M_n$. Cette approche revient à chercher le Z-estimateur associé à la fonction $\psi(\theta,x) = \nabla_\theta m(\theta,x)$. ♢

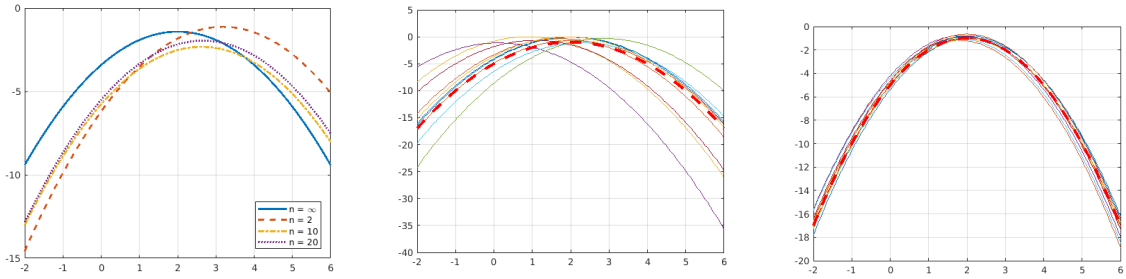

FIGURE II-2.1 – Cas d'un modèle de translation gaussien. [gauche] On trace la fonction M_{θ_0} dans le cas $\theta_0 = 2$; et une réalisation du processus aléatoire M_n. Trois processus aléatoires sont considérés, pour différentes valeurs de n. [centre, droite] On trace la fonction M_{θ_0} en trait pointillé rouge, ainsi que 10 réalisations du processus M_n dans le cas $n = 2$ au centre, et $n = 50$ à droite.

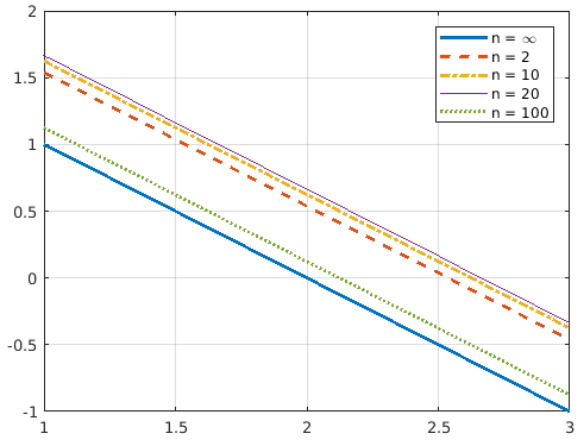

FIGURE II-2.2 – Cas d'un modèle de translation gaussien. On trace la fonction Ψ_{θ_0} dans le cas $\theta_0 = 2$ (cela correspond à $n = \infty$ dans la légende) ; et une réalisation du processus aléatoire Ψ_n. Quatre processus aléatoires sont considérés, pour différentes valeurs de n.

II-2.1.2 Consistance des Z-estimateurs

Soient $\psi_j : \Theta \times \mathsf{X} \to \mathbb{R}$, pour $j = 1, \cdots, d$, des fonctions mesurables. On pose $\boldsymbol{\psi} := (\psi_1, \cdots, \psi_d)^\top \in \mathbb{R}^d$. Supposons que *(i)* pour tout $\theta, \theta_0 \in \Theta$, $\mathbb{E}_{\theta_0}[\|\boldsymbol{\psi}(\theta, X_1)\|] < \infty$, et *(ii)* pour tout $\theta_0 \in \Theta$, $\mathbb{E}_{\theta_0}[\boldsymbol{\psi}(\theta_0, X_1)] = 0$.

On pose

$$\Psi_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\theta, X_i) \, .$$

D'après la loi des grands nombres (Théorème IV-5.18), pour tout $\theta, \theta_0 \in \Theta$,

$$\Psi_n(\theta) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \Psi_{\theta_0}(\theta) := \mathbb{E}_{\theta_0}\left[\boldsymbol{\psi}(\theta, X_1)\right] \, .$$

Sous des hypothèses appropriées, il est raisonnable d'espérer que les solutions du système $\Psi_n(\theta) = 0$ convergent vers les solutions de $\Psi_{\theta_0}(\theta) = 0$. Ici encore, la convergence ponctuelle ne suffit pas et il faudra supposer que la convergence est uniforme. Si θ_0 est une solution de $\Psi_{\theta_0}(\theta) = 0$ et que cette solution est isolée, alors les solutions approchées de $\Psi_n(\theta) = 0$ convergent vers θ_0.

Nous donnons ci-dessous un jeu de conditions suffisantes pour la consistance des Z-estimateurs ; noter qu'il n'est pas supposé que le processus stochastique $\theta \mapsto \Psi_n(\theta)$ est de la forme $\theta \mapsto n^{-1} \sum_{i=1}^{n} \psi(\theta, X_i)$.

H *II-2.5 (Hypothèses pour la consistance d'un Z-estimateur).*

(i) Pour tout $\theta_0 \in \Theta$ et $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0} \left(\sup_{\theta \in \Theta} \left\| \Psi_n(\theta) - \Psi_{\theta_0}(\theta) \right\| \geq \varepsilon \right) = 0 \, .$$

(ii) Pour tout $\theta_0 \in \Theta$, $\Psi_{\theta_0}(\theta_0) = 0$, et pour tout $\varepsilon > 0$,

$$\inf_{\|\theta - \theta_0\| \geq \varepsilon} \left\| \Psi_{\theta_0}(\theta) \right\| > 0 \, .$$

(iii) La suite d'estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$ vérifie : pour tout $\theta_0 \in \Theta$ et tout $\varepsilon > 0$,

$$\mathbb{P}_{n,\theta_0} \left(\left\| \Psi_n(\hat{\theta}_n) \right\| \geq \varepsilon \right) = 0 \, . \qquad \diamondsuit$$

Théorème II-2.6 (Convergence des Z-estimateurs). *Supposons **H** II-2.5. Alors la suite de Z-estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est consistante : pour tout $\theta_0 \in \Theta$ et $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0} \left(\left\| \hat{\theta}_n - \theta_0 \right\| \geq \varepsilon \right) = 0 \, .$$

Démonstration. Le résultat découle du théorème II-2.2 avec $M_n(\theta) \leftarrow - \|\Psi_n(\theta)\|$ et $M_{\theta_0}(\theta) \leftarrow - \left\|\Psi_{\theta_0}(\theta)\right\|$.□

II-2.1.3 Loi faible des grands nombres uniforme

Dans cette partie, nous considérons le cas où M_n ou Ψ_n sont de la forme

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

et nous donnons quelques éléments permettant de comprendre le type de résultats nécessaires pour obtenir la convergence uniforme de $\theta \mapsto M_n(\theta)$ ou $\theta \mapsto \Psi_n(\theta)$ sous $\mathbb{P}_{n,\theta_0}$ vers resp. $M_{\theta_0}(\theta)$ et $\Psi_{\theta_0}(\theta)$. Il existe bien entendu des techniques beaucoup plus sophistiquées.

Lemme II-2.7. *Considérons le modèle statistique* $(\mathsf{X}, \mathscr{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta \subset K\})$ *où K est un ensemble compact de* $\mathbb{R}^d$. *Soit* $\phi : K \times \mathsf{X} \to \mathbb{R}$, $(\theta, x) \mapsto \phi(\theta, x)$ *une fonction. On définit le module de continuité de* $\theta \mapsto \phi(\theta, x)$, *défini par : pour tout* $\delta > 0$, $x \in \mathsf{X}$,

$$w_\delta(x) := \sup_{\{\theta, \vartheta \in K\,:\, \|\theta - \vartheta\| \le \delta\}} |\phi(\theta, x) - \phi(\vartheta, x)|\,. \tag{II-2.5}$$

Supposons que

(i) pour tout $x \in \mathsf{X}$, *la fonction* $\theta \mapsto \phi(\theta, x)$ *est continue sur K,*

(ii) pour tout $\theta_0 \in \Theta$,

$$\int \sup_{\theta \in K} |\phi(\theta, x)| \mathbb{Q}_{\theta_0}(\mathrm{d}x) < \infty\,.$$

Alors, pour tout $\theta_0 \in \Theta$, *la fonction*

$$\theta \mapsto W_{\theta_0}(\theta) := \int \phi(\theta, x)\ \mathbb{Q}_{\theta_0}(\mathrm{d}x)$$

est continue sur K et

$$\lim_{\delta \to 0^+} \int_{\mathsf{X}} w_\delta(x)\ \mathbb{Q}_{\theta_0}(\mathrm{d}x) = 0\,.$$

Démonstration. Soit $\theta_0 \in \Theta$. Etablissons le premier résultat : soit $\theta \in K$ et $\{\theta_n,\ n \in \mathbb{N}\}$ une suite d'éléments de K convergeant vers θ. Nous allons vérifier les conditions du théorème de convergence dominée. On pose, pour tout $x \in \mathsf{X}$, $\varphi(x) := \sup_{\theta \in K} |\phi(\theta, x)|$ (rappelons que toute fonction continue sur un ensemble compact atteint son maximum sur ce compact). Par l'hypothèse *(i)*, pour tout $x \in \mathsf{X}$ nous avons

$$\lim_{n \to \infty} \phi(\theta_n, x) = \phi(\theta, x)\,.$$

D'autre part, pour tout $n \in \mathbb{N}$, $|\phi(\theta_n, x)| \le \varphi(x)$ et $\int \varphi(x)\ \mathbb{Q}_{\theta_0}(\mathrm{d}x) < \infty$ par l'hypothèse *(ii)*. Le théorème de convergence dominée montre que $W_{\theta_0}(\theta_n) \to W_{\theta_0}(\theta)$.

Prouvons le second résultat, là encore par application du théorème de convergence dominée. Toute fonction continue sur un compact est uniformément continue ; nous avons donc, pour tout $x \in \mathsf{X}$,

$$\lim_{\delta \to 0^+} w_\delta(x) = 0\,.$$

D'autre part, nous avons, pour tout $\delta > 0$, $w_\delta(x) \le 2\varphi(x)$. Pour tout suite $\{\delta_n,\ n \in \mathbb{N}\}$ telle que $\lim_{n \to \infty} \delta_n = 0$, le théorème de convergence dominée montre que

$$\lim_{n \to \infty} \int w_{\delta_n}(x) \mathbb{Q}_{\theta_0}(\mathrm{d}x) = 0\,.$$

□

Théorème II-2.8. *Considérons une suite de d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} = \mathbb{P}_\theta^{\otimes n}\,,\ \theta \in \Theta \subset K\right\}\right)$$

où K est un ensemble compact de $\mathbb{R}^d$. *Soit* $\phi : K \times \mathsf{X} \to \mathbb{R}$, $(\theta, x) \mapsto \phi(\theta, x)$ *une fonction. Supposons que*

(i) pour tout $x \in \mathsf{X}$, *la fonction* $\theta \mapsto \phi(\theta, x)$ *est continue sur K.*

(ii) pour tout $\theta_0 \in K$, $\int \sup_{\theta \in K} |\phi(\theta, x)| \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty$.

Posons pour tout $\theta, \theta_0 \in \Theta \subset K$ *et* $n \ge 1$,

$$W_n(\theta) := \frac{1}{n} \sum_{i=1}^n \phi(\theta, X_i) \quad \textit{et} \quad W_{\theta_0}(\theta) := \mathbb{E}_{\theta_0}[\phi(\theta, X_1)]\,.$$

Alors, pour tout $\theta_0 \in K$ *et* $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}_{n,\theta_0}\left(\sup_{\theta \in K} |W_n(\theta) - W_{\theta_0}(\theta)| \ge \varepsilon\right) = 0\,.$$

Démonstration. Soient $\varepsilon > 0$ et $\theta_0 \in K$. Le lemme II-2.7 montre que

$$\exists \delta > 0, \qquad \text{tel que} \quad \int w_\delta(x) \mathbb{P}_{\theta_0}(\mathrm{d}x) \leq \varepsilon \,,$$

où $w_\delta(x)$ est défini en (II-2.5). L'ensemble K étant compact, on peut extraire du recouvrement $\bigcup_{\theta \in K} \mathrm{B}(\theta, \delta)$ un sous-recouvrement fini : il existe $m \in \mathbb{N}$ et $\{\theta^{(\ell)}\}_{\ell=1}^m$ tels que $K \subset \bigcup_{\ell=1}^m \mathrm{B}(\theta^{(\ell)}, \delta)$. Notons que, pour tout $\theta, \vartheta \in \Theta$ tels que $\|\theta - \vartheta\| \leq \delta$, nous avons

$$|W_{\theta_0}(\theta) - W_{\theta_0}(\vartheta)| \leq \int |\phi(\theta, x) - \phi(\vartheta, x)| \mathbb{P}_{\theta_0}(\mathrm{d}x) \leq \int w_\delta(x) \mathbb{Q}_{\theta_0}(\mathrm{d}x) \,.$$

ce qui implique

$$\sup_{\{\theta, \vartheta \in K : \, \|\theta - \vartheta\| \leq \delta\}} |W_{\theta_0}(\theta) - W_{\theta_0}(\vartheta)| \leq \varepsilon \,. \tag{II-2.6}$$

Nous écrivons

$$\sup_{\theta \in K} |W_n(\theta) - W_{\theta_0}(\theta)| = \max_{\ell \in \{1, \cdots, m\}} \sup_{\theta \in \mathrm{B}(\theta^{(\ell)}, \delta)} |W_n(\theta) - W_{\theta_0}(\theta)| \,.$$

Soit $\ell \in \{1, \cdots, m\}$. Pour tout $\theta \in \mathrm{B}(\theta^{(\ell}, \delta)$, nous avons

$$|W_n(\theta) - W_{\theta_0}(\theta)| \leq |W_n(\theta) - W_n(\theta^{(\ell)}) + |W_n(\theta^{(\ell)}) - W_{\theta_0}(\theta^{(\ell)})| + |W_{\theta_0}(\theta^{(\ell)}) - W_{\theta_0}(\theta)|$$

ce qui implique

$$\begin{aligned} \sup_{\theta \in \mathrm{B}(\theta^{(\ell)}, \delta)} |W_n(\theta) - W_{\theta_0}(\theta)| \leq & \sup_{\theta \in \mathrm{B}(\theta^{(\ell)}, \delta)} \Big\{ |W_n(\theta) - W_n(\theta^{(i)})| \Big\} \\ & + |W_n(\theta^{(\ell)}) - W_{\theta_0}(\theta^{(\ell)})| + \sup_{\theta \in \mathrm{B}(\theta^{(\ell)}, \delta)} |W_{\theta_0}(\theta^{(\ell)}) - W_{\theta_0}(\theta)| \,. \end{aligned}$$

L'inégalité (II-2.6) montre que uniformément en ℓ,

$$\sup_{\theta \in \mathrm{B}(\theta^{(\ell)}, \delta)} |W_{\theta_0}(\theta^{(\ell)}) - W_{\theta_0}(\theta)| \leq \varepsilon \,.$$

De plus

$$\begin{aligned} \sup_{\theta \in \mathrm{B}(\theta^{(\ell)}, \delta)} \Big\{ |W_n(\theta) - W_n(\theta^{(\ell)})| \Big\} & \leq \frac{1}{n} \sum_{i=1}^n \sup_{\theta \in \mathrm{B}(\theta^{(\ell)}, \delta)} |\phi(\theta, X_i) - \phi(\theta^{(\ell)}, X_i)| \\ & \leq \frac{1}{n} \sum_{i=1}^n w_\delta(X_i) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \int w_\delta(x) \, \mathbb{P}_{\theta_0}(\mathrm{d}x) \leq \varepsilon \,. \end{aligned}$$

Enfin,la loi des grands nombres implique

$$\max_{1 \leq \ell \leq m} |W_n(\theta^{(\ell)}) - W_{\theta_0}(\theta^{(\ell)})| \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0 \,. \qquad \square$$

On applique maintenant ces résultats généraux pour établir la consistance des M-estimateurs dans le cas où Θ est compact.

Théorème II-2.9 (Consistance d'un M-estimateur (cas compact)). *Considérons une suite de d'expériences statistiques produit*

$$\big(\mathsf{X}^n, \mathscr{X}^n, \big\{\mathbb{P}_{n,\theta} = \mathbb{P}_\theta^{\otimes n}, \; \theta \in \Theta\big\}\big)$$

où Θ est un ensemble compact de $\mathbb{R}^d$. Supposons que

(i) Pour tout $x \in \mathsf{X}$, la fonction $\theta \mapsto m(\theta, x)$ est continue.

(ii) Pour tout $\theta_0 \in \Theta$,

$$\int \sup_{\theta \in \Theta} |m(\theta, x)| \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty \, .$$

(iii) Pour tout $\theta \neq \theta_0 \in \Theta$, $M_{\theta_0}(\theta) < M_{\theta_0}(\theta_0)$ où $M_{\theta_0}(\theta) := \int m(\theta, x) \mathbb{P}_{\theta_0}(\mathrm{d}x)$.

Supposons de plus qu'il existe une suite de variables aléatoires positives $\{\rho_n,\ n \in \mathbb{N}\}$ et une suite de variables aléatoires $\{\hat{\theta}_n, n \in \mathbb{N}\} \subset \Theta$ telles que pour tout $\theta_0 \in \Theta$

$$\lim_{n \to \infty} \mathbb{P}_{n,\theta_0} \left(\rho_n \geq \varepsilon \right) = 0 \, , \quad \textit{pour tout } \varepsilon > 0 \tag{II-2.7}$$

$$\liminf_{n \to \infty} \mathbb{P}_{n,\theta_0} \left(M_n(\hat{\theta}_n) \geq M_n(\theta_0) - \rho_n \right) = 1 \, , \tag{II-2.8}$$

où

$$M_n(\theta) := n^{-1} \sum_{i=1}^{n} m(\theta, X_i) \, .$$

Alors $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est une suite consistante d'estimateurs.

Démonstration. Nous appliquons le théorème II-2.2 à $M_n(\theta) = n^{-1} \sum_{i=1}^{n} m(\theta, X_i)$. Nous allons vérifier **H** II-2.1. Sous les conditions (i) et (ii), le théorème II-2.8 montre que, pour tout $\theta_0 \in \Theta$,

$$\sup_{\theta \in \Theta} |M_n(\theta) - M_{\theta_0}(\theta)| \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0 \, ,$$

et donc que **H** II-2.1-(i) est vérifiée. Le lemme II-2.7 montre que la fonction $\theta \mapsto M_{\theta_0}(\theta)$ est continue. Le lemme II-2.3 montre que **H** II-2.1-(ii) est vérifiée. □

L'hypothèse de compacité est souvent contraignante. Il est dans certains cas possible de relâcher cette hypothèse en utilisant le résultat suivant :

Théorème II-2.10 (Consistance d'un M-estimateur (cas non-compact)). *Considérons une suite de d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{ \mathbb{P}_{n,\theta} = \mathbb{P}_{\theta}^{\otimes n},\ \theta \in \mathbb{R}^d \right\} \right)$$

(i) Pour tout $\theta_0 \in \mathbb{R}^d$, la fonction $\theta \mapsto m(\theta, x)$ est continue.

(ii) Pour tout compact $K \subset \mathbb{R}^d$ et pour tout $\theta_0 \in \mathbb{R}^d$,

$$\int \sup_{\theta \in K} |m(\theta, x)| \, \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty \, .$$

(iii) Pour tout $\theta \neq \theta_0 \in \mathbb{R}^d$,

$$M_{\theta_0}(\theta) < M_{\theta_0}(\theta_0) \, .$$

(iv) Pour tout $\theta_0 \in \mathbb{R}^d$, il existe $a > 0$ tel que

$$\int \sup_{\|\theta\| \geq a} |m(\theta, x)| \, \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty \, .$$

(v) Pour tout $\theta_0 \in \Theta$ et tout $x \in \mathsf{X}$

$$\lim_{b \to \infty} \sup_{\|\theta\| > b} m(\theta, x) = -\infty \, .$$

Supposons de plus qu'il existe une suite de variables aléatoires positives $\{\rho_n,\ n\in\mathbb{N}\}$ *et une suite de variables aléatoires* $\{\hat\theta_n, n\in\mathbb{N}\}\subset\Theta$ *telles que*

$$\lim_{n\to\infty}\mathbb{P}_{n,\theta_0}\left(\rho_n\geq\varepsilon\right)=0\,,\quad \textit{pour tout } \varepsilon>0 \tag{II-2.9}$$

$$\liminf_{n\to\infty}\mathbb{P}_{n,\theta_0}\left(M_n(\hat\theta_n)\geq M_n(\theta_0)-\rho_n\right)=1\,, \tag{II-2.10}$$

où

$$M_n(\theta):=n^{-1}\sum_{i=1}^n m(\theta,X_i)\,.$$

Alors $\{\hat\theta_n, n\in\mathbb{N}\}$ *est une suite consistante d'estimateurs.*

Démonstration. Notons que, pour tout $b>a$, nous avons

$$0\leq\sup_{\|\theta\|>a}m(\theta,x)-\sup_{\|\theta\|>b}m(\theta,x)\,.$$

Sous (v), le théorème de convergence monotone montre que

$$\lim_{b\to\infty}\int_{\mathsf{X}}\left\{\sup_{\|\theta\|>b}m(\theta,x)\right\}\mathbb{P}_{\theta_0}(\mathrm{d}x)=-\infty\,.$$

Soient $\theta_0\in\mathbb{R}^d$ et b tel que

$$\int\left\{\sup_{\|\theta\|>b}m(\theta,x)\right\}\mathbb{P}_{\theta_0}(\mathrm{d}x)<M_{\theta_0}(\theta_0)\,. \tag{II-2.11}$$

Remarquons que, par construction $b>\|\theta_0\|$. La loi forte des grands nombres montre que

$$\begin{aligned}\sup_{\|\theta\|>b}M_n(\theta)-M_n(\theta_0)&\leq\frac{1}{n}\sum_{i=1}^n\sup_{\|\theta\|>b}m(\theta,X_i)-m(\theta_0,X_i)\\&\overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow}\int\left\{\sup_{\|\theta\|>b}m(\theta,x)\right\}\mathbb{P}_{\theta_0}(\mathrm{d}x)-M_{\theta_0}(\theta_0)<0\,,\end{aligned}$$

et donc

$$\lim_{n\to\infty}\mathbb{P}_{n,\theta_0}\left(\sup_{\|\theta\|>b}M_n(\theta)-M_n(\theta_0)+\rho_n\geq 0\right)=0\,.$$

On pose $K:=\overline{\mathrm{B}(0,b)}$. Par construction, l'ensemble K est compact. On a

$$\begin{aligned}\{\hat\theta_n\notin K\}&\subset\{M_n(\hat\theta_n)\leq M_n(\theta_0)-\rho_n\}\cup\{\hat\theta_n\notin K, M_n(\hat\theta_n)\geq M_n(\theta_0)-\rho_n\}\\&\subset\{M_n(\hat\theta_n)\leq M_n(\theta_0)-\rho_n\}\cup\{\sup_{\|\theta\|\geq b}M_n(\theta)\geq M_n(\theta_0)-\rho_n\}\,.\end{aligned}$$

Donc

$$\limsup_{n\to\infty}\mathbb{P}_{n,\theta_0}(\hat\theta_n\notin K)=0\,.$$

Nous concluons en appliquant le théorème II-2.9 à l'estimateur $\hat\theta_n\mathbb{1}_K$. □

Exemple II-2.11. Soit F une fonction de répartition vérifiant $\int|x|F(\mathrm{d}x)<\infty$, $F(x)=1-F(-x)$ pour tout $x\in\mathbb{R}$ et $F(\eta)>1/2$ pour tout $\eta>0$. Pour $\theta\in\Theta=\mathbb{R}$ et $x\in\mathsf{X}$, notons

$$F_\theta(x)=F(x-\theta)\,,$$

et notons par $\mathbb{Q}_\theta$ la loi sur $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ de fonction de répartition F_θ. Soit une suite d'expériences statistiques produits

$$\left(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n),\left\{\mathbb{P}_{n,\theta}:=\mathbb{Q}_\theta^{\otimes n},\ \theta\in\Theta\right\}\right)\ .$$

Nous considérons le M-estimateur associé à la fonction $m(\theta,x)=-|x-\theta|$, i.e., pour tout $\theta,\theta_0\in\Theta$,

$$M_n(\theta)=-\frac{1}{n}\sum_{i=1}^n|X_i-\theta|\quad\text{et}\quad M_{\theta_0}(\theta)=-\int|x-\theta|F_{\theta_0}(\mathrm{d}x)\ .$$

Notons que, pour tout $x\in\mathbb{R}$, $\theta\mapsto m(\theta,x)$ est continue et donc la condition théorème II-2.10-(i) est satisfaite. Nous avons d'autre part $|m(\theta,x)|\le|\theta|+|x|$ et, pour tout $\theta_0\in\mathbb{R}$ et tout sous-ensemble compact $K\subset\mathbb{R}$,

$$\int\left\{\sup_{\theta\in K}|m(\theta,x)|\right\}F_{\theta_0}(\mathrm{d}x)\le 2\mathrm{diam}(K)+\int|x|F(\mathrm{d}x)<\infty\ ,$$

ce qui établit la condition théorème II-2.10-(ii). Pour tout $a\ge 0$ et $x\in\mathsf{X}$ on a

$$m(a,x)\le\sup_{\|\theta\|\ge a}m(\theta,x)\le 0$$

et donc

$$\left|\sup_{\|\theta\|\ge a}m(\theta,x)\right|\le|m(a,x)|\ .$$

Comme, pour tout $\theta_0\in\mathbb{R}$, $\int|m(a,x)|F_{\theta_0}(\mathrm{d}x)<\infty$, la condition théorème II-2.10-(iv) est satisfaite. Comme $|\theta|-|x|\le|\theta-x|$, nous avons $-|\theta-x|\le|x|-|\theta|$ ce qui implique

$$\sup_{\|\theta\|\le b}m(\theta,x)\le|x|-b\ .$$

Par conséquent, pour tout $x\in\mathbb{R}$, $\lim_{b\to\infty}\sup_{\|\theta\|\le b}m(\theta,x)=-\infty$ et la condition théorème II-2.10-(v) est satisfaite.

Considérons maintenant **H** II-2.19-(ii). Ici, pour tout $\theta,\theta_0\in\Theta$,

$$\begin{aligned}M_{\theta_0}(\theta)&=-\int|x-\theta|F_\theta(\mathrm{d}x)=-\int|x+(\theta_0-\theta)|F(\mathrm{d}x)\\&=-\int_0^\infty\{|x+(\theta_0-\theta)|+|x-(\theta_0-\theta)|\}F(\mathrm{d}x)\ .\end{aligned}$$

Un calcul élémentaire montre que, pour tout $\theta,\theta_0\in\mathbb{R}$,

$$\int_0^\infty\{|x+(\theta-\theta_0)|+|x-(\theta-\theta_0)|\}F(\mathrm{d}x)=\int_{-\infty}^\infty|x|F(\mathrm{d}x)+2\int_0^{|\theta-\theta_0|}(|\theta-\theta_0|-x)F(\mathrm{d}x)\ .$$

Comme, pour tout $\eta>0$, $F(\eta)>0$, $\int_0^\eta(\eta-x)F(\mathrm{d}x)>0$, on a donc, pour tout $\theta\ne\theta_0$, $M_{\theta_0}(\theta)<M_{\theta_0}(\theta_0)$. ♢

II-2.2 Normalité asymptotique des Z- et M-estimateurs

Nous précisons les résultats de la section précédente, en cherchant une vitesse de convergence $\alpha_n\to\infty$ de sorte que l'erreur normalisée

$$\alpha_n(\hat\theta_n-\theta)$$

converge en loi vers une loi limite non-dégénérée. Nous donnons des hypothèses suffisantes sur les fonctions ψ – pour les Z-estimateurs – et m – pour les M-estimateurs – de sorte qu'on ait une convergence en loi vers une gaussienne avec la normalisation $\alpha_n=\sqrt{n}$. A l'inverse de la section précédente, nous partons d'un résultat sur les Z-estimateurs pour en déduire un résultat sur les M-estimateurs.

II-2.2.1 Cas des Z-estimateurs

Nous démontrons les résultats dans le cas $\Theta\subset\mathbb{R}$ pour simplifier (de sorte que $d=1$ et $\boldsymbol{\psi}=\psi$. Étant données, d'une part une fonction $\psi:\Theta\times\mathbb{R}\to\mathbb{R}$ définissant un Z-estimateur, et d'autre part un modèle statistique, on considère le jeu d'hypothèses suivant :

***H* II-2.12 (Hypothèse pour la normalité asymptotique des Z-estimateurs : cas scalaire).** Pour tout point θ_0 dans l'intérieur Θ^o de Θ,

(i) il existe un voisinage $\mathscr{V}(\theta_0)$ tel que pour tout $x \in \mathsf{X}$, la fonction $\theta \mapsto \psi(\theta,x)$ est continûment différentiable sur $\mathscr{V}(\theta_0)$. Posons

$$\dot{\psi}(\theta,x) := \frac{\partial \psi}{\partial \theta}(\theta,x) \, .$$

(ii) Il existe une fonction mesurable g, telle que, pour tout $x \in \mathsf{X}$,

$$\sup_{\theta \in \mathscr{V}(\theta_0)} |\dot{\psi}(\theta,x)| \leq g(x), \quad \text{et} \quad \mathbb{E}_{\theta_0}\big[g(X)\big] < +\infty.$$

De plus, $\int \dot{\psi}(\theta_0,x)\mathbb{P}_{\theta_0}(\mathrm{d}x) \neq 0$.

(iii) On a

$$\int \psi^2(\theta_0,x)\mathbb{P}_{\theta_0}(\mathrm{d}x) < +\infty \, , \qquad \int \psi(\theta_0,x)\mathbb{P}_{\theta}(\mathrm{d}x) = 0 \, . \qquad \diamondsuit$$

Remarque II-2.13. Le jeu d'hypothèses II-2.12 est local : comme le suggère **H** II-2.12-(i), on doit pouvoir contrôler le comportement de la fonction $\psi(\theta,x)$ dans un voisinage de θ_0, pour tout $\theta_0 \in \Theta^o$. Ceci exclut les paramètres de la frontière de Θ dans le cas où Θ n'est pas un ouvert. $\diamondsuit$

Remarque II-2.14. La condition II-2.12-(i) est vérifiée dans de nombreux exemples, mais n'est pas vérifiée par exemple si $\psi(\theta,x) = \mathrm{signe}(x-\theta)$. La normalité asymptotique de la médiane peut être établie, mais en utilisant d'autres méthodes (voir (IV-2.24)). $\diamondsuit$

Sous ce jeu d'hypothèses, on a le comportement asymptotique suivant pour les Z-estimateurs

Théorème II-2.15 (Loi limite des Z-estimateurs : cas scalaire). *Soit une suite d'expériences statistiques produit*

$$\big(\mathsf{X}^n, \mathscr{X}^n, \{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n}, \ \theta \in \Theta \subseteq \mathbb{R}\}\big) \, .$$

*Supposons **H** II-2.12. Soit $\{\hat{\theta}_n, n \in \mathbb{N}\}$ une suite d'estimateurs consistante telle que, pour tout $\theta_0 \in \Theta^o$,*

$$n^{1/2} \left\| \Psi_n(\hat{\theta}_n) \right\| \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0 \, . \tag{II-2.12}$$

Pour tout $\theta_0 \in \Theta^o$, nous avons

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}\big(0, v_\psi(\theta_0)\big)$$

où pour tout $\theta \in \Theta$,

$$v_\psi(\theta) := \frac{\mathbb{E}_\theta\big[\psi^2(\theta,X_1)\big]}{\Big(\mathbb{E}_\theta\big[\dot{\psi}(\theta,X_1)\big]\Big)^2} = \frac{\int \psi^2(\theta,x)\mathbb{P}_\theta(\mathrm{d}x)}{\big(\int \dot{\psi}(\theta,x)\mathbb{P}_\theta(\mathrm{d}x)\big)^2} \, .$$

Démonstration. Soit $\theta_0 \in \Theta^o$. Soit $x \in \mathsf{X}$ un point tel que $\theta \to \psi(\theta,x)$ soit continûment dérivable sur $\mathscr{V}(\theta_0)$. Pour tout t tel que $\theta_0 + t \in \mathscr{V}(\theta_0)$, nous avons

$$\psi(\theta_0+t,x) = \psi(\theta_0,x) + \dot{\psi}(\theta_0,x)t + r(t,x)t \tag{II-2.13}$$

où

$$r(t,x) := \int_0^1 \{\dot{\psi}(\theta_0+wt,x) - \dot{\psi}(\theta_0,x)\}\,\mathrm{d}w \, . \tag{II-2.14}$$

Pour $\delta > 0$, définissons

$$R_\delta(x) = \sup_{|t| \leq \delta} |r(t,x)| \, .$$

Nous avons $R_\delta(x) \le 2g(x)$ et $\lim_{\delta\to 0} R_\delta(x) = 0$. Le théorème de convergence dominée implique que $\lim_{\delta\to 0}\int R_\delta(x)\mathbb{P}_{\theta_0}(\mathrm{d}x) = 0$. En utilisant (II-2.13), nous avons, $\mathbb{P}_{\theta_0}$-presque sûrement

$$\Psi_n(\theta_0+t) = \frac{1}{n}\sum_{i=1}^n \psi(\theta_0,X_i) + t\int \dot\psi(\theta_0,x)\mathbb{P}_{\theta_0}(\mathrm{d}x) + tR_n(t)\ , \tag{II-2.15}$$

où nous avons posé $R_n(t) = R_n^{(1)} + R_n^{(2)}(t)$ avec

$$R_n^{(1)} := \frac{1}{n}\sum_{i=1}^n \left\{ \dot\psi(\theta_0,X_i) - \int \dot\psi(\theta_0,x)\mathbb{P}_{\theta_0}(\mathrm{d}x)\right\}$$
$$R_n^{(2)}(t) := \frac{1}{n}\sum_{i=1}^n r(t,X_i)\ .$$

Sous **H** II-2.12-(i), la loi faible des grands nombres (Théorème IV-5.18) montre que

$$R_n^{(1)} \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0\ . \tag{II-2.16}$$

Pour toute suite $\{\delta_n,\ n\in\mathbb{N}\}$ telle que $\lim_{n\to\infty}\delta_n = 0$, nous avons

$$\sup_{|t|\le\delta_n} |R_n^{(2)}(t)| \le \frac{1}{n}\sum_{i=1}^n R_{\delta_n}(X_i) \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{n,\theta_0}[R_{\delta_n}(X_i)] = \mathbb{E}_{\theta_0}[R_{\delta_n}(X_1)] \to_{n\to\infty} 0\ .$$

Par l'inégalité de Markov (lemme IV-1.1), nous avons donc

$$\sup_{|t|\le\delta_n} |R_n^{(2)}(t)| \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0\ . \tag{II-2.17}$$

Posons $\Delta_n = \hat\theta_n - \theta_0$. Par hypothèse, $\Delta_n = o_P(1)$, donc il existe une suite $\{\delta_n,\ n\in\mathbb{N}\}$ telle que $\lim_{n\to\infty}\mathbb{P}_{n,\theta_0}(|\Delta_n|\ge \delta_n) = 0$. Comme, pour tout $\varepsilon > 0$,

$$\mathbb{P}_{n,\theta_0}(|R_n^{(2)}(\Delta_n)|\ge\varepsilon) \le \mathbb{P}_{n,\theta_0}(\sup_{|t|\le\delta_n}|R_n^{(2)}(t)|\ge\varepsilon) + \mathbb{P}_{n,\theta_0}(|\Delta_n|\ge\delta_n)\ ,$$

on a $R_n^{(2)}(\Delta_n) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0$. En combinant ce résultat avec (II-2.17), nous obtenons

$$R_n(\Delta_n) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0\ , \tag{II-2.18}$$

et donc, en utilisant (II-2.15),

$$\sqrt{n}\Psi_n(\hat\theta_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(\theta_0,X_i) + \sqrt{n}(\hat\theta_n-\theta_0)\left\{\mathbb{E}_{\theta_0}[\dot\psi(\theta_0,X_1)] + R_n(\Delta_n)\right\}\ . \tag{II-2.19}$$

Le T.L.C. (Théorème IV-5.39) montre que, sous **H** II-2.12-(iii),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(\theta_0,X_i) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}(0,\mathbb{E}_{\theta_0}[\psi^2(\theta_0,X_1)])\ .$$

On conclut en appliquant le lemme IV-5.33. □

Remarque II-2.16. Le théorème II-2.15 montre que le " bon" ordre de grandeur de l'erreur $\hat\theta_n - \theta$ est $n^{-1/2}$. En effet, la convergence vers une loi non-dégénérée [1] avec la normalisation $\sqrt{n}$ implique que si l'on choisit une autre normalisation $\alpha_n\to\infty$, alors l'erreur normalisée

$$\alpha_n(\hat\theta_n-\theta)$$

tend vers 0 en probabilité si $\alpha_n/\sqrt{n}\to 0$ et " explose [2] " si $\alpha_n/\sqrt{n}\to\infty$. ♢

1. C'est-à-dire une loi gaussienne de matrice de variance-covariance $V(\theta)$ non singulière.
2. Dans le sens suivant : pour tout $M>0$, $\liminf_{n\to\infty}\mathbb{P}_{n,\theta}\left[|\alpha_n(\hat\theta_n-\theta)|\ge M\right] > 0$.

La preuve précédente peut être étendue sans difficulté au cas où le paramètre est vectoriel. Lorsque le paramètre est d-dimensionnel, nous utilisons d équations d'estimation, $\boldsymbol{\psi}$ est une fonction $\mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^d$. Considérons les hypothèses suivantes.

***H* II-2.17** *(Hypothèses pour la normalité asymptotique des Z-estimateurs : cas vectoriel).* Pour tout point $\theta_0 \in \Theta^o$

(i) Pour tout $x \in \mathsf{X}$, la fonction $\theta \mapsto \psi(\theta, x)$ est continûment différentiable sur un voisinage $\mathscr{V}(\theta_0)$. Posons

$$\mathrm{J}_{\psi}(\theta_0) := \left[\int \left[\frac{\partial \psi_r}{\partial \theta^{(s)}}(\theta_0, x) \right] \mathbb{P}_{\theta_0}(\mathrm{d}x) \right]_{1 \le r \le d, 1 \le s \le d} \tag{II-2.20}$$

(ii) Il existe une fonction mesurable g telle que, pour tout $x \in \mathsf{X}$,

$$\sup_{\theta \in \mathscr{V}(\theta_0)} \sum_{i=1}^{d} \left\| \frac{\partial \boldsymbol{\psi}}{\partial \theta^{(i)}}(\theta, x) \right\| \le g(x) \quad \text{et} \quad \int g(x) \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty \, .$$

De plus, la matrice $\mathrm{J}_{\psi}(\theta_0)$ est inversible.

(iii) On a

$$\int \|\boldsymbol{\psi}(\theta_0, x)\|^2 \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty \, , \qquad \int \boldsymbol{\psi}(\theta_0, x) \mathbb{P}_{\theta_0}(\mathrm{d}x) = 0 \, . \qquad \diamondsuit$$

Théorème II-2.18 (Loi limite des *Z*-estimateurs : cas vectoriel). *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{ \mathbb{P}_{n,\theta} := \mathbb{P}_{\theta}^{\otimes n}, \, \theta \in \Theta \subseteq \mathbb{R}^d \right\} \right) \, .$$

Supposons ***H*** *II-2.17. Soit* $\{\hat{\theta}_n, n \in \mathbb{N}\}$ *une suite d'estimateurs consistante telle que, pour tout* $\theta_0 \in \Theta$

$$\sqrt{n} \Psi_n(\hat{\theta}_n) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0 \, . \tag{II-2.21}$$

Alors, pour tout $\theta_0 \in \Theta^o$*, nous avons*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}\left(0, [\mathrm{J}_{\boldsymbol{\psi}}(\theta_0)]^{-1} \, G_{\boldsymbol{\psi}}(\theta_0) \, [\mathrm{J}_{\boldsymbol{\psi}}(\theta_0)]^{-\top} \right) \, , \tag{II-2.22}$$

où $\mathrm{J}_{\boldsymbol{\psi}}(\theta_0)$ est définie en (II-2.20) et

$$G_{\boldsymbol{\psi}}(\theta_0) := \int \boldsymbol{\psi}(\theta_0, x) \, (\boldsymbol{\psi}(\theta_0, x))^{\top} \mathbb{P}_{\theta_0}(\mathrm{d}x) \, . \tag{II-2.23}$$

Démonstration. La preuve est similaire au cas scalaire, et est omise. □

II-2.2.2 Cas des *M*-estimateurs

L'adaptation au cas des M-estimateurs est élémentaire (du moins, sous les hypothèses très restrictives que nous utilisons). Rappelons que, si la la fonction $\theta \mapsto m(\theta, x)$ est différentiable sur Θ^o pour tout $x \in \mathsf{X}$ et si la fonction $\theta \mapsto M_n(\theta) = n^{-1} \sum_{i=1}^n m(\theta, X_i)$ atteint son maximum en un point $\hat{\theta}_n \in \Theta^o$, alors le gradient de la fonction $\theta \mapsto M_n(\theta)$ s'annule en ce point, $\nabla M_n(\hat{\theta}_n) = \mathbf{0}_{d \times 1}$. Par conséquent, $\hat{\theta}_n$ est aussi un Z-estimateur, solution (approchée) des équations d'estimation

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

où nous avons posé

$$\nabla m(\theta,x) := \left[\frac{\partial m}{\partial \theta^{(1)}}(\theta,x),\ldots,\frac{\partial m}{\partial \theta^{(d)}}(\theta,x)\right]^\top \in \mathbb{R}^d\,. \tag{II-2.24}$$

Par conséquent, en posant $\psi(\theta,x) := \nabla m(\theta,x)$, nous pouvons déduire la loi limite des M-estimateurs de la loi limite des Z-estimateurs.

***H* II-2.19 (Hypothèse pour la normalité asymptotique des M-estimateurs).** Pour tout $\theta_0 \in \Theta^o$,

(i) Pour tout $x \in \mathsf{X}$, la fonction $\theta \mapsto m(\theta,x)$ est deux fois continûment différentiable sur un voisinage $\mathscr{V}(\theta_0)$. Posons

$$\mathrm{H}_m(\theta,x) := \left[\frac{\partial^2 m}{\partial \theta^{(i)}\partial \theta^{(j)}}(\theta,x)\right]_{1\le i\le d, 1\le j\le d}\,.$$

(ii) Il existe une fonction mesurable g telle que, pour tout $x \in \mathsf{X}$,

$$\sup_{\theta\in\mathscr{V}(\theta_0)} \|\mathrm{H}_m(\theta,x)\| \le g(x) \quad \text{et} \quad \int g(x)\mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty\,,$$

De plus, la matrice $\int \mathrm{H}_m(\theta_0,x)\mathbb{P}_{\theta_0}(\mathrm{d}x)$ est inversible.

(iii) On a

$$\int \|\nabla m(\theta_0,x)\|^2\,\mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty\,, \qquad \int \nabla m(\theta_0,x)\mathbb{P}_{\theta_0}(\mathrm{d}x) = 0\,. \qquad \diamondsuit$$

Théorème II-2.20 (Loi limite des M-estimateurs : cas vectoriel). *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta\in\Theta\subseteq\mathbb{R}^d\right\}\right)\,.$$

*Supposons **H** II-2.19. Soit $\{\hat\theta_n, n\in\mathbb{N}\}$ une suite consistante d'estimateurs telle que, pour tout $\theta_0\in\Theta^o$,*

$$\sqrt{n}\,\nabla M_n(\hat\theta_n) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0\,. \tag{II-2.25}$$

Alors, pour tout $\theta_0\in\Theta^o$, nous avons

$$\sqrt{n}(\hat\theta_n-\theta_0) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}\left(0, U_m(\theta_0)\,G_m(\theta_0)\,U_m^\top(\theta_0)\right) \tag{II-2.26}$$

où

$$G_m(\theta_0) := \int \left[\nabla m(\theta_0,x)\,(\nabla m(\theta_0,x))^\top\right]\mathbb{P}_{\theta_0}(\mathrm{d}x)\,, \tag{II-2.27}$$

$$U_m(\theta_0) := \left(\int \mathrm{H}_m(\theta_0,x)\mathbb{P}_{\theta_0}(\mathrm{d}x)\right)^{-1}\,. \tag{II-2.28}$$

Démonstration. Comme indiqué plus haut, on applique le théorème II-2.18 à $\psi(\theta,x) := \nabla m(\theta,x)$. □

Chapitre II-3

Maximum de vraisemblance, information statistique et optimalité

Ce dernier chapitre est essentiellement consacré à l'étude des propriétés asymptotiques de l'estimateur du Maximum de Vraisemblance (MV). On introduit aussi *l'efficacité asymptotique*, méthode de comparaison d'estimateurs asymptotiquement normaux et basée sur la comparaison de leur variance asymptotique. Un résultat fondamental est que l'estimateur MV est optimal parmi la classes des Z-estimateurs réguliers. Nous terminons ce chapitre en montrant qu'il est possible de modifier un estimateur asymptotiquement normal pour qu'il atteigne la même optimalité que l'estimateur MV.
Les résultats importants sont la consistance de l'estimateur du Maximum de Vraisemblance (MV), sa normalité asymptotique ainsi que l'expression de sa matrice de variance-covariance asymptotique, et son efficacité asymptotique dans la classe des Z-estimateurs pour des modèles statistiques réguliers.

Dans toute la suite de ce chapitre, nous supposerons disposer d'une suite d'expériences statistiques produit

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta \in \Theta \subseteq \mathbb{R}^d\right\}\right) .$$

Nous noterons $X_i, i \geq 1$, les observations canoniques.

II-3.1 Consistance de l'estimateur du Maximum de Vraisemblance

La méthode d'estimation au sens du Maximum de Vraisemblance (MV) est un cas particulier de M-estimation (Equation (I-2.20)). Soit $\mu \in \mathbb{M}_+(X)$ une mesure σ-finie sur l'espace mesurable $(\mathsf{X}, \mathscr{X})$.

Considérons disposer d'une suite d'expériences statistiques produit

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta \in \Theta \subseteq \mathbb{R}^d\right\}\right) , \qquad \mathbb{P}_{n,\theta} := p_\theta \cdot \mathrm{d}\mu .$$

L'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$, s'il existe, peut s'interpréter comme le M-estimateur associé à la fonction

$$\ell(\theta, x) := \log p_\theta(x) .$$

Le théorème I-2.23 montre que, pour tout $\theta_0 \in \Theta$, la valeur $\theta = \theta_0$ maximise la fonction

$$\theta \mapsto M_{\theta_0}(\theta) := \int \log p_\theta(x) \mathbb{P}_{\theta_0}(\mathrm{d}x) = -\,\mathrm{KL}(\mathbb{P}_{\theta_0}, \mathbb{P}_\theta) + \int \log p(\theta_0, x) \mathbb{P}_{\theta_0}(\mathrm{d}x) . \tag{II-3.1}$$

Si, pour μ-presque tout $x \in \mathsf{X}$, la fonction $\theta \to \log p_\theta(x)$ est différentiable, alors le maximum de vraisemblance peut être aussi vu comme un Z-estimateur associé à la fonction

$$\psi(\theta,x) = \nabla \log p_\theta(x) = \begin{bmatrix} \frac{\partial \log p}{\partial \theta^{(1)}}(\theta,x) \\ \cdots \\ \frac{\partial \log p}{\partial \theta^{(d)}}(\theta,x) \end{bmatrix} \quad \theta \in \Theta, \quad x \in \mathbb{R} .$$

Par suite, nous pouvons appliquer les résultats du chapitre II-2 et en déduire une étude asymptotique de l'estimateur MV. Le théorème II-2.9 permet de donner des conditions sous lesquelles l'estimateur du maximum de vraisemblance est consistant.

Théorème II-3.1 (Consistance de l'estimateur du M.V. (cas compact)). *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta \in \Theta\right\}\right) , \qquad \mathbb{P}_{n,\theta} := p_\theta \cdot \mathrm{d}\mu .$$

Supposons que Θ est un compact de $\mathbb{R}^d$ et que,

(i) pour tout $x \in \mathsf{X}$, la fonction $\theta \mapsto \log p_\theta(x)$ est continue.

(ii) pour tout $\theta_0 \in \Theta$,

$$\int \sup_{\theta \in \Theta} |\log p_\theta(x)| \, \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty .$$

(iii) pour tout $\theta \neq \theta_0 \in \Theta$, $\mathbb{P}_\theta \neq \mathbb{P}_{\theta_0}$.

Alors, pour tout $\theta_0 \in \Theta$, la fonction

$$\theta \mapsto \int \log p_\theta(x) \mathbb{P}_{\theta_0}(\mathrm{d}x)$$

a un maximum unique au point $\theta = \theta_0$. De plus, si pour tout $\theta_0 \in \Theta$, l'estimateur $\{\hat\theta_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ vérifie

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}(\ell_n(\hat\theta_n^{\mathrm{MV}}) \geq \ell_n(\theta_0)) = 1 , \quad \textit{où} \quad \ell_n(\theta) := n^{-1} \sum_{i=1}^n \log p_\theta(X_i) ,$$

alors $\{\hat\theta_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ est consistant.

Démonstration. Il suffit d'appliquer le théorème II-2.9 avec

$$m(\theta,x) := \log p_\theta(x) .$$

Il faut donc vérifier les conditions théorème II-2.9-(i)-(ii) découlent des conditions (i) et (ii). Par le Théorème I-2.23 et (II-3.1), sous la condition (iii), $M_{\theta_0}(\theta_0) > M_{\theta_0}(\theta)$. □

On peut étendre la consistance de l'estimateur du maximum de vraisemblance au cas non compact en utilisant le théorème II-2.10

Théorème II-3.2 (Consistance de l'estimateur du M.V. (cas non compact)). *Soit une suite d'expériences statistiques produit*

$$\left(\mathsf{X}^n, \mathscr{X}^n, \left\{\mathbb{P}_{n,\theta} := \mathbb{P}_\theta^{\otimes n},\ \theta \in \Theta \subseteq \mathbb{R}^d\right\}\right) , \qquad \mathbb{P}_{n,\theta} := p_\theta \cdot \mathrm{d}\mu .$$

Supposons que

(i) pour μ-presque tout x, la fonction $\theta \mapsto \log p_\theta(x)$ est continue sur $\mathbb{R}^d$.

(ii) pour tout $\theta_0 \in \mathbb{R}^d$ et tout compact K de $\mathbb{R}^d$,

$$\int \sup_{\theta\in K} |\log p_\theta(x)|\, \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty\,.$$

(iii) pour tout $\theta \neq \theta_0 \in \mathbb{R}^d$, $\mathbb{P}_\theta \neq \mathbb{P}_{\theta_0}$.

(iv) pour tout $\theta_0 \in \mathbb{R}^d$, il existe $a > 0$ tel que

$$\int \left| \sup_{\|\theta\|\geq a} \log p_\theta(x) \right| \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty\,.$$

(v) Pour μ-presque tout x,

$$\lim_{b\to\infty} \sup_{\|\theta\|\geq b} \log p_\theta(x) = -\infty\,.$$

Si pour tout $\theta_0 \in \mathbb{R}^d$, l'estimateur $\{\hat\theta_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ vérifie

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}\left(\ell_n(\hat\theta_n^{\mathrm{MV}}) \geq \ell_n(\theta_0)\right) = 1\,, \quad \text{où} \quad \ell_n(\theta) := n^{-1}\sum_{i=1}^n \log p_\theta(X_i)\,,$$

alors $\{\hat\theta_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ est consistante.

Démonstration. C'est une application immédiate du théorème II-2.10. □

II-3.2 Loi limite de l'estimateur du Maximum de Vraisemblance

aux modèles $(\mathsf{X}, \mathscr{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$ *réguliers*, au sens de la définition I-4.16.

Rappelons que, pour un modèle régulier $(\mathsf{X}, \mathscr{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$, le gradient de la log-vraisemblance, $\nabla\ell(\theta, x)$ est bien défini. Il est appelé *fonction score* ou simplement *score de Fisher*

Définition II-3.3 (Fonction score (ou score de Fisher)). *Pour tout $\theta \in \Theta$ et $x \in \mathsf{X}$, on pose*

$$\ell(\theta, x) := \log p_\theta(x)\,.$$

Le gradient de la fonction $\theta \mapsto \ell(\theta, x)$, lorsqu'il existe,

$$\nabla\ell(\theta, x) := \begin{bmatrix} \frac{\partial \log p}{\partial \theta^{(1)}} \\ \cdots \\ \frac{\partial \log p}{\partial \theta^{(d)}} \end{bmatrix}\,,$$

s'appelle fonction score *ou simplement* score de Fisher.

Le résultat fondamental de cette section est que pour des modèles réguliers, l'estimateur MV est asymptotiquement normal et sa matrice de variance-covariance asymptotique est l'information de Fisher $\mathbb{I}(\theta)$ (définition I-4.18).

Théorème II-3.4 (Normalité asymptotique de l'estimateur du M.V.). *Soit une suite d'expériences statistiques produit, du modèle* $(\mathsf{X}, \mathscr{X}, \{p_\theta \cdot \mu\,,\ \theta, \theta \in \Theta \subseteq \mathbb{R}^d\})$ *supposé régulier (voir définition I-4.16).*

Soit $\{\hat\theta_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ *un estimateur consistant vérifiant, pour tout* $\theta_0 \in \Theta$*,*

$$\sqrt{n}\,\nabla \ell_n(\hat\theta_n^{\mathrm{MV}}) \overset{\mathbb{P}_{n,\theta_0}-\mathrm{prob}}{\longrightarrow} 0\,, \quad \textit{où} \ \ \ell_n(\theta) := n^{-1}\sum_{i=1}^n \ell(\theta, X_i)\,.$$

Alors, pour tout $\theta_0 \in \Theta$*,*

$$\sqrt{n}(\hat\theta_n^{\mathrm{MV}} - \theta_0) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}(0, \mathbb{I}^{-1}(\theta_0))\,.$$

Démonstration. Il s'agit d'une application directe du théorème II-2.20, en notant que, par la Proposition I-4.19 :

$$\mathbb{I}(\theta) = \int \nabla\left(\ell(\theta,x)\right)\,\left(\nabla\ell(\theta,x)\right)^\top\,\mathbb{P}_\theta(\mathrm{d}x) = -\int \mathrm{H}_\ell(\theta,x)\,\mathbb{P}_\theta(\mathrm{d}x)\,, \tag{II-3.2}$$

cette matrice $d \times d$ étant en outre inversible. □

Exemple II-3.5. Considérons le cas où

$$p_\theta(x) := \begin{cases} (\theta_1+\theta_2)^{-1}\mathrm{e}^{-x/\theta_1} & x > 0 \\ (\theta_1+\theta_2)^{-1}\mathrm{e}^{x/\theta_2} & x \le 0, \end{cases}$$

où $\theta = (\theta_1, \theta_2) \in \Theta := \mathbb{R}_+^* \times \mathbb{R}_+^*$. On note

$$T_{1,n} := n^{-1}\sum_{i=1}^n X_i \mathbb{1}_{]0,\infty[}(X_i) \quad \text{et} \quad T_{2,n} := -n^{-1}\sum_{i=1}^n X_i \mathbb{1}_{]-\infty,0]}(X_i)\,.$$

La log-vraisemblance normalisée des observations est donnée par

$$\theta \mapsto \ell_n(\theta) := -\log(\theta_1+\theta_2) - \frac{T_{1,n}}{\theta_1} - \frac{T_{2,n}}{\theta_2}\,.$$

Le gradient de la log-vraisemblance et sa Hessienne sont données par

$$\nabla\ell_n(\theta) = \begin{bmatrix} -\frac{1}{\theta_1+\theta_2} + \frac{T_{1,n}}{\theta_1^2} \\ -\frac{1}{\theta_1+\theta_2} + \frac{T_{2,n}}{\theta_2^2} \end{bmatrix} \qquad \mathrm{H}_{\ell_n}(\theta) = \begin{bmatrix} \frac{1}{(\theta_1+\theta_2)^2} - \frac{2T_{1,n}}{\theta_1^3} & \frac{1}{(\theta_1+\theta_2)^2} \\ \frac{1}{(\theta_1+\theta_2)^2} & \frac{1}{(\theta_1+\theta_2)^2} - \frac{2T_{2,n}}{\theta_2^3} \end{bmatrix}\,.$$

Nous vérifions aisément que les équations de vraisemblance admettent une unique solution, qui correspond à l'estimateur du maximum de vraisemblance, donné par

$$\hat\theta_n^{\mathrm{MV}} = \begin{bmatrix} \sqrt{T_{1,n}T_{2,n}} + T_{1,n} \\ \sqrt{T_{1,n}T_{2,n}} + T_{2,n} \end{bmatrix}\,.$$

On vérifie ici directement que l'estimateur du maximum de vraisemblance est consistant : en effet, par application de la loi faible des grands nombres, pour tout $\theta \in \Theta$,

$$T_{1,n} \overset{\mathbb{P}_{n,\theta}-\mathrm{prob}}{\longrightarrow} \frac{\theta_1^2}{\theta_1+\theta_2} \qquad T_{2,n} \overset{\mathbb{P}_{n,\theta}-\mathrm{prob}}{\longrightarrow} \frac{\theta_2^2}{\theta_1+\theta_2}$$

et donc $\hat\theta_n^{\mathrm{MV}} \overset{\mathbb{P}_{n,\theta}-\mathrm{prob}}{\longrightarrow} (\theta_1,\theta_2)$. Une application directe du théorème II-3.4 montre que

$$\sqrt{n}(\hat\theta_n^{\mathrm{MV}} - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \mathbb{I}^{-1}(\theta))$$

où la matrice d'information de Fisher est donnée par

$$\mathbb{I}(\theta) := \begin{pmatrix} 1 + \frac{2\theta_2}{\theta_1} & -1 \\ -1 & 1 + \frac{2\theta_1}{\theta_2} \end{pmatrix}\,.$$

♢

II-3.2.1 Trois pivots asymptotiques

Supposons que les hypothèses du théorème II-3.4 soient satisfaites. Pour tout $\theta \in \Theta^o$, nous avons donc, en appliquant le théorème II-3.4,

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{MV}} - \theta) \stackrel{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \mathbb{I}^{-1}(\theta)) \,,$$

où $\hat{\theta}_n^{\mathrm{MV}}$ est l'estimateur du maximum de vraisemblance de θ basé sur les n observations $X_1, \ldots, X_n$. Par conséquent, nous avons

$$\sqrt{n}\mathbb{I}^{1/2}(\theta)(\hat{\theta}_n^{\mathrm{MV}} - \theta) \stackrel{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,\, \mathrm{I}_d) \,, \tag{II-3.3}$$

où $\mathbb{I}^{1/2}(\theta)$ est une racine carrée de la matrice $\mathbb{I}(\theta)$ (c'est à dire une matrice vérifiant $\mathbb{I}^{1/2}(\theta)\mathbb{I}^{1/2}(\theta)^T = \mathbb{I}(\theta)$). Par conséquent, la fonction

$$G_n(X_1, \ldots, X_n, \theta) := \sqrt{n}\mathbb{I}^{-1/2}(\theta)(\hat{\theta}_n^{\mathrm{MV}} - \theta) \,.$$

est un pivot asymptotique (voir la définition II-1.14). Nous avons aussi, pour tout $\theta \in \Theta$,

$$n(\hat{\theta}_n^{\mathrm{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\mathrm{MV}} - \theta) \stackrel{\mathbb{P}_{n,\theta}}{\Longrightarrow} \chi^2(d) \tag{II-3.4}$$

où $\chi^2(d)$ est la loi du χ^2 centré à d degrés de liberté. Si $\chi^2_{1-\alpha}(d)$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(d)$, alors pour tout $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta}(n(\hat{\theta}_n^{\mathrm{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\mathrm{MV}} - \theta) \leq \chi^2_{1-\alpha}(d)) = 1 - \alpha \,.$$

Considérons l'ensemble (aléatoire)

$$S_{n,\alpha} := \left\{ \theta \in \Theta : \, n(\hat{\theta}_n^{\mathrm{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\mathrm{MV}} - \theta) \leq \chi^2_{1-\alpha}(d) \right\} \,. \tag{II-3.5}$$

Nous avons, pour tout $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta}(\theta \in S_{n,\alpha}) = 1 - \alpha.$$

La suite d'ensembles aléatoires $\{S_{n,\alpha}, n \in \mathbb{N}\}$ définit une suite de régions de confiance de θ de niveau de couverture asymptotique $1-\alpha$. Bien qu'elle soit parfaitement légitime, cette façon de construire des régions (asymptotiques) de confiance n'est pas très fréquemment utilisée. L'information de Fisher est souvent difficile à calculer de façon analytique et, quand l'information de Fisher a une expression explicite, celle-ci est le plus souvent une fonction compliquée du paramètre rendant la détermination de $S_{n,\alpha}$ délicate.

Comme nous allons le voir ci-dessous, on peut simplifier considérablement la détermination des régions de confiance en remplaçant l'information de Fisher au point θ par une approximation. Nous allons considérer deux types d'approximations. Supposons tout d'abord que nous disposions d'une expression explicite de la matrice d'information de Fisher et que la fonction $\theta \mapsto \mathbb{I}(\theta)$ est continue. Comme la suite $\{\hat{\theta}_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ est consistante, alors $\mathbb{I}(\hat{\theta}_n^{\mathrm{MV}}) \stackrel{\mathbb{P}_{n,\theta_0}-\mathrm{prob}}{\longrightarrow} \mathbb{I}(\theta)$. Par conséquent, en appliquant le théorème II-3.4 et le lemme IV-5.33, nous obtenons, pour tout $\theta \in \Theta$,

$$\sqrt{n}\mathbb{I}^{1/2}(\hat{\theta}_n^{\mathrm{MV}})(\hat{\theta}_n^{\mathrm{MV}} - \theta) \stackrel{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \mathrm{I}_d) \,,$$

ce qui implique $\lim_{n\to\infty} \mathbb{P}_{n,\theta}(\theta \in \tilde{S}_{n,\alpha}) = 1 - \alpha$ où $\tilde{S}_\alpha$ est l'ellipsoïde

$$\tilde{S}_{n,\alpha} = \left\{ \theta \in \Theta : \, n(\hat{\theta}_n^{\mathrm{MV}} - \theta)^T \mathbb{I}(\hat{\theta}_n^{\mathrm{MV}})(\hat{\theta}_n^{\mathrm{MV}} - \theta) < \chi^2_{1-\alpha}(d) \right\} \,. \tag{II-3.6}$$

Le calcul des régions de confiance (II-3.5) et (II-3.6) requiert le calcul de la matrice d'information de Fisher. Nous ne disposons pas toujours d'expressions explicites pour cette quantité, mais la proposition suivante montre que l'on peut en calculer une approximation. Pour tout $\theta \in \Theta^o$ posons

$$\mathbb{I}_n(\theta) := -n^{-1} \sum_{i=1}^{n} \mathrm{H}_\ell(\theta, X_i) \,. \tag{II-3.7}$$

Proposition II-3.6. *Supposons définition I-4.16.*

(i) Pour toute suite $\{\delta_n,\ n\in\mathbb{N}\}$ *positive telle que* $\lim_{n\to\infty}\delta_n = 0$ *et tout* $\theta_0\in\Theta^o$,

$$\lim_{n\to\infty}\mathbb{E}_{\theta_0}\left[\sup_{\|t\|\le\delta_n}\|\mathbb{I}_n(\theta_0+t)-\mathbb{I}_n(\theta_0)\|\right] = 0\,. \tag{II-3.8}$$

(ii) Si $\{\hat\theta_n, n\in\mathbb{N}\}$ *est une suite consistante d'estimateurs, alors, pour tout* $\theta_0\in\Theta^o$,

$$\mathbb{I}_n(\hat\theta_n) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \mathbb{I}(\theta_0)\,.$$

Démonstration. (i) Pour tout $\delta>0$ tel que $\mathrm{B}(\theta_0,\delta)\subset\mathscr{V}(\theta_0)$,

$$\sup_{t\in\mathrm{B}(\theta_0,\delta)}\|\mathrm{H}_\ell(\theta_0+t,x)-\mathrm{H}_\ell(\theta_0,x)\|\le 2g(x)\,.$$

D'autre part, comme $\theta\mapsto\ell(\theta,x)$ est deux fois continûment différentiable au point θ_0 pour μ-presque tout $x\in\mathsf{X}$, nous avons $\lim_{\delta\to0}R_\delta(x)=0$ où

$$R_\delta(x) := \sup_{|t|\le\delta}\|\mathrm{H}_\ell(\theta_0+t,x)-\mathrm{H}_\ell(\theta_0,x)\|\,.$$

Le théorème de convergence dominée montre que $\lim_{\delta\to0}\mathbb{E}_{\theta_0}[R_\delta(X)]=0$. Par conséquent, pour toute suite $\{\delta_n,\ n\in\mathbb{N}\}$ telle que $\lim_{n\to\infty}\delta_n=0$, on a

$$\mathbb{E}_{\theta_0}\left[\sup_{\|t\|\le\delta_n}\|\mathbb{I}_n(\theta_0+t)-\mathbb{I}_n(\theta_0)\|\right]\le\mathbb{E}_{\theta_0}\left[n^{-1}\sum_{i=1}^n R_{\delta_n}(X_i)\right]=\mathbb{E}_{\theta_0}[R_{\delta_n}(X_1)]\to_{n\to\infty}0\,.$$

(ii) On écrit $\mathbb{I}_n(\hat\theta_n)=\mathbb{I}(\theta_0)+\mathbb{I}_n(\theta_0)-\mathbb{I}(\theta_0)+\mathbb{I}_n(\hat\theta_n)-\mathbb{I}_n(\theta_0)$. La loi des grands nombres montre que

$$\mathbb{I}_n(\theta_0)-\mathbb{I}(\theta_0)\overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow}0\,.$$

Comme $\{\hat\theta_n,n\in\mathbb{N}\}$ est consistante, il existe une suite $\{\delta_n,\ n\in\mathbb{N}\}$ telle que $\lim_{n\to\infty}\delta_n=0$ et $\lim_{n\to\infty}\mathbb{P}_{n,\theta_0}(\|\hat\theta_n-\theta_0\|\ge\delta_n)=0$. Par conséquent, nous avons pour tout $\varepsilon>0$, en utilisant (i),

$$\begin{aligned}&\mathbb{P}_{n,\theta_0}\left(\|\mathbb{I}_n(\hat\theta_n)-\mathbb{I}_n(\theta_0)\|\ge\varepsilon\right)\\&\qquad\le\mathbb{P}_{n,\theta_0}\left(\|\hat\theta_n-\theta_0\|\ge\delta_n\right)+\mathbb{P}_{n,\theta_0}\left(\sup_{\|t\|\le\delta_n}\|\mathbb{I}_n(\theta_0+t)-\mathbb{I}_n(\theta_0)\|\ge\varepsilon\right)\to_{n\to\infty}0\,.\end{aligned}$$

On en déduit que $\mathbb{I}_n(\hat\theta_n)-\mathbb{I}(\theta_0)\overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow}0$. □

On appelle *Information de Fisher observée* la quantité

$$\mathbb{I}_n(\hat\theta_n^{\mathrm{MV}}) := -\frac{1}{n}\sum_{i=1}^n\mathrm{H}_\ell(\hat\theta_n^{\mathrm{MV}},X_i)\,. \tag{II-3.9}$$

En combinant (II-3.3) et la proposition II-3.6, nous obtenons donc que, pour tout $\theta_0\in\Theta^o$,

$$\sqrt{n}\,\mathbb{I}_n^{1/2}(\hat\theta_n^{\mathrm{MV}})(\hat\theta_n^{\mathrm{MV}}-\theta_0)\overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow}\mathrm{N}(0,\mathrm{I}_d)\,. \tag{II-3.10}$$

En raisonnant comme précédemment, nous avons donc pour tout $\theta\in\Theta$, $\lim_{n\to\infty}\mathbb{P}_{n,\theta}(\theta\in\check S_{n,\alpha})=1-\alpha$, où $\check S_{n,\alpha}$ est l'ellipsoïde défini par

$$\check S_{n,\alpha} := \left\{\theta\in\Theta:\ n(\hat\theta_n^{\mathrm{MV}}-\theta)^T\mathbb{I}_n(\hat\theta_n^{\mathrm{MV}})(\hat\theta_n^{\mathrm{MV}}-\theta)\le\chi^2_{1-\alpha}(d)\right\}\,. \tag{II-3.11}$$

La suite $\{\check{S}_{n,\alpha}, n \in \mathbb{N}\}$ définit une suite de régions de confiance de couverture asymptotique $1-\alpha$. Cette région de confiance dépend uniquement de la log-vraisemblance et de sa Hessienne au point $\hat{\theta}_n^{\rm MV}$. On peut construire des intervalles de confiance qui dépendent de façon plus subtile du comportement de la log-vraisemblance en son maximum. En développant la log-vraisemblance en son maximum $\hat{\theta}_n^{\rm MV}$, nous obtenons

$$\ell_n(\theta) = \ell_n(\hat{\theta}_n^{\rm MV}) - \frac{1}{2}(\hat{\theta}_n^{\rm MV} - \theta)^T \mathbb{I}_n(\theta_n^*)(\hat{\theta}_n^{\rm MV} - \theta) \,, \tag{II-3.12}$$

où θ_n^* est un élément du segment $[\theta, \hat{\theta}_n^{\rm MV}]$ (à condition que $\nabla \ell_n \, \hat{\theta}_n^{\rm MV} = 0$ ce qui est vérifié si $\hat{\theta}_n^{\rm MV}$ est le maximum de $\theta \to \ell_n(\theta)$ et $\hat{\theta}_n^{\rm MV} \in \Theta^o$). La relation (II-3.10) implique que, pour tout $\theta_0 \in \Theta^o$,

$$n(\hat{\theta}_n^{\rm MV} - \theta_0)\mathbb{I}_n(\hat{\theta}_n^{\rm MV})(\hat{\theta}_n^{\rm MV} - \theta_0) \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \chi^2(d) \,,$$

ce qui implique, en utilisant que $\mathbb{I}_n(\theta_n^*) - \mathbb{I}_n(\hat{\theta}_n^{\rm MV}) \stackrel{\mathbb{P}_{n,\theta_0}-{\rm prob}}{\longrightarrow} \mathbf{0}_d$,

$$2n\,\ell_n(\hat{\theta}_n^{\rm MV}) - 2n\,\ell_n(\theta_0) \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \chi^2(d) \,.$$

Par conséquent, nous avons, pour tout $\theta_0 \in \Theta^o$,

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}(2n\,\ell_n(\hat{\theta}_n^{\rm MV}) - 2n\,\ell_n(\theta_0) < \chi^2_{1-\alpha}(d)) = 1-\alpha.$$

Si nous définissons la région

$$\check{S}_{n,\alpha} := \left\{\theta \in \Theta : \; 2n\,\ell_n(\hat{\theta}_n^{\rm MV}) - 2n\,\ell_n(\theta) \le \chi^2_{1-\alpha}(d)\right\} \tag{II-3.13}$$

alors, pour tout $\theta \in \Theta$, nous avons $\lim_{n\to\infty} \mathbb{P}_{n,\theta}(\theta \in \check{S}_{n,\alpha}) = 1-\alpha$.

Si g est une fonction de $\mathbb{R}^d \to \mathbb{R}$, on peut aussi être amené à construire des intervalles de confiance pour $g(\theta)$. Soit $\theta_0 \in \Theta^o$. Si g est différentiable au point θ_0, alors (II-3.3) implique, en utilisant la méthode-δ (théorème IV-5.47), que

$$\sqrt{n}\;\sigma_g^{-1}(\hat{\theta}_n^{\rm MV})\;\{g(\theta_0) - g(\hat{\theta}_n^{\rm MV})\} \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} {\rm N}(0,1) \quad \text{où} \quad \sigma_g(\theta)^2 := [\nabla g(\theta)]^T \mathbb{I}^{-1}(\theta) \nabla g(\theta) \,.$$

Par conséquent,

$$\left[g(\hat{\theta}_n^{\rm MV}) - n^{-1/2} z_{1-\alpha/2}\sigma_g(\hat{\theta}_n^{\rm MV}), \hat{\theta}_n^{\rm MV} + n^{-1/2} z_{1-\alpha/2}\sigma_g(\hat{\theta}_n^{\rm MV})\right] \tag{II-3.14}$$

est un intervalle de confiance de niveau de couverture asymptotique $1-\alpha$ pour $g(\theta_0)$. Lorsque l'information de Fisher n'est pas calculable, nous remplaçons cette quantité par l'information de Fisher observée. En utilisant (II-3.10), nous avons

$$\sqrt{n}\;\hat{\sigma}_{n,g}^{-1}(\hat{\theta}_n^{\rm MV})\;\{g(\theta_0) - g(\hat{\theta}_n^{\rm MV})\} \stackrel{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} {\rm N}(0,1) \quad \text{où} \quad \hat{\sigma}_{n,g}^2(\theta) := [\nabla g(\theta)]^T \mathbb{I}_n^{-1}(\theta) \nabla g(\theta) \,.$$

ce qui conduit à un intervalle de confiance asymptotique

$$\left[g(\hat{\theta}_n^{\rm MV}) - n^{-1/2} z_{1-\alpha/2}\hat{\sigma}_{n,g}(\hat{\theta}_n^{\rm MV}), \hat{\theta}_n^{\rm MV} + n^{-1/2} z_{1-\alpha/2}\hat{\sigma}_{n,g}(\hat{\theta}_n^{\rm MV})\right] \tag{II-3.15}$$

II-3.3 Efficacité asymptotique

Nous avons étudié au chapitre précédent la construction d'estimateurs basés sur la maximisation d'un critère ,

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} n^{-1} \sum_{i=1}^n m(\theta, X_i), \quad m : \Theta \times \mathsf{X} \to \mathbb{R}$$

ou sur la résolution d'un système d'équations

$$n^{-1}\sum_{i=1}^{n}\psi(\theta,X_i)=0,\quad \psi:\Theta\times\mathsf{X}\to\mathbb{R}^d\ .$$

Nous avons montré que, sous des hypothèses de régularité, la suite d'estimateurs $\{\hat\theta_n, n\in\mathbb{N}\}$ était consistante et asymptotiquement normale

$$\sqrt{n}\big(\hat\theta_n-\theta\big)\overset{\mathbb{P}_{n,\theta}}{\Longrightarrow}\mathrm{N}\big(0,V(\theta)\big) \tag{II-3.16}$$

avec $V(\theta)$ une matrice symétrique. Par exemple, sous des hypothèses de régularité sur le modèle statistique, l'estimateur du maximum de vraisemblance est asymptotiquement normal de variance l'inverse de l'information de Fisher. Nous allons montrer que cette variance est minimale parmi la classe des Z-estimateurs (ou M-estimateurs réguliers) et ce résultat nous fournira une notion d'optimalité associée aux modèles réguliers.

Dans cette section

— nous nous plaçons dans le cas de la dimension 1, avec $\Theta\subseteq\mathbb{R}$ pour simplifier. Les extensions au cas multidimensionnel se font de la même manière que pour la section II-3.2.

— nous nous restreignons à la classe des estimateurs asymptotiquement normaux, c'est-à-dire les estimateurs $\hat\theta_n$ pour lesquels, pour tout $\theta\in\Theta$, il existe une *variance asymptotique* $v(\theta)>0$ telle que

$$\sqrt{n}\big(\hat\theta_n-\theta\big)\overset{\mathbb{P}_{n,\theta}}{\Longrightarrow}\mathrm{N}\big(0,v(\theta)\big)\ .$$

On suppose de plus :

***H* II-3.7.** L'application $\theta\mapsto v(\theta)$ est continue sur Θ. ♢

Sous des hypothèses de régularité, on a vu que les M-estimateurs sont asymptotiquement normaux et vérifient **H** II-3.7. En particulier, pour l'estimateur du maximum de vraisemblance,

$$v(\theta)=\frac{1}{\mathbb{I}(\theta)}.$$

On a la règle de comparaison suivante.

Définition II-3.8. *Si $\{\widehat\theta_{n,1}, n\in\mathbb{N}\}$ et $\{\widehat\theta_{n,2}, n\in\mathbb{N}\}$ sont deux suites d'estimateurs asymptotiquement normaux de variances asymptotiques respectives $v_1(\theta)$ et $v_2(\theta)$ et vérifiant **H** II-3.7, on dit que $\widehat\theta_{n,1}$ est* plus efficace *que $\widehat\theta_{n,2}$ si pour tout $\theta\in\Theta$,*

$$v_1(\theta)\le v_2(\theta)$$

et si de plus, il existe un point $\widetilde\theta\in\Theta$ tel que

$$v_1(\widetilde\theta)<v_2(\widetilde\theta).$$

Une suite d'estimateurs $\{\hat\theta_n, n\in\mathbb{N}\}$ est asymptotiquement efficace *s'il n'existe pas d'autre estimateurs (dans la classe considérée) plus efficace que $\hat\theta_n$.*

Remarque II-3.9. L'hypothèse de normalité asymptotique en tout point $\theta\in\Theta$ permet en particulier d'exclure les estimateurs artificiels de la forme $\hat\theta_n=\theta_0$ pour un point $\theta_0\in\Theta$ arbitraire, qui sont catastrophiques pour le risque quadratique en dehors d'un "petit " voisinage de θ_0 mais qui ont un risque nul en θ_0. ♢

II-3.3.1 Efficacité asymptotique du maximum de vraisemblance

Dans cette section, on considère une expérience statistique régulière (au sens de la Définition I-4.16), et on se restreint à la classe des Z-estimateurs, qui contient en particulier les M-estimateurs.

Théorème II-3.10 (Efficacité asymptotique du M.V. dans la classe des Z-estimateurs). *Soit une suite d'expériences statistiques produit, du modèle $(\mathsf{X}, \mathscr{X}, \{p_\theta \cdot \mu\,,\ \theta \in \Theta \subseteq \mathbb{R}\})$ supposé régulier (voir définition I-4.16). Soit $\hat{\theta}_n$ un Z-estimateur associé à une fonction ψ vérifiant **H** II-2.12.*

La variance asymptotique de $\hat{\theta}_n$, donnée par le théorème II-2.15, vérifie

$$v_\psi(\theta) = \frac{\int \psi^2(\theta,x)\mathbb{P}_\theta(\mathrm{d}x)}{\left(\int \psi'(\theta,x)\,\mathbb{P}_\theta(\mathrm{d}x)\right)^2} \geq \frac{1}{\mathbb{I}(\theta)}\,.$$

Corollaire II-3.11. *Dans un modèle régulier, l'estimateur du maximum de vraisemblance est asymptotiquement efficace parmi les Z-estimateurs réguliers.*

Démonstration. Notons $\psi'(\theta,x)$ la dérivée de $\theta \mapsto \psi(\theta,x)$; et $\dot{\ell}(\theta,x)$ la dérivée de $\theta \mapsto \ell(\theta,x)$.

Puisque $\mathbb{E}_\theta\left[\psi(\theta,X_1)\right] = 0$ pour tout $\theta \in \Theta^o$, la dérivée est nulle, ce qui donne

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \psi'(\theta,x)\, p_\theta(x)\mu(\mathrm{d}x) + \int_{\mathbb{R}} \psi(\theta,x)\frac{\partial p_\theta(x)}{\partial\theta}\mu(\mathrm{d}x) \\ &= \int_{\mathbb{R}} \psi'(\theta,x)\, p_\theta(x)\mu(\mathrm{d}x) + \int_{\mathbb{R}} \psi(\theta,x)\dot{\ell}(\theta,x)p_\theta(x)\mu(\mathrm{d}x), \end{aligned}$$

c'est-à-dire

$$\int \psi'(\theta,x)\mathbb{P}_\theta(\mathrm{d}x) = -\int \psi(\theta,x)\,\dot{\ell}(\theta,x)\mathbb{P}_\theta(\mathrm{d}x)\,.$$

En appliquant l'inégalité de Cauchy-Schwarz, on obtient

$$\left(\int \psi'(\theta,x)\mathbb{P}_\theta(\mathrm{d}x)\right)^2 \leq \left(\int \psi(\theta,x)^2\mathbb{P}_\theta(\mathrm{d}x)\right)\left(\int (\dot{\ell}(\theta,x))^2\mathbb{P}_\theta(\mathrm{d}x)\right)\,,$$

ce qui implique

$$v_\psi(\theta)^{-1} = \frac{\left(\int \psi'(\theta,x)\mathbb{P}_\theta(\mathrm{d}x)\right)^2}{\int \psi(\theta,x)^2\mathbb{P}_\theta(\mathrm{d}x)} \leq \int \left(\dot{\ell}(\theta,X)\right)^2\mathbb{P}_\theta(\mathrm{d}x) = \mathbb{I}(\theta)\,. \qquad \square$$

II-3.3.2 La conjecture de Fisher

En 1922, Fisher conjectura que, pour un modèle régulier, dans un sens comparable avec celui donné à la définition I-4.16,

(i) l'estimateur du maximum de vraisemblance converge et a pour variance asymptotique $\frac{1}{\mathbb{I}(\theta)}$.

(ii) si, pour une suite d'estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$, on a $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}\left(0, v(\theta)\right)$, alors nécessairement $v(\theta) \geq \frac{1}{\mathbb{I}(\theta)}$.

Le programme de Fisher aurait permis, parmi une classe d'estimateurs raisonnables, de clore le débat sur l'optimalité asymptotique. On a vu que le point (i) de la conjecture de Fisher est vérifié. On a montré que le point (ii) est vrai parmi la classe restreinte des Z-estimateurs réguliers.

Mais la conjecture de Fisher est fausse en général : *pour tout estimateur asymptotiquement normal, on peut construire un estimateur modifié plus efficace*. Une construction classique, le contre-exemple de Hodges-Lehmann montre que l'on peut trouver un estimateur qui est asymptotiquement "meilleur" que l'estimateur du maximum de vraisemblance.

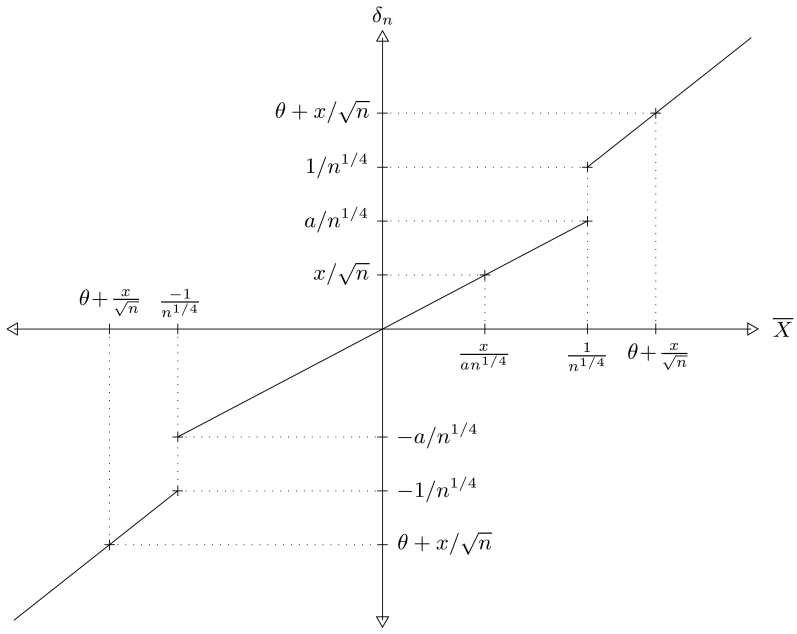

FIGURE II-3.1 – Estimateur de Hodges-Lehmann

Soit une suite d'expériences statistiques produit, du modèle gaussien

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}), \{p_\theta \cdot \lambda_{\text{Leb}}\,,\ \theta \in \Theta := \mathbb{R}\})$$

où p_θ est la densité d'une loi $\mathrm{N}(\theta, 1)$.

Posons $\overline{X}_n := n^{-1}\sum_{i=1}^n X_i$. Définissons l'estimateur δ_n

$$\delta_n := \begin{cases} \overline{X}_n, & |\overline{X}_n| \geq 1/n^{1/4}; \\ a\overline{X}_n, & |\overline{X}_n| < 1/n^{1/4}, \end{cases}$$

où a est une constante de $]0,1[$.

• Pour tout $\theta \neq 0$, déterminons la distribution de $\sqrt{n}(\delta_n - \theta)$ sous $\mathbb{P}_{n,\theta}$. Supposons tout d'abord que $\theta < 0$. Fixons x et considérons

$$\mathbb{P}_{n,\theta}(\sqrt{n}(\delta_n - \theta) \leq x) = \mathbb{P}_{n,\theta}(\delta_n \leq \theta + x/\sqrt{n})\ .$$

Comme $\theta + x/\sqrt{n} \to \theta < 0$ et $-1/n^{1/4} \to 0$, pour n suffisamment grand, $\theta + x/\sqrt{n} < -1/n^{1/4}$, et donc

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta}(\sqrt{n}(\delta_n - \theta) \leq x) = \lim_{n\to\infty} \mathbb{P}_{n,\theta}(\overline{X}_n \leq \theta + x/\sqrt{n}) = \Phi(x)\ ,$$

où Φ désigne la fonction de répartition d'une loi $\mathrm{N}(0,1)$. Par conséquent, $\sqrt{n}(\delta_n - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1)$. Un calcul similaire montre que $\sqrt{n}(\delta_n - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1)$ pour $\theta > 0$.

• Supposons maintenant que $\theta = 0$. Fixons x et considérons

$$\mathbb{P}_{n,0}(\sqrt{n}\delta_n \leq x) = \mathbb{P}_{n,0}(\delta_n \leq x/\sqrt{n})\ .$$

On a

$$\sqrt{n}\delta_n = \sqrt{n}\bar{X}_n(1-a)\mathbb{1}_{\{|\sqrt{n}\bar{X}_n| \geq n^{1/4}\}} + a\sqrt{n}\bar{X}_n\ .$$

Or, sous $\mathbb{P}_{n,0}$, $\sqrt{n}\bar{X}_n \sim \mathrm{N}(0,1)$; cela entraine que $\mathbb{1}_{\{|\sqrt{n}\bar{X}_n| \geq n^{1/4}\}} \overset{\mathbb{P}_{n,0}-\text{prob}}{\longrightarrow} 0$, puis, par le lemme de Slutsky, (voir lemme IV-5.33) on a

$$\sqrt{n}\delta_n \overset{\mathbb{P}_{n,0}}{\Longrightarrow} \mathrm{N}(0, a^2)\ .$$

• Par conséquent, nous avons, pour tout $\theta \in \mathbb{R}$,

$$\sqrt{n}(\delta_n - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,\ \sigma^2(\theta))\,, \tag{II-3.17}$$

où

$$\sigma^2(\theta) = \begin{cases} 1, & \theta \neq 0 \\ a^2, & \theta = 0 \end{cases}\,. \tag{II-3.18}$$

Cet estimateur est "super efficace" car sa variance asymptotique pour $\theta = 0$ est inférieure à $1/\mathbb{I}(\theta) = 1$.

• Comme $\sqrt{n}(\overline{X}_n - \theta) \sim \mathrm{N}(0,\ 1)$, (II-3.17) semble suggérer que δ_n pourrait être asymptotiquement un meilleur estimateur que $\overline{X}_n$, au moins lorsque n est grand. Pour comprendre ce phénomène, nous allons évaluer le risque de ces deux estimateurs pour la perte quadratique. Comme $R_n(\theta,\ \overline{X}_n) = \mathbb{E}_{n,\theta}[(\overline{X}_n - \theta)^2] = 1/n$, $nR_n(\theta,\ \overline{X}_n) = 1$. Un calcul élémentaire montre que

$$nR_n(\theta,\ \delta_n) \to \begin{cases} 1, & \theta \neq 0 \\ a^2, & \theta = 0 \end{cases}\,,$$

comme le suggère d'ailleurs (II-3.18). La comparaison de δ_n et $\overline{X}_n$ par leur variance asymptotique ou par leur risque quadratique (normalisé par n) ne donne pas une vision complète, car la convergence n'est pas uniforme en θ. Notons en effet (voir fig. II-3.1) que l'estimateur δ_n ne prend jamais de valeurs dans l'intervalle

$$\left]\, \frac{a}{n^{1/4}},\ \frac{1}{n^{1/4}} \right[\,.$$

Si nous définissons

$$\theta_n := \frac{1+a}{2n^{1/4}}$$

le milieu de cet intervalle, alors $|\delta_n - \theta_n|$ est toujours plus grand que la moitié de la longueur de cet intervalle, et donc

$$(\delta_n - \theta_n)^2 \geq (\frac{1-a}{2n^{1/4}})^2 = \frac{(1-a)^2}{4\sqrt{n}}.$$

Nous en déduisons

$$nR_n(\theta_n,\ \delta_n) \geq n\frac{(1-a)^2}{4\sqrt{n}} = \frac{(1-a)^2}{4}\sqrt{n} \to_{n\to\infty} \infty.$$

Ceci montre que, lorsque $n \geq 1$, le risque de l'estimateur δ_n au point θ_n est considérablement plus grand que le risque de $\overline{X}_n$ au point θ_n (voir fig. II-3.2). L'amélioration du risque quadratique au point $\theta = 0$ se paye donc par un accroissement considérable du risque dans un voisinage de 0.

En conclusion, le développement d'une théorie satisfaisante de la comparaison des estimateurs passe par la comparaison des risques non pas en un point, mais sur des voisinages de ces points, pour éviter les comportements pathologiques.

II-3.4 Pour aller plus loin : Méthode du score de Fisher

Dans un modèle régulier, l'estimateur du maximum de vraisemblance est " meilleur " que n'importe quel autre Z-estimateur au sens de l'efficacité asymptotique. Pourtant, il est parfois plus facile de mettre en oeuvre un Z-estimateur donné (ou d'ailleurs un M-estimateur) plutôt que l'estimateur du maximum de vraisemblance, voir l'exemple I-2.3 du modèle de Cauchy.

On peut modifier un estimateur $\hat{\theta}_n$ consistant et asymptotiquement normal de sorte qu'il ait asymptotiquement le même comportement que l'estimateur du maximum de vraisemblance.

Lemme II-3.12. *Supposons le modèle* $(\mathsf{X}, \mathscr{X}, \{p_\theta \cdot \mathrm{d}\mu\,,\ \theta \in \Theta\})$ *régulier. Soit* $\{\hat{\theta}_n, n \in \mathbb{N}\}$ *une suite d'estimateurs asymptotiquement normale. Alors pour tout* $\theta_0 \in \Theta$ *et* $\varepsilon > 0$*,*

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta_0}\left(n^{1/2}|\nabla \ell_n(\hat{\theta}_n) - \{\nabla \ell_n(\theta_0) - \mathbb{I}(\theta_0)(\hat{\theta}_n - \theta_0)\}| \geq \varepsilon\right) = 0\,,$$

où $\nabla \ell_n(\theta) := n^{-1}\sum_{i=1}^n \nabla \ell(\theta, X_i)$.

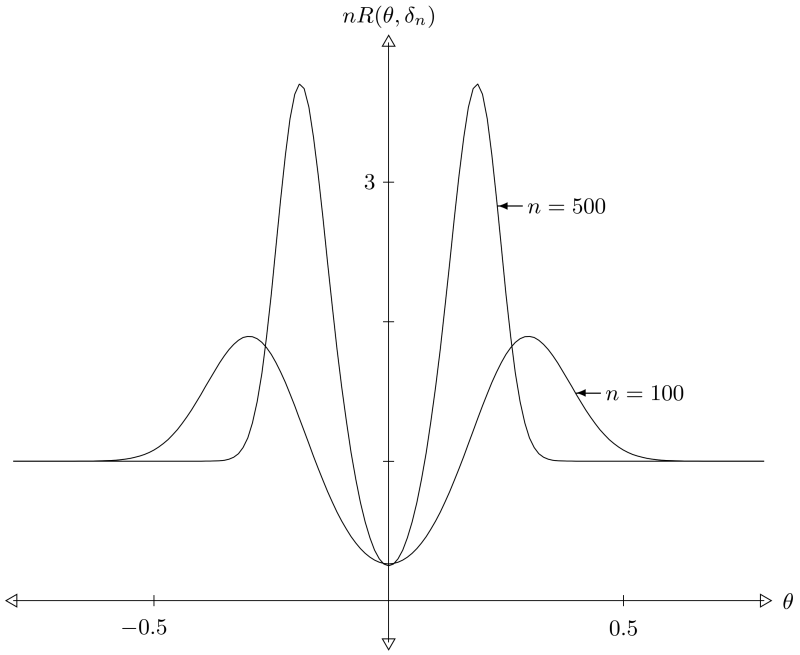

FIGURE II-3.2 – $\theta \mapsto n\,\mathbb{E}_{n,\theta}[(\delta_n - \theta)^2]$ pour $n = 100$ et $n = 500$, $a = 1/2$

Démonstration. Nous avons

$$\begin{aligned}
\nabla \ell_n(\hat\theta_n) &= n^{-1}\sum_{i=1}^n \nabla\ell(\hat\theta_n, X_i) \\
&= n^{-1}\sum_{i=1}^n \nabla\ell(\theta_0, X_i) + n^{-1}\sum_{i=1}^n \mathrm{H}_\ell(\theta_0, X_i)(\hat\theta_n - \theta_0) \\
&\quad + n^{-1}\sum_{i=1}^n \left[\int_0^1 \{\mathrm{H}_\ell(\theta_0 + w(\hat\theta_n - \theta_0), X_i) - \mathrm{H}_\ell(\theta_0, X_i)\}\mathrm{d}w\right](\hat\theta_n - \theta_0) \\
&= n^{-1}\sum_{i=1}^n \nabla\ell(\theta_0, X_i) - \mathbb{I}_n(\theta_0)\{\hat\theta_n - \theta_0\} + \Delta_n(\hat\theta_n - \theta_0)\ ,
\end{aligned}$$

où

$$\Delta_n := -\int_0^1 \{\mathbb{I}_n(\theta_0 + w(\hat\theta_n - \theta_0)) - \mathbb{I}_n(\theta_0)\}\mathrm{d}w\ .$$

Par conséquent, en utilisant la proposition II-3.6, nous avons, pour tout $\varepsilon > 0$,

$$\mathbb{P}_{n,\theta_0}(\|\Delta_n\| \geq \varepsilon) \leq \mathbb{P}_{n,\theta_0}(\left\|\hat\theta_n - \theta_0\right\| \geq \delta_n) + \mathbb{P}_{n,\theta_0}\left(\sup_{\|t\|\leq\delta_n} \|\mathbb{I}_n(\theta_0 + t) - \mathbb{I}_n(\theta_0)\| \geq \varepsilon\right) \to_{n\to\infty} 0\ ,$$

ce qui montre que $\Delta_n \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0$. Comme $\{\hat\theta_n, n \in \mathbb{N}\}$ est asymptotiquement normal, nous avons, en utilisant le lemme IV-5.33, que

$$\sqrt{n}\Delta_n(\hat\theta_n - \theta_0) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0\ ,$$

ou de façon plus concise $\Delta_n(\hat\theta_n - \theta) = o_P(n^{-1/2})$.

Par la loi des grands nombres, nous avons, pour tout $\theta_0 \in \Theta$, $\mathbb{I}_n(\theta_0) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \mathbb{I}(\theta_0)$ et donc, en utilisant le lemme IV-5.33,

$$\sqrt{n}\{\mathbb{I}_n(\theta_0) - \mathbb{I}(\theta_0)\}(\hat\theta_n - \theta) \overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} 0\ .$$

□

Théorème II-3.13 (Score de Fisher). *Supposons le modèle* $(\mathsf{X}, \mathscr{X}, \{p_\theta \cdot \mathrm{d}\mu, \theta \in \Theta\})$ *régulier. Soit* $\{\hat\theta_n, n \in \mathbb{N}\}$ *un estimateur asymptotiquement normal. Considérons l'estimateur*

$$\tilde\theta_n = \hat\theta_n + \mathbb{I}_n(\hat\theta_n)^{-1} \nabla \ell_n(\hat\theta_n) \,,$$

où $\mathbb{I}_n(\hat\theta_n)$ *est l'information de Fisher observée au point* $\hat\theta_n$ *et* $\nabla \ell_n(\theta) := n^{-1} \sum_{i=1}^n \nabla \ell(\theta, X_i)$.

Alors, pour tout $\theta_0 \in \Theta$,

$$\sqrt{n}\big(\widetilde{\theta}_n - \theta_0\big) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}\Big(0, \mathbb{I}^{-1}(\theta_0)\Big). \tag{II-3.19}$$

Le choix initial pourra donc être un M- ou Z-estimateur consistant et asymptotiquement normal, sans que l'on ait besoin de se soucier de sa variance asymptotique (Définition II-3.8).

Démonstration. Soit $\theta_0 \in \Theta$. Le lemme II-3.12 montre que,

$$n^{1/2} \left(\nabla \ell_n(\hat\theta_n) - \{ \nabla \ell_n(\theta_0) - \mathbb{I}(\theta_0)(\hat\theta_n - \theta_0) \} \right) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} 0 \,. \tag{II-3.20}$$

Comme $\sqrt{n} \nabla \ell_n(\theta_0) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} \mathrm{N}(0, \mathbb{I}(\theta_0))$ (noter que par la proposition I-4.19, $\mathbb{E}_{\theta_0}[\nabla \ell(\theta_0, X)] = 0$), nous avons, en appliquant lemme IV-5.58,

$$\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P}_{n,\theta_0}(\sqrt{n} \|\nabla \ell_n(\theta_0)\| \geq M) = 0 \,.$$

Nous avons de même, comme $\{\hat\theta_n, n \in \mathbb{N}\}$ est asymptotiquement normal, $\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}_{n,\theta_0}(\sqrt{n}\|\hat\theta_n - \theta_0\| \geq M) = 0$. Nous en déduisons que

$$\lim_{M \to \infty} \mathbb{P}_{n,\theta_0}(n^{1/2} \|\nabla \ell_n(\hat\theta_n)\| \geq M) = 0 \,. \tag{II-3.21}$$

Comme la fonction $A \mapsto A^{-1}$ est continue en toute matrice A_0 inversible, le théorème IV-5.6 et la proposition II-3.6 montrent que

$$\mathbb{I}_n^{-1}(\hat\theta_n) - \mathbb{I}_n^{-1}(\theta_0) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} 0 \,. \tag{II-3.22}$$

En combinant (II-3.21) et (II-3.22) nous obtenons donc

$$\begin{aligned} \tilde\theta_n &= \hat\theta_n + \mathbb{I}^{-1}(\theta_0) \nabla \ell_n(\hat\theta_n) + \{\mathbb{I}^{-1}(\hat\theta_n) - \mathbb{I}^{-1}(\theta_0)\} \nabla \ell_n(\hat\theta_n) \\ &= \hat\theta_n + \mathbb{I}^{-1}(\theta_0) \nabla \ell_n(\theta_0) - (\hat\theta_n - \theta_0) + \{\mathbb{I}^{-1}(\hat\theta_n) - \mathbb{I}^{-1}(\theta_0)\} \nabla \ell_n(\hat\theta_n) \,. \end{aligned}$$

Nous en déduisons

$$\sqrt{n}(\tilde\theta_n - \theta_0) - \mathbb{I}^{-1}(\theta_0) n^{1/2} \nabla \ell_n(\theta_0) \overset{\mathbb{P}_{n,\theta_0}}{\Longrightarrow} 0 \,, \tag{II-3.23}$$

ce qui montre (II-3.19). □

Exemple II-3.14. Une source émet des particules de type A avec probabilité θ et de type B avec probabilité $1-\theta$, où $\theta \in \Theta = (0,1)$. On mesure l'énergie des particules, qui est distribuée selon une densité f_1 pour les particules de type A et f_2 pour les particules de type B. Les densités f_1, f_2 sont connues. Si l'on détecte n particules avec des énergies $X_1, \ldots, X_n$, quelle est la valeur de θ ?

Nous allons tout d'abord formaliser de façon plus précise les hypothèses. Soient f_1 et f_2 deux densités par rapport à la mesure de Lebesgue sur $\mathbb{R}$ telles que $\int |f_1(x) - f_2(x)| \mathrm{d}x > 0$.

Soit une suite d'expériences statistiques produit du modèle $\{p_\theta(x) \cdot \lambda_{\mathrm{Leb}} \,:\, \theta \in \Theta := [0,1]\}$, où

$$p_\theta(x) := \theta f_1(x) + (1-\theta) f_2(x) \,.$$

La fonction de vraisemblance de

$$\mathrm{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^n \big(\theta f_1(X_i) + (1-\theta) f_2(X_i) \big),$$

de sorte que pour tout $\theta \in]0,1[$

$$\dot{\ell}_n(\theta, X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}(\theta, X_i) \tag{II-3.24}$$

où $\dot{\ell}(\theta, x)$ est le score

$$\dot{\ell}(\theta, x) := \frac{f_1(x) - f_2(x)}{\theta f_1(x) + (1-\theta) f_2(x)} = \frac{f_1(x) - f_2(x)}{\theta \{f_1(x) - f_2(x)\} + f_2(x)} \,. \tag{II-3.25}$$

Le score $\theta \mapsto \dot{\ell}(\theta, x)$ est une fonction décroissante sur $[0,1]$. Nous allons montrer que pour tout $\theta_0 \in [0,1]$ nous avons

$$\int \dot{\ell}(0,x) \mathbb{P}_{\theta_0}(\mathrm{d}x) > 0 \,,$$

(cette quantité étant éventuellement infinie). Pour que cette quantité soit bien définie, nous devons tout d'abord montrer que $\int \{\dot{\ell}(0,x)\}^- \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty$. Pour tout $\theta_0 \in]0,1[$,

$$\begin{aligned} \int \frac{\{f_1(x) - f_2(x)\}^-}{f_2(x)} \mathbb{P}_{\theta_0}(\mathrm{d}x) &= \int \frac{\{f_1(x) - f_2(x)\}^-}{f_2(x)} \{\theta_0 f_1(x) + (1-\theta_0) f_2(x)\} \, \mathrm{d}x \\ &\leq \theta_0 \int \{f_1(x) - f_2(x)\}^- \mathrm{d}x + (1-\theta_0) \int \{f_1(x) - f_2(x)\}^- \mathrm{d}x \,. \end{aligned}$$

Comme $\int \{f_1(x) - f_2(x)\} \mathrm{d}x = 0$, nous avons

$$\int \{f_1(x) - f_2(x)\}^+ \mathrm{d}x = \int \{f_1(x) - f_2(x)\}^- \mathrm{d}x \,,$$

ce qui implique que $\int \{f_1(x) - f_2(x)\}^- \mathrm{d}x = (1/2) \int |f_1(x) - f_2(x)| \mathrm{d}x$. Nous obtenons donc

$$\int \frac{\{f_1(x) - f_2(x)\}^-}{f_2(x)} \mathbb{P}_{\theta_0}(\mathrm{d}x) \leq \frac{1}{2} \int |f_1(x) - f_2(x)| \mathrm{d}x \,.$$

Un calcul élémentaire montre que

$$\begin{aligned} \int \dot{\ell}(0, X_1) \mathbb{P}_{\theta_0}(\mathrm{d}x) &= \int \frac{f_1(x) - f_2(x)}{f_2(x)} \{\theta_0 f_1(x) + (1-\theta_0) f_2(x)\} \mathrm{d}x \\ &= \theta_0 \int \frac{\{f_1(x) - f_2(x)\}^2}{f_2(x)} \mathrm{d}x \,, \end{aligned}$$

où nous avons utilisé que $\int \{f_1(x) - f_2(x)\} \mathrm{d}x = 0$.

De façon tout à fait similaire, nous pouvons montrer que $\int \{\dot{\ell}(1,x)\}^+ \mathbb{P}_{\theta_0}(\mathrm{d}x) < \infty$ et que

$$\int \dot{\ell}(1,x) \mathbb{P}_{\theta_0}(\mathrm{d}x) = -(1-\theta_0) \int \frac{\{f_1(x) - f_2(x)\}^2}{f_1(x)} \mathrm{d}x < 0 \,.$$

En appliquant la loi des grands nombres (Théorème IV-5.19), nous avons donc, pour tout $\theta_0 \in]0,1[$ que

$$\begin{aligned} n^{-1} \sum_{k=1}^{n} \dot{\ell}(0, X_k) &\overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \int \dot{\ell}(0,x) \mathbb{P}_{\theta_0}(\mathrm{d}x) > 0 \\ n^{-1} \sum_{k=1}^{n} \dot{\ell}(1, X_k) &\overset{\mathbb{P}_{n,\theta_0}-\text{prob}}{\longrightarrow} \int \dot{\ell}(1,x) \mathbb{P}_{\theta_0}(\mathrm{d}x) < 0 \end{aligned}$$

Notons que $\int \dot{\ell}(0,x) \mathbb{P}_{\theta_0}(\mathrm{d}x)$ peut être égal à $+\infty$ et $\int \dot{\ell}(1,x) \mathbb{P}_{\theta_0}(\mathrm{d}x)$ peut être égal à $-\infty$. Par conséquent, pour tout $\theta_0 \in \Theta$, l'équation de vraisemblance (II-3.24) admet une solution unique, qui correspond à un maximum, notée $\hat{\theta}_n^{\mathrm{MV}}$. On vérifie aisément que les conditions de régularité (Définition I-4.16) sont satisfaites. La quantité d'information de Fisher est donnée, pour tout $\theta \in \Theta$, par

$$\mathbb{I}(\theta) = \frac{1}{\theta(1-\theta)} \left[1 - \int \frac{f_1(x) f_2(x)}{\theta f_1(x) + (1-\theta) f_2(x)} \right]$$

On remarque que l'information de Fisher est maximale quand $f_1(x) f_2(x) =$ pour tout $x \in \mathbb{R}$, i.e. que les supports des lois du mélange sont disjointes. Dans ce cas, chaque observation X_i nous donne une information exacte sur la composante du mélange qui a été choisie. La quantité d'information correspond dans ce cas à celle d'une loi de Bernoulli de paramètre de succès θ.

La résolution de l'équation de vraisemblance associée est d'autant plus difficile que n est grand. Supposons que $\int_{\mathbb{R}} \big(F_1(x)-F_2(x)\big)^2 \mathrm{d}x < +\infty$, où $F_i(x) := \int_{-\infty}^{x} f_i(t)dt$, $i = 1,2$ est la fonction de répartition associée à la densité f_i. Soit $\hat{\theta}_n$ l'estimateur qui minimise

$$\theta \mapsto \int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_\theta(x)\big)^2 \mathrm{d}x,$$

où

$$F_\theta(x) := \theta F_1(x) + (1-\theta)F_2(x)\,, \qquad \widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le x}\,.$$

En dérivant par rapport à la variable θ, on obtient

$$\int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_\theta(x)\big)\big(F_1(x)-F_2(x)\big)\mathrm{d}x = 0,$$

d'où

$$\hat{\theta}_n = \frac{\int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_2(x)\big)\big(F_1(x)-F_2(x)\big)\mathrm{d}x}{\int_{\mathbb{R}} \big(F_1(x)-F_2(x)\big)^2 \mathrm{d}x}.$$

En utilisant le T.L.C. (Théorème IV-5.39), on peut montrer que $\hat{\theta}_n$ est asymptotiquement normal. Alors l'estimateur modifié

$$\widetilde{\theta}_n = \hat{\theta}_n - \frac{\dot{\ell}_n(\hat{\theta}_n, X_1, \ldots, X_n)}{\ddot{\ell}_n(\hat{\theta}_n, X_1, \ldots, X_n)}$$

où

$$\ddot{\ell}_n(\theta, X_1, \ldots, X_n) := -\frac{1}{n}\sum_{i=1}^{n} \frac{\big(f_1(X_i)-f_2(X_i)\big)^2}{\big(\theta f_1(X_i) + (1-\theta) f_2(X_i)\big)^2}$$

est asymptotiquement efficace, et sa variance asymptotique est l'information de Fisher du modèle

$$\mathbb{I}(\theta) = \int_{\mathbb{R}} \frac{\big(f_1(x)-f_2(x)\big)^2}{\theta f_1(x) + (1-\theta) f_2(x)} \mathrm{d}x.$$

♢

Troisième partie

Fondamentaux de l'apprentissage statistique

Chapitre III-1

Classification supervisée

III-1.1 La classification binaire

La *reconnaissance des formes* (ou *classification*) consiste à prédire la classe inconnue Y associée à une observation X. Nous considérons dans ce chapitre le problème de classification le plus simple, où le nombre de classes est égal à deux (chaque observation a deux étiquettes possibles). Il s'agit en fait d'un problème très classique. Par exemple, les courriels peuvent être classés comme étant des courriels légitimes ou des pourriels. Un patient dans un hôpital est en bonne santé ou malade, etc. Dans tous ces scénarios, l'ensemble des étiquettes possibles est donc $\mathsf{Y} = \{0,1\}$ ou $\mathsf{Y} = \{-1,+1\}$ (ce qui simplifie parfois certaines écritures). Pour classer une observation, nous devons utiliser des *attributs* de l'observation (en anglais, les "features"). L'espace des attributs X dépend évidemment du problème spécifique considéré :

— Pour la classification des courriels, il peut s'agir d'un *ensemble de mots* (ou *bag of words*). Dans la version la plus simple, un document est représenté par un vecteur de la taille du dictionnaire, la composante i indique le nombre d'occurrences du i-ème mot du dictionnaire dans le document.

— Pour le diagnostic du patient, il peut s'agir d'un ensemble de mesures recueillies par les médecins pour ce patient, par exemple, la fréquence cardiaque, la tension artérielle, la présence de symptômes spécifiques.

Par souci de simplicité, nous supposerons que l'espace des attributs est un sous-ensemble $\mathsf{X} \subset \mathbb{R}^d$ (bien que cela ne soit pas si important dans l'exposé qui suit ; on peut être amené à travailler par exemple avec des "attributs" plus structurés). L'extraction des attributs est une tâche difficile qui dépend fortement de la tâche de classification considérée. Dans ce court exposé, nous considérons que ces attributs sont connus.

Pour modéliser le problème d'apprentissage, nous considérons un cadre probabiliste. Nous supposons que l'observation (X,Y) (où X est le vecteur de attributs et Y est l'étiquette) est un vecteur aléatoire défini sur un espace de probabilité $(\Omega,\mathscr{F},\mathbb{P})$ à valeurs dans $\mathbb{R}^d \times \{0,1\}$. Nous notons par μ la loi du vecteur aléatoire X des attributs : pour tout ensemble mesurable $A \in \mathscr{B}(\mathbb{R}^d)$,

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

Notons $p = \mathbb{P}(Y = 1)$ la probabilité a priori de la classe 1 (nous avons donc $\mathbb{P}(Y = 0) = 1 - p$). Nous appellerons μ_0 et μ_1 les lois conditionnelles du vecteur des attributs sachant la classe : pour tout $A \in \mathscr{B}(\mathbb{R}^d)$ (en supposant pour éviter les trivialités que $p \in]0,1[$),

$$\mu_i(A) = \mathbb{P}(X \in A, Y = i)/\mathbb{P}(Y = i), \qquad \forall i \in \{0,1\}\ .$$

Nous avons donc, pour tout $A \in \mathscr{B}\mathbb{R}^d$

$$\mu(A) = p\mu_0(A) + (1-p)\mu_1(A)\ .$$

Remarquons que $\mu(A) = 0$ implique $\mu_0(A) = 0$ et $\mu_1(A) = 0$; par conséquent $\mu_0 \ll \mu$ et $\mu_1 \ll \mu$. Nous notons par $p_i = \mathrm{d}\mu_i/\mathrm{d}\mu$, $i \in \{0,1\}$, la densité de μ_i par rapport à μ. Nous avons, pour tout $A \in \mathscr{B}(\mathbb{R}^d)$,

$$\mu_i(A) = \int \mathbb{1}_A(x) p_i(x) \mu(\mathrm{d}x)\ , \quad i \in \{0,1\}\ .$$

La loi jointe du couple (X,Y) admet donc une densité $h : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}_+$ par rapport à la mesure produit $\mu \otimes c$, où c est la mesure de comptage sur $\{0,1\}$:

$$h_{X,Y}(x,y) = (1-p)p_0(x)\mathbb{1}_{\{0\}}(y) + pp_1(x)\mathbb{1}_{\{1\}}(y) .$$

En utilisant la règle de Bayes, nous pouvons définir la densité conditionnelle donnée, pour tout $x \in \left\{x \in \mathbb{R}^d : pp_0(x) + (1-p)p_1(x) \right.$ par

$$h_{Y|X}(y|x) = \frac{h_{X,Y}(x,y)}{(1-p)p_0(x) + pp_1(x)} , \quad y \in \{0,1\} .$$

Nous notons, pour tout $x \in \mathbb{R}^d$,

$$\eta(x) = h_{Y|X}(1|x) = \frac{pp_1(x)}{(1-p)p_0(x) + pp_1(x)} . \tag{III-1.1}$$

Cette fonction est appelée *probabilité a posteriori*.

De façon générale, soit λ une mesure σ-finie sur $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Supposons que la loi du couple (X,Y) admet une densité $h_{X,Y}$ par rapport à $\mu \otimes \mathrm{c}$, où c est la mesure de comptage sur $\{0,1\}$. Pour toute fonction $f : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}_+$, nous avons par définition

$$\mathbb{E}[f(X,Y)] = \iint f(x,y)h_{X,Y}(x,y)\lambda(\mathrm{d}x)\,\mathrm{c}(\mathrm{d}y) = \int f(x,0)h_{X,Y}(x,0)\lambda(\mathrm{d}x) + \int f(x,1)h_{X,Y}(x,1)\lambda(\mathrm{d}x) .$$

La loi marginale des attributs X admet une densité par rapport à mesure λ donnée par

$$h_X(x) = \int h_{X,Y}(x,y)\,\mathrm{c}(\mathrm{d}y) = h_{X,Y}(x,0) + h_{X,Y}(x,1) .$$

La loi marginale du label Y est donnée, pour $y \in \{0,1\}$ par

$$h_Y(y) = \int h_{X,Y}(x,y)\lambda(\mathrm{d}x) .$$

Ainsi, $\mathbb{P}(Y=1) = h_Y(1) = \int h_{X,Y}(x,1)\lambda(\mathrm{d}x)$ et $\mathbb{P}(Y=0) = h_Y(0) = \int h_{X,Y}(x,0)\lambda(\mathrm{d}x)$. Pour tout $x \in \mathbb{R}^d$ tel que $h_X(x) \neq 0$, nous notons

$$\eta(x) = \frac{h_{X,Y}(x,1)}{h_X(x)} . \tag{III-1.2}$$

Par convention, si $h_X(x) = 0$, nous posons $\eta(x) = 0$. Nous avons, pour tout $f : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}_+$,

$$\begin{aligned}
\mathbb{E}[f(X,Y)] &= \iint f(x,y)h_{X,Y}(x,y)\lambda(\mathrm{d}x)\,\mathrm{c}(\mathrm{d}y) \\
&= \int_{\left\{x \in \mathbb{R}^d : h_X(x) \neq 0\right\}} h_X(x) \int f(x,y)\frac{h_{X,Y}(x,y)}{h_X(x)}\,\mathrm{c}(\mathrm{d}y)\lambda(\mathrm{d}x) \\
&= \int_{\left\{x \in \mathbb{R}^d : h_X(x) \neq 0\right\}} h_X(x)\left\{f(x,1)\eta(x) + f(x,0)(1-\eta(x))\right\}\lambda(\mathrm{d}x) .
\end{aligned}$$

Définition III-1.1 (Classifieur, probabilité d'erreur). *Une fonction $g : \mathbb{R}^d \to \mathsf{Y}$ (où $\mathsf{Y} = \{-1,+1\}$ ou $\mathsf{Y} = \{0,1\}$) définit un* classifieur *ou une* règle de classification. *Le résultat du classifieur est erroné si $g(X) \neq Y$, et la* probabilité d'erreur *ou* risque de classification *(que l'on abrégera en* risque *lorsqu'il n'y a pas d'ambiguïté) pour un classifieur g est donnée par*

$$R(g) = \mathbb{P}\left(g(X) \neq Y\right) . \tag{III-1.3}$$

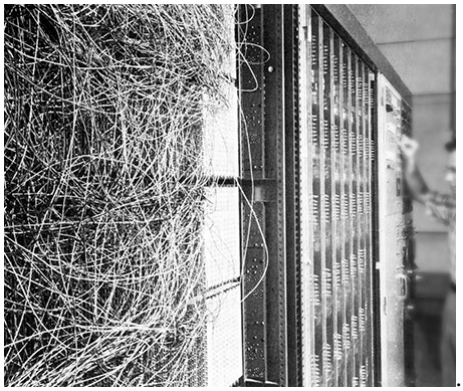

FIGURE III-1.1 – F. Rosenblatt avec le capteur d'image du Mark I Perceptron (source : Arvin Carlspan, Advanced technology center).

Exemple III-1.2 (Discrimination linéaire). Un discriminateur linéaire divise l'espace des attributs par un hyperplan et assigne une classe différente à chaque demi-espace. De telles règles offrent d'énormes avantages - elles sont faciles à interpréter. Le discriminateur linéaire est le classifieur donné par

$$g(x) = \sigma\left(a_0 + \sum_{i=1}^{d} a_i x^{(i)}\right) = \sigma(\mathbf{a}^T \mathbf{x}) \,,$$

où $\mathbf{x} = (1, x^{(1)}, \ldots, x^{(d)})$ est le vecteur des attributs, $\mathbf{a} = (a_0, \ldots, a_d)$ est le vecteur de *poids synaptiques* et σ est la fonction d'activation que nous prenons ici égale à la fonction de Heaviside

$$\text{heaviside}(x) = \begin{cases} 0 & x \leq 0 \\ +1 & x > 0 \end{cases} . \tag{III-1.4}$$

♢

Le coefficient a_0 est souvent appelé le biais. On trouve souvent une variante de ce modèle dans lequel la sortie prend les valeurs -1 et $+1$: il suffit d'utiliser la fonction $\sigma(x) = 2\,\text{heaviside}(x) - 1$. Le vecteur des poids synaptiques détermine l'importance relative des coordonnées du vecteur de attributs. La décision est également facile à mettre en œuvre - dans une solution logicielle standard, le temps de décision est proportionnel à d.

Rosenblatt ([**?**, **?**]) a réalisé l'énorme potentiel de telles règles linéaires et les a appelées *perceptrons*. Il a proposé des règles permettant de modifier le vecteur de poids au fur et à mesure que de nouvelles données arrivent, permettant ainsi à ces algorithmes d'apprendre de façon séquentielle et donnant naissance au premier algorithme d'intelligence artificielle.

Exemple III-1.3 (Réseau de neurones). Le discriminateur linéaire (perceptron monocouche) prend une décision $g(\mathbf{x}) = \sigma(\psi(\mathbf{x}))$ où σ est la fonction d'activation et

$$\psi(\mathbf{x}) = c_0 + \sum_{i=1}^{d} c_i x^{(i)} \,. \tag{III-1.5}$$

Le perceptron monocouche est un cas particulier d'un réseau de neurones sans couche cachée (appelé aussi *perceptron multicouche* ou *réseau connexionniste* ; voir [**?**, **?**]).

L'unité de traitement élémentaire dans un réseau de neurones n'est capable de réaliser que certaines opérations simples. Ces unités sont souvent appelées *neurones formels* pour leur similitude grossière avec les neurones du cerveau. Les modèles de réseaux connexionnistes qui nous intéressent particulièrement, les réseaux multicouches, classent les unités selon qu'elles sont des neurones d'entrée, cachés, ou de sortie.

— Un neurone d'entrée ou, simplement, une entrée, est une unité chargée de transmettre une composante du vecteur $\mathbf{x}$ des attributs .

— Un neurone de sortie est une unité qui fournit une hypothèse d'apprentissage, par exemple dans un problème de classification binaire, une valeur 0 ou 1.

— Enfin, un neurone caché est un neurone qui n'est ni un neurone d'entrée, ni un neurone de sortie. Sa fonction est de faire des traitements intermédiaires.

Dans un perceptron monocouche (ou discriminateur linéaire), il n'y pas de neurone caché et les neurones d'entrée et de sortie sont confondus. Considérons maintenant un réseau connexionniste avec une couche cachée. Dans ce cas, nous aurons deux types de neurones, des neurones d'entrée et des neurones de sortie, mais pas de neurones cachés. Supposons que nous avons k neurones d'entrée. Chaque neurone d'entrée $i \in \{1,\dots,k\}$ calcule une valeur

$$u_i = \sigma(\psi_i(\mathbf{x})) \,, \tag{III-1.6}$$

où chaque fonction ψ_i est de la forme donnée dans (III-1.5) :

$$\psi_i(\mathbf{x}) = b_i + \sum_{j=1}^{d} a_{i,j} x^{(j)} \,,$$

pour des poids $a_{i,j}$; σ est la *fonction d'activation*. Cette fonction d'activation est supposée être croissante et vérifier $\lim_{z\to-\infty}\sigma(z) = 0$, $\lim_{z\to+\infty}\sigma(z) = +1$. Des exemples de fonctions d'activation incluent le *signe*, $\sigma(x) = \text{signe}(x)$, la fonction logistique (ou sigmoïde)

$$\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}} \,, \tag{III-1.7}$$

l'arc tangente

$$\sigma(x) = (1/2)\left\{\frac{2}{\pi}\arctan(x) + 1\right\} \,.$$

Le neurone de sortie effectue une combinaison linéaire des sorties calculées par chaque neurone d'entrée

$$\psi(\mathbf{u}) = c_0 + \sum_{i=1}^{k} c_i u_i \,, \qquad \mathbf{u} = (u_1,\dots,u_k) \,,$$

où c_0 est le biais et $(c_1,\dots,c_k)$ sont des poids synaptiques ; puis il prend la décision en appliquant à ce résultat la fonction heaviside. La règle de décision pour un réseau de neurones à une couche cachée peut s'écrire

$$g(\mathbf{x}) = \text{heaviside}\left(c_0 + \sum_{i=1}^{k} c_i \sigma(\psi_i(\mathbf{x}))\right) \,. \tag{III-1.8}$$

Nous pouvons maintenant itérer ce procédé et introduire des couches de neurones cachés entre les neurones d'entrée et de sortie. On peut créer de cette façon des réseaux connexionnistes multicouches. Considérons un réseau à deux couches cachées avec k neurones d'entrée et ℓ neurones cachés. Nous calculons tout d'abord les sorties des k neurones d'entrée, pour $i \in \{1,\dots,k\}$,

$$u_i = \sigma\left(b_i + \sum_{j=1}^{d} a_{i,j} x^{(j)}\right) \,,$$

où σ est une fonction d'activation. Les valeurs $(u_1,\dots,u_k)$ sont les entrées des neurones cachés qui calculent pour $i \in \{1,\dots,\ell\}$,

$$z_i = \sigma\left(d_{i,0} + \sum_{j=1}^{k} d_{i,j} u_j\right) \,,$$

où $d_{i,0}$ est le biais et $(d_{i,1},\dots,d_{i,k})$ sont les poids synaptiques du i-ème neurone caché. Les sorties de ces neurones cachés sont les entrées des neurones de sortie ; le neurone de sortie calcule

$$g(\mathbf{x}) = \text{heaviside}\left(c_0 + \sum_{i=1}^{\ell} c_i z_i\right) \,.$$

L'analyse théorique des réseaux neuronaux est basée sur un théorème classique de Kolmogorov (1957) et Lorentz (1976) qui montre que chaque fonction continue f sur $[0,1]^d$ peut s'écrire de la façon suivante :

$$f(\mathbf{x}) = \sum_{i=1}^{2d+1} F_i\left(\sum_{j=1}^{d} G_{i,j}(x^{(j)})\right) \,,$$

où F_i et $G_{i,j}$ sont des fonctions continues qui dépendent de f. Dans la pratique, les réseaux de neurones multi-couches permettent d'approcher toutes les fonctions mesurables avec une précision arbitraire, même si leurs fonctions d'activation sont fixées. ♢

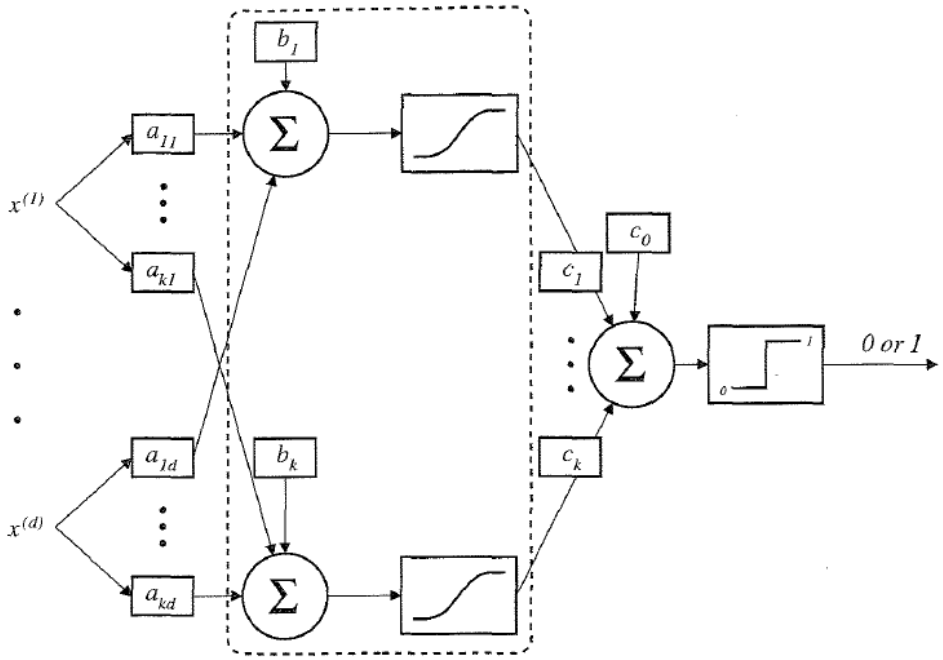

FIGURE III-1.2 – Un réseau neuronal avec une couche cachée.

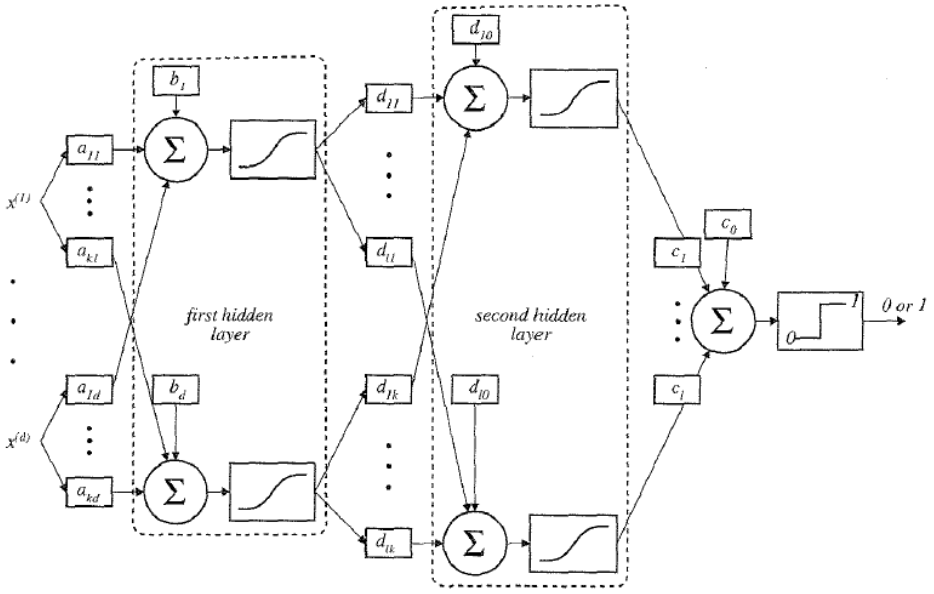

FIGURE III-1.3 – Un réseau neuronal avec deux couches cachées.

III-1.2 Classification bayésienne

La théorie bayésienne de la décision est une approche statistique fondamentale pour définir des règles de décision. Elle est basée sur l'hypothèse que le problème de décision est posé dans un cadre probabiliste, et que la loi des observations est **connue**.

Définition III-1.4 (Risque bayésien, règle bayésienne). *Le risque bayésien de classification est l'infimum du risque de classification (ou de la probabilité d'erreur) pour tous les classifieurs.*

$$R^* = \inf_{g \in \text{measurable}} R(g) \,. \tag{III-1.9}$$

Une règle de classification qui atteint le risque bayésien est appelée classifieur bayésien.

Nous allons montrer que toute règle de classification qui satisfait

$$g^*(x) = \begin{cases} 1 & \text{si} \quad \eta(x) > 1/2 \\ 0 & \text{autrement,} \end{cases}$$

où $\eta(X) = \mathbb{P}(Y = 1 \,|\, X)$ est la loi a posteriori définie par (III-1.1) est bayésienne. Avant de démontrer rigoureusement ce résultat, nous allons tout d'abord nous convaincre qu'une telle règle est en fait très intuitive. Si, pour un $X = x$ donné, nous savons que la probabilité que $Y = 1$ est plus grande que celle de $Y = 0$, nous prédisons que l'étiquette est 1, et vice-versa. Ainsi, le classifieur bayésien choisit l'étiquette $Y = 1$ si $\eta(X) = \mathbb{P}(Y = 1 \,|\, X) \geq \mathbb{P}(Y = 0 \,|\, X)$, et l'étiquette zéro sinon.

Exemple III-1.5. Considérons le problème élémentaire suivant : nous supposons que l'attribut X est scalaire, distribué suivant une loi uniforme sur $[-1,+1]$. Nous supposons d'autre part que la loi conditionnelle du label Y est donnée par

$$\eta(x) = \begin{cases} 0.9 & \text{si } x \geq 0 \\ 0.1 & \text{sinon} \end{cases} .$$

La loi jointe du couple (X,Y) a donc une densité par rapport à la mesure $\lambda_{\text{Leb}} \otimes \text{c}$ donnée par

$$\begin{aligned} h_{X,Y}(x,1) &= \frac{0.9}{2} \mathbb{1}_{[0,1]}(x) + \frac{0.1}{2} \mathbb{1}_{[-1]1}(x) \mathbb{1}_{[-1,0[}(x) \\ h_{X,Y}(x,0) &= \frac{0.1}{2} \mathbb{1}_{[0,1]}(x) + \frac{0.9}{2} \mathbb{1}_{[-1]1}(x) \mathbb{1}_{[-1,0[}(x) \,. \end{aligned}$$

Le classifieur bayésien est donc donné par

$$g^*(x) = \mathbb{1}_{[0,1]}(x) \,.$$

Le risque du classifieur bayésien est donné par

$$\begin{aligned} \mathbb{P}(g^*(X)) \neq Y) &= \mathbb{P}(g^*(X) = 1, Y = 0) + \mathbb{P}(g^*(X) = 0, Y = 1) \\ &= \mathbb{P}(X \geq 0, Y = 0) + \mathbb{P}(X < 0, Y = 1) \\ &= \int_0^1 h_{X,Y}(x,0) \lambda_{\text{Leb}}(\text{d}x) + \int_{-1}^0 h_{X,Y}(x,1) \lambda_{\text{Leb}}(\text{d}x) = \frac{0.1}{2} + \frac{0.1}{2} = 0.1 \,. \end{aligned}$$ ♢

Nous sommes maintenant en mesure d'énoncer notre premier résultat.

Théorème III-1.6. *Pour tout classifieur $g : \mathbb{R}^d \to \{0,1\}$,*

$$\mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y) \,.$$

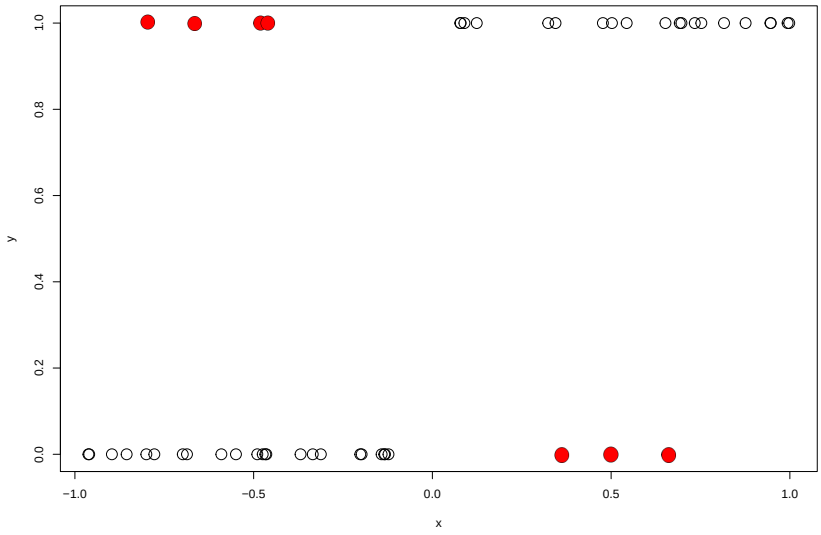

FIGURE III-1.4 – 40 échantillons engendrés par le modèle de l'Exemple III-1.5. En rouge, les points mal classés par le classifieur bayésien.

Démonstration. La probabilité d'erreur conditionnelle de toute règle de classification g peut être exprimée comme suit

$$\begin{aligned}\mathbb{P}(g(X)\neq Y\,|\,X) &= 1-\mathbb{P}(Y=g(X)\,|\,X)\\ &= 1-\{\mathbb{P}(Y=1,g(X)=1\,|\,X)+\mathbb{P}(Y=0,\ g(X)=0\,|\,X)\}\\ &= 1-\left\{\mathbb{1}_{\{g(X)=1\}}\eta(X)+\mathbb{1}_{\{g(X)=0\}}(1-\eta(X))\right\}\ ,\end{aligned}$$

où $\mathbb{1}_A$ est l'indicatrice de l'ensemble A et $\eta(X)=\mathbb{P}(Y=1\,|\,X)$ (voir (III-1.1)). Nous avons donc

$$\begin{aligned}&\mathbb{P}(g(X)\neq Y\,|\,X)-\mathbb{P}(g^*(X)\neq Y\,|\,X)\\ &=\eta(X)(\mathbb{1}_{\{g^*(X)=1\}}-\mathbb{1}_{\{g(X)=1\}})+(1-\eta(X))(\mathbb{1}_{\{g^*(X)=0\}}-\mathbb{1}_{\{g(X)=0\}})\\ &=\ (2\eta(X)-1)(\mathbb{1}_{\{g^*(X)=1\}}-\mathbb{1}_{\{g(X)=1\}})\geq\ 0\,,\end{aligned}\qquad\text{(III-1.10)}$$

en utilisant que $\mathbb{1}_{\{g=0\}}=1-\mathbb{1}_{\{g=1\}}$ dans la dernière égalité, et la définition de la règle de classification bayésienne g^* pour l'inégalité. Le résultat en découle en intégrant les deux parties par rapport à la loi μ des attributs X. □

La preuve ci-dessus montre que

$$R(g)=1-\mathbb{E}\left[\mathbb{1}_{\{g(X)=1\}}\eta(X)+\mathbb{1}_{\{g(X)=0\}}(1-\eta(X))\right],\qquad\text{(III-1.11)}$$

et en particulier

$$R^*=1-\mathbb{E}\left[\mathbb{1}_{\{\eta(X)>1/2\}}\eta(X)+\mathbb{1}_{\{\eta(X)\leq 1/2\}}(1-\eta(X))\right]=\mathbb{E}[\min(\eta(X),1-\eta(X))]\,.$$

Notez que g^* dépend de la distribution de (X,Y). Si cette distribution est connue, g^* peut être calculé. Le plus souvent, la distribution de (X,Y) et donc le classifieur bayésien g^* sont inconnus.

Remarque III-1.7. Observons que la probabilité a posteriori

$$\eta(x)=\mathbb{P}(Y=1\,|\,X=x)=\mathbb{E}\,[Y\,|\,X=x]$$

minimise l'erreur quadratique de prédiction de l'étiquette Y par $g(X)$ où $g : \mathbb{R}^d \to [0,1]$ est une fonction mesurable

$$\mathbb{E}[(\eta(X)-Y)^2] \leq \mathbb{E}[(g(X)-Y)^2]\,.$$

En effet,

$$\begin{aligned}\mathbb{E}\left[(g(X)-Y)^2\,\middle|\,X\right] &= \mathbb{E}\left[(g(X)-\eta(X)+\eta(X)-Y)^2\,\middle|\,X\right]\\ &= (g(X)-\eta(X))^2 + 2(g(X)-\eta(X))\,\mathbb{E}\left[\eta(X)-Y\,|X\right] + \mathbb{E}\left[(\eta(X)-Y)^2\,\middle|\,X\right]\\ &= (g(X)-\eta(X))^2 + \mathbb{E}\left[(\eta(X)-Y)^2\,\middle|\,X\right] \geq \mathbb{E}\left[(\eta(X)-Y)^2\,\middle|\,X\right]. \qquad \diamondsuit\end{aligned}$$

III-1.3 Risque moyen, excès de risque

Nous allons maintenant chercher à utiliser les données pour construire une "bonne" règle de prédiction. Notre objectif est de construire un classifieur dont le risque (probabilité d'erreur en classification binaire) est aussi petit que possible. Comme nous ne connaissons pas la distribution sous-jacente, il faut recourir à l'utilisation des données d'apprentissage pour estimer les probabilités d'erreur pour les classifieurs de $\mathscr{C}$.

Supposons que nous disposions d'un ensemble d'apprentissage $\{(X_i,Y_i)\}_{i=1}^n$, où n est le nombre d'exemples et que $\{(X_i,Y_i)\}_{i=1}^n$ est une suite de variables aléatoires i.i.d. de même loi que (X,Y). Comme nous ne connaissons pas la distribution jointe des attributs et des labels, il n'est pas possible de calculer le risque (probabilité d'erreur) d'une règle de classification. Il est par contre possible d'évaluer le nombre d'erreurs commises sur l'ensemble d'apprentissage.

Définition III-1.8 (Risque empirique). *Le risque empirique associé à l'ensemble d'apprentissage* $\{(X_i,Y_i)\}_{i=1}^n$ *est donné par*

$$\widehat{\mathrm{R}}_n(g) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{g(X_i)\neq Y_i\}}.$$

La loi des grands nombres montre que, pour toute règle de prédiction g, le risque empirique $\widehat{\mathrm{R}}_n(g)$ converge en probabilité vers le risque $L(g)$.

Exemple III-1.9 (Classifieurs linéaires). Nous suposons ici $\mathsf{Y} = \{-1,1\}$ et considérons l'ensemble des classifieurs linéaires sur $\mathbb{R}^d$,

$$\mathscr{C} = \left\{\mathbf{x}\mapsto g(\mathbf{x}) = \mathrm{signe}(\mathbf{w}^T\mathbf{x})\ :\ \mathbf{w}\in\mathbb{R}^d\right\},$$

où par convention $\mathrm{signe}(x) = 2\,\mathrm{heaviside}(x) - 1$. Dans ce cas, le classifieur qui minimise le risque empirique est obtenu en déterminant un vecteur de poids qui minimise le risque empirique

$$\mathbf{w}_n^* \in \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\mathrm{signe}(\mathbf{w}^T X_i)\neq Y_i\}}\,.$$

Ce problème n'admet pas de façon générale une unique solution. Le classifieur minimisant le risque empirique est alors donné par $g_n(\mathbf{x}) = \mathrm{heaviside}(\mathbf{w}_n^T\mathbf{x})$.

L'algorithme du perceptron travaille directement sur le vecteur des poids synaptiques **a** qui caractérise l'hyperplan de discrimination. L'algorithme du perceptron utilise un protocole d'apprentissage itératif : il prend les données d'apprentissage les unes après les autres, chacune étant choisie soit par un passage systématique dans l'ensemble d'apprentissage (version "non stochastique"), soit par un tirage au hasard dans celui-ci (version 'stochastique").

Supposons qu'il existe un hyperplan qui sépare parfaitement les classes : il existe un vecteur de poids $\mathbf{w}_*\in\mathbb{R}^d$ tel que

$$\sum_{i=1}^n \mathbb{1}_{\{\mathrm{signe}(\mathbf{w}_*^T\mathbf{X}_i)\neq Y_i\}} = 0.$$

On peut déterminer $\mathbf{w}_*$ en utilisant l'algorithme du perceptron.

La version stochatique de l'algorithme du perceptron procède de la façon suivante. Nous normalisons tout d'abord tous les vecteurs d'attributs de sorte de $\|\mathbf{X}_i\| = 1$, pour $i \in \{1,\dots,n\}$ et nous initialisons le vecteur de poids $\mathbf{w}_0 = 0$. Nous appelons $\mathbf{w}_k$ le vecteur de poids à la k-ème itération de l'algorithme. A chaque itération k, nous sélectionnons un élément de l'ensemble d'apprentissage $I_k \in \{1,\dots,n\}$ au hasard (uniformément dans l'ensemble d'apprentissage). Nous mettons à jour le vecteur de poids $\mathbf{w}_k$ en utilisant la règle suivante

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \gamma \text{signe}(Y_{I_k}) \mathbf{X}_{I_k} \mathbb{1}_{\{\text{signe}(\mathbf{w}_k^T \mathbf{X}_{I_k}) \neq Y_{I_k}\}} ,$$

où $\gamma \in]0,1]$ est un pas. Nous pouvons interpréter la récursion précédente de la façon suivante :

— Si l'observation $(\mathbf{X}_{I_k}, Y_{I_k})$ est bien classée, alors on ne modifie pas les poids synaptiques.

— Si l'observation $(\mathbf{X}_{I_k}, Y_{I_k})$ est incorrectement classé, alors nous modifions les poids en prenant compte le signe du label. Si le label est $+1$, on ajoute au vecteur des poids synaptiques le vecteur d'attributs $\mathbf{X}_{I_k}$ multiplié par un pas d'apprentissage γ. Si le label est -1, alors on soustrait au vecteur courant $\gamma \mathbf{X}_{I_k}$.

Si les données d'apprentissage ne peuvent pas être séparées par un hyperplan, on peut montrer que la complexité de la minimisation du risque empirique croît de manière exponentielle avec le nombre d'observations. ♢

III-1.4 Sur-apprentissage

Plus la classe $\mathscr{C}$ est grande, plus petite est la quantité $\inf_{g \in \mathscr{C}} R(g)$. Toutefois, choisir une classe $\mathscr{C}$ trop grande ne permet pas nécessairement de réduire le risque d'un minimiseur du risque empirique.

Prenons un exemple extrême : supposons que $\mathscr{C}$ est l'ensemble de toutes les fonctions mesurables. Alors, le classifieur bayésien (qui minimise le risque de classification) est donné par $g^*(\mathbf{x}) = \mathbb{1}_{\{\eta(\mathbf{x}) > 1/2\}}$ où $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \,|\, X = \mathbf{x})$. Considérons la règle de décision g_{bad} définie par

$$g_{\text{bad}}(\mathbf{x}) = \begin{cases} Y_i & \text{si } \mathbf{x} = X_i \text{ pour } i \in \{1,\dots,n\} , \\ 0 & \text{sinon} . \end{cases}$$

Le risque empirique de cette règle de décision est 0, elle minimise donc le risque empirique. Cependant, g_{bad} prédit l'étiquette zéro pour les valeurs des attributs qui n'ont pas été observées pendant l'apprentissage, ce qui a toutes les chances d'être très mauvais.

Exemple III-1.10 (Sur-apprentissage). Considérons les données et le classifieur représentés dans la Figure III-1.5. La dimension de l'espace des attributs est $d = 2$. Les données ont été générées par un mélange de deux distributions gaussiennes centrées dans les quadrants supérieur gauche et inférieur droit de l'espace objet, et chacune de ces deux composantes correspond à une étiquette différente :

$$p(\mathbf{x}|y = i) = \frac{1}{2\pi|\Gamma|^{1/2}} \exp\left(-1/2(\mathbf{x} - \mu_i)^T \Gamma^{-1} (\mathbf{x} - \mu_i)\right) .$$

La matrice de covariance "intra-classe" Γ est supposée être la même pour les deux classes. Si nous avions connaissance de la distribution des observations, alors la règle de décision optimale (le classifieur bayésien) serait le classifieur linéaire, voir la Section III-1.1. Cependant, nous pouvons construire une infinité de classifieurs qui classent sans erreur les données d'apprentissage. La théorie nous enseigne que ces règles de classification sont sous-optimales par rapport au classifieur linéaire (voir le Théorème III-1.6). ♢

La performance de g_n est mesurée par la probabilité d'erreur conditionnelle (ou risque de classification conditionnel), qui est le risque de classification (voir Définition III-1.1) du classifieur g_n :

$$R_n = R(g_n) = \mathbb{P}(g_n(X; D_n) \neq Y \,|\, D_n) , \quad D_n = \{(X_i, Y_i)\}_{i=1}^n .$$

La probabilité d'erreur conditionnelle R_n (ou risque de classification) du classifieur g_n est une variable aléatoire qui dépend des données d'apprentissage D_n. Le conditionnement dans la probabilité d'erreur par D_n permet de distinguer l'aléa qui provient de l'échantillon d'apprentissage de celui d'un couple générique (X,Y), indépendant de l'ensemble d'apprentissage. Cette quantité est d'un intérêt théorique, car la loi du couple (X,Y) est inconnue. L'espérance de cette quantité, appelée *risque moyen*, est donnée par

$$\mathbb{E}[R(g_n)] = \mathbb{P}(g_n(X; D_n) \neq Y) .$$

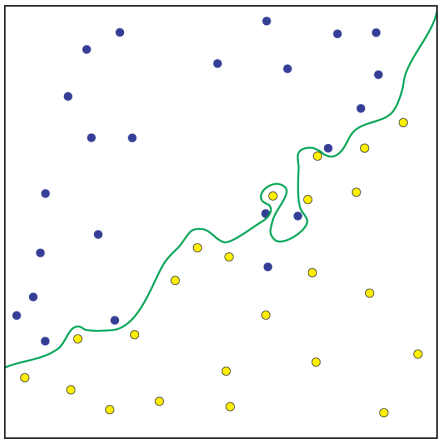

FIGURE III-1.5 – Exemple de sur-apprentissage. La frontière de décision du classifieur s'adapte afin d'étiqueter correctement toutes les données d'apprentissage, alors que le classifieur Bayésien optimal est un hyperplan.

Pour mesurer la performance de la règle de classification g_n, il est intéressant de considérer la différence entre le risque moyen $\mathbb{E}[R(g_n)]$ et le risque bayésien R^*. Cette différence est appelée l'*excès de risque* :

$$\mathbb{E}[R(g_n)] - R^* \,. \tag{III-1.12}$$

L'excès de risque peut se décomposer en deux termes :

— le *risque d'estimation*

$$\mathbb{E}[R(g_n)] - \inf_{g\in\mathscr{C}} R(g) \,, \tag{III-1.13}$$

— le *risque d'approximation*

$$\inf_{g\in\mathscr{C}} R(g) - R^* \,. \tag{III-1.14}$$

Le risque d'approximation quantifie la perte associée au choix de la classe $\mathscr{C}$. Il ne dépend pas des données d'apprentissage, mais seulement du choix de la famille de classifieurs $\mathscr{C}$ et de la distribution des observations (X,Y).

Le risque d'estimation reflète comment le classifieur g_n (construit à partir des données d'apprentissage) approche le classifieur optimal dans la classe $\mathscr{C}$. Il dépend de la classe $\mathscr{C}$ (plus cette classe est "grande", plus l'erreur sera importante) et de la "méthode" de construction du classifieur g_n (l'"algorithme" utilisé pour construire le classifieur à partir des données d'apprentissage).

Il est intéressant de comprendre ce qui se passe dans des cas extrêmes. Si la classe $\mathscr{C}$ est petite (le cas extrême étant que $\mathscr{C}$ contient uniquement une règle de classification), alors l'erreur d'estimation est faible (dans le cas extrême où $\mathscr{C}$ est un singleton, cette erreur est nulle), mais l'erreur d'approximation sera très importante. Si la classe $\mathscr{C}$ est grande (le cas extrême étant que $\mathscr{C}$ contient l'ensemble de toutes les règles de classification), alors l'erreur d'approximation est nulle mais l'erreur d'estimation sera importante.

Exemple III-1.11. Considérons le problème élémentaire suivant : nous supposons que l'attribut X est scalaire, distribué suivant une loi uniforme sur $[-1,+1]$. Nous supposons d'autre part que la loi conditionnelle du label Y est donnée par

$$\eta(x) = \begin{cases} 0.9 & \text{si } x \in [-1,1]\setminus]-1/2,1/2[\\ 0.1 & \text{sinon} \end{cases} .$$

Le classifieur bayésien est donné par

$$g^*(x) = \mathbb{1}_{[-1/2,1/2]}(x) \,.$$

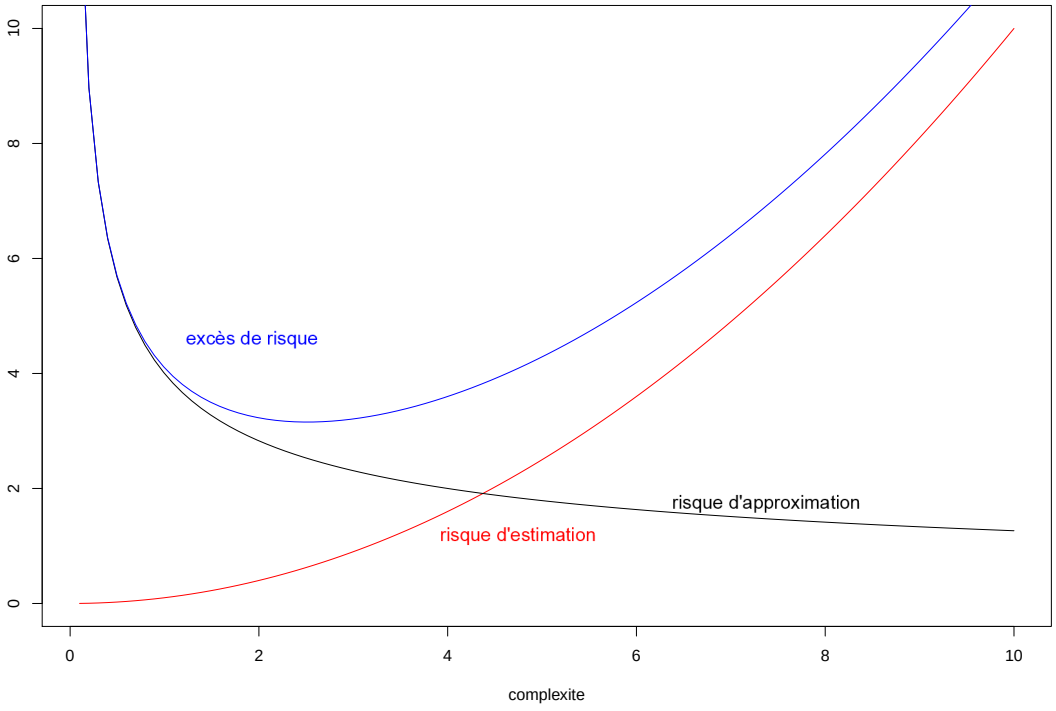

FIGURE III-1.6 – Illustration du compromis entre le risque d'approximation et le risque d'estimation

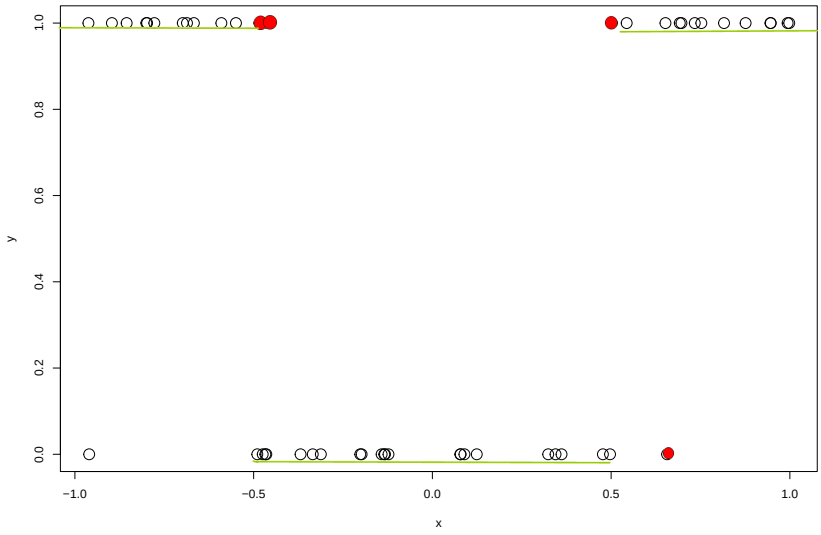

FIGURE III-1.7 – 50 échantillons engendrés par le modèle de l'Exemple III-1.5. En rouge les points mal classés par le classifieur bayésien

.

Le risque du classifieur bayésien est donné par

$$\begin{aligned}\mathbb{P}(g^*(X)) \neq Y) &= \mathbb{P}(X \notin [-1/2,1/2], Y=0) + \mathbb{P}(X \in [-1/2,1/2], Y=1) \\ &= \frac{0.1}{2} + \frac{0.1}{2} = 0.1\,.\end{aligned}$$

Nous considérons tout d'abord la classe d'estimateurs

$$\mathscr{C}_R = \left\{ g(x) = \text{heaviside}(b+ax) \,:\, (a,b) \in \mathbb{R}^2 \right\}\,.$$

Nous déterminons les poids a, b en minimisant le risque empirique

$$\widehat{\mathrm{R}}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g(X_i)\neq Y_i\}}\,.$$

Nous considérons maintenant la classe d'estimateurs quadratiques

$$\mathscr{C}_Q = \left\{ g(x) = \text{heaviside}(b+a_1x+a_2x^2) \,:\, (a_1,a_2,b) \in \mathbb{R}^3 \right\}\,.$$ ♢

Nous pouvons de la même façon déterminer les poids en minimisant le risque empirique.

Il faut donc définir une stratégie permettant d'éviter :

— le surapprentissage : le classifieur apprend par coeur les observations, ce qui l'empêche de généraliser correctement.

— le sousapprentissage : la classe $\mathscr{C}$ est trop petite pour contenir une bonne approximation du classifieur de Bayes.

Deux approches sont utilisées pour réaliser ce compromis :

— Restreindre la classe $\mathscr{C}$ pour contrôler l'erreur d'estimation. Nous pouvons par exemple nous limiter à des discriminateurs linéaires, ou considérer des architectures de réseau neuronaux "simples", de façon à contrôler l'erreur d'estimation.

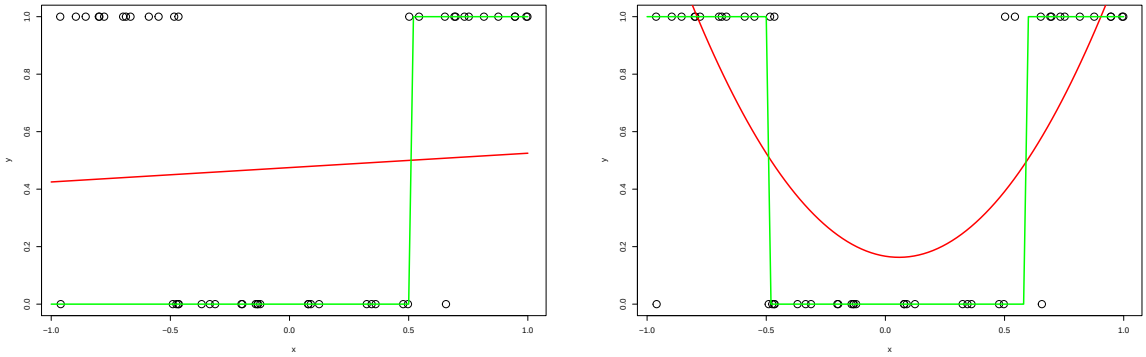

FIGURE III-1.8 – Panneau de gauche : classifieur linéaire ; Panneau de droite : classifieur quadratique

— Ajouter au risque empirique une “pénalité" - dépendant éventuellement des données d'apprentissage - mesurant la “complexité" du modèle (dans un sens que nous définirons dans un chapitre ultérieur) permettant d'éviter les classifieurs trop complexes.

Il existe de nombreuses façon de restreindre les classes de règles de classification ou de pénaliser le risque empirique. Nous en étudierons quelques unes dans le Chapitre III-3. Nous allons maintenant les introduire de façon informelle.

La méthode du crible (ou *method of sieves*) : L'approche la plus simple consiste à ajuster la complexité de la classe de modèles au nombre d'exemples d'apprentissage. Plus le nombre d'exemples d'apprentissage est grand, plus nous pouvons considérer des modèles “complexes". Dans la méthode du crible, nous considérons une collection de modèles $\{\mathscr{C}_k\}_{k=1}^K$, dont la “complexité" croît avec l'indice k : par exemple $\mathscr{C}_k$ peut être l'ensemble des classifieurs de la forme $g(x) = \mathbb{1}_{\{a(x)\geq 0\}}$ où $a(x)$ est un polynôme multivarié de degré maximum k ($\mathscr{C}_1$ est la classe des discriminateurs linéaires, $\mathscr{C}_2$ la classe des discriminateurs quadratiques, etc..). La méthode du crible procède en deux étapes.

Le première étape consiste à calculer pour chaque classe $\mathscr{C}_k$ une règle de classification $\hat{g}_k$ (par exemple, la règle minimisant le risque empirique sur cette classe $\hat{g}_k^* = \arg\min_{g\in\mathscr{C}_k} \widehat{\mathrm{R}}_n(g)$). Nous disposons ainsi pour chaque classe $\mathscr{C}_k$ d'une règle de classification $\hat{g}_k$.

La deuxième étape consiste à déterminer, dans cet ensemble de règles de classification, une règle de classification $\hat{g}_n^{**} = \hat{g}_{n,K_n}$, où K_n est un index aléatoire, dépendant de l'ensemble d'apprentissage. Idéalement, l'indice K_n devrait être choisi de façon à ce que l'excès de risque soit minimal, i.e. on voudrait $K_n \in \arg\min_{k\in\mathbb{N}^*} \{\mathbb{E}[R(\hat{g}_k)] - R^*\}$. Comme les quantités $\mathbb{E}[R(\hat{g}_k)] - R^*$ ne sont pas calculables, on va chercher K_n de façon à garantir que l'excès de risque $\mathbb{E}[R(\hat{g}_n^{**})] - R^*$ soit proche de $\min_{k\in\mathbb{N}^*} \{\mathbb{E}[R(\hat{g}_k)] - R^*\}$.

Méthode de validation simple ("hold-out" method) L'idée de base de la méthode de validation simple est de diviser les données disponibles $D_n = \{(X_i,Y_i)\}_{i=1}^n$ en un ensemble d'apprentissage $D_T = \{(X_i,Y_i)\}_{i=1}^m$ et un ensemble de validation $D_V = \{(X_i,Y_i)\}_{i=m+1}^n$. Supposons que nous disposions d'une collection de modèles $\{\mathscr{C}_k : k\in\mathbb{N}\}$. Pour chaque classe, nous calculons un classifieur $\hat{g}_k$ en utilisant uniquement les données d'apprentissage $\{(X_i,Y_i)\}_{i=1}^m$ (on peut par exemple considéré $\hat{g}_k^*$ minimisant l'erreur empirique sur l'ensemble d'apprentissage $\hat{g}_k^* = \arg\min_{g\in\mathscr{C}_k} \widehat{\mathrm{R}}_m(g)$, où $\widehat{\mathrm{R}}_m(g) = m^{-1}\sum_{i=1}^m \mathbb{1}_{\{g(X_i)\neq Y_i\}}$. Notons que ces classifieurs ont été obtenus en utilisant uniquement les données d'apprentissage. Nous pouvons maintenant déterminer la performance de ces règles de classification sur les données de validation. Nous définissons l'erreur “hold-out"

$$\hat{R}_V(g) = \frac{1}{n-m}\sum_{i=m+1}^{n} \mathbb{1}_{\{g(X_i)\neq Y_i\}} \,.$$

Nous sélectionnons enfin un classifieur minimisisant l'erreur “hold-out"

$$g_n^{**} \in \arg\min_k \hat{R}_V(\hat{g}_k^*) \,.$$

Nous utilisons l'ensemble d'apprentissage pour entraîner une famille de règles de classification et l'ensemble de validation pour choisir un classifieur dans cette famille.

III-1.5 Consistance

Comme en statistique paramétrique, il est possible de définir une notion de *consistance* d'une suite de règles de classification. Intuitivement, une suite de classifieurs est "bonne" quand lson risque de classification converge, quelle que soit la loi du couple (X,Y), vers le risque bayésien, lorsque le nombre d'exemples d'apprentissage tend vers l'infini.

Définition III-1.12 (Consistance d'une suite de classifieurs). *Soit $\{(X_k,Y_k),\ k\in\mathbb{N}^*\}$ une suite de variables aléatoires i.i.d. de même loi $\mu_{X,Y}$. Soit $\{\hat{g}_n,\ n\in\mathbb{N}\}$ une suite de classifieurs, où pour tout $n\in\mathbb{N}^*$ et $x\in\mathbb{R}^d$, $\hat{g}_n(x)=\hat{g}_n(x;D_n)$ où $D_n=\{(X_k,Y_k)\}_{k=1}^n$. Nous dirons que la suite $\{\hat{g}_n,\ n\in\mathbb{N}\}$ est faiblement consistante si*

$$R(\hat{g}_n)\xrightarrow{\mathbb{P}-\text{prob}} R^*\ . \tag{III-1.15}$$

où R^* est le risque bayésien (voir Définition III-1.4-Eq.(III-1.9)*).*

Remarquons que la condition (III-1.15) équivaut à

$$\lim_{n\to\infty}\mathbb{E}[R(\hat{g}_n)]=R^*\ , \tag{III-1.16}$$

En effet, pour tout $\delta>0$, en appliquant l'inégalité de Markov,

$$\mathbb{P}\left(R(\hat{g}_n)-R^*\geq\delta\right)\leq\delta^{-1}\{\mathbb{E}[R(\hat{g}_n)]-R^*\}\to_{n\to\infty}0\ .$$

D'autre part, puisque $R^*\leq R(\hat{g}_n)\leq 1$,

$$\begin{aligned}\mathbb{E}[R(\hat{g}_n)]-R^* &= \mathbb{E}[\{R(\hat{g}_n)-R^*\}\mathbb{1}_{\{R(\hat{g}_n)-R^*\geq\delta\}}]+\mathbb{E}[\{R(\hat{g}_n)-R^*\}\mathbb{1}_{\{R(\hat{g}_n)-R^*\leq\delta\}}]\\ &\leq \mathbb{P}(R(\hat{g}_n)-R^*\geq\delta)+\delta\end{aligned}$$

et donc la condition (III-1.15) implique (III-1.16).

Chapitre III-2

Apprentissage PAC

Développer des algorithmes qui permettent d'apprendre à partir d'exemples soulève des questions fondamentales. Qu'est-il possible d'apprendre ? Peut-on quantifier la difficulté d'apprentissage ? De combien d'exemples devons nous disposer pour apprendre de façon fiable ? Existe-t-il une façon universelle d'apprendre ? Dans ce chapitre introductif à la théorie de l'apprentissage, nous allons chercher (très modestement nous le verrons) à apporter des réponses à ces questions difficiles. Nous allons montrer dans des cas simples comment l'erreur d'approximation est liée à la complexité de la classe de classifieurs. Nous nous concentrerons sur l'exemple simple des classes $\mathscr{C}$ finies : la complexité est ici donnée par le "nombre" d'éléments de cette classe. Nous étendrons dans le chapitre suivant ces résultats à des classes infinies.

Nous utiliserons dans ce chapitre le cadre théorique introduit par [**?**]. Nous disposons d'un ensemble d'apprentissage $D_n = \{(X_i, Y_i)\}_{i=1}^n$, qui sont des variables aléatoires i.i.d. de loi inconnue.

Définition III-2.1 (bornes PAC ; [?]). *Soit $\varepsilon > 0$, $\delta \in]0,1[$ et $\mathscr{C}$ une famille de règles de classification. Nous dirons qu'une règle $g_n \in \mathscr{C}$ est Probablement Approximativement Correcte avec une précision ε et une confiance $1-\delta$ (où plus simplement, g_n est (ε,δ)-PAC) si*

$$\mathbb{P}\left(R(g_n) - \inf_{g\in\mathscr{C}} R(g) \geq \varepsilon\right) \leq \delta\,, \tag{III-2.1}$$

où, pour $g \in \mathscr{C}$, $R(g) = \mathbb{P}(g(X) \neq Y)$ est le risque de classification.

De façon équivalente, la règle de classification g_n est (ε,δ)-PAC si

$$R(g_n) - \inf_{g\in\mathscr{C}} R(g) \leq \varepsilon \quad \text{avec une probabilité } 1-\delta. \tag{III-2.2}$$

Ici $\varepsilon > 0$ est la précision : nous cherchons à construire une règle dont l'erreur d'approximation soit inférieure ε. Nous demandons que cette précision soit atteinte non pas uniformément sur les exemples d'apprentissage, mais avec une probabilité $1-\delta$. L'introduction de la confiance $1-\delta$ peut sembler surprenante. Il semblerait judicieux d'exiger que la règle de classification atteigne la précision cible ε quelque soit l'ensemble d'apprentissage, i.e.pouvoir prendre $\delta = 0$. Nous verrons dans la suite que cette exigence est excessive : il faut accepter que la règle de classification est une précision ε pour tous les exemples d'apprentissage à l'exception d'un ensemble d'exemples, dont la probabilité est contrôlée par $\delta > 0$. Il est aussi important de remarquer que la garantie PAC ((III-2.1) ou (III-2.2)) doit être satisfaite pour toutes les lois de probabilité.

III-2.1 Minimiseur du risque empirique

Nous allons maintenant obtenir des bornes PAC pour le classifieur minimisant le risque empirique. Nous allons dans un premier temps établir une inégalité générale, reliant le risque d'approximation et la déviation uniforme du risque. Soit $\mathscr{C}$ une classe de règles de classification. Notons par $\hat{g}_n^*$ une règle de classification de $\mathscr{C}$ qui minimise le risque empirique dans $\mathscr{C}$. Le résultat suivant montre que l'erreur conditionnelle de $\hat{g}_n$ est contrôlée par la déviation uniforme $\sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|$ sur la classe $\mathscr{C}$.

Théorème III-2.2. *Soit $\hat{g}_n^* \in \mathscr{C}$ un minimiseur du risque empirique sur $\mathscr{C}$:*

$$\widehat{\mathrm{R}}_n(\hat{g}_n^*) \leq \widehat{\mathrm{R}}_n(g) \quad \textit{pour tout } g \in \mathscr{C}.$$

Alors, l'excès de risque ainsi que la différence entre le risque empirique et le risque de $\hat{g}_n^*$ satisfont

$$R(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) \leq 2 \sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| \,, \tag{III-2.3}$$

$$|\widehat{\mathrm{R}}_n(\hat{g}_n^*) - R(\hat{g}_n^*)| \leq \sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| \,. \tag{III-2.4}$$

Démonstration. Pour tout $\varepsilon > 0$, il existe $g_\varepsilon \in \mathscr{C}$ tel que $R(g_\varepsilon) - \varepsilon \leq \inf_{g\in\mathscr{C}} R(g)$. Par conséquent,

$$\begin{aligned} R(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) = \; & R(\hat{g}_n^*) - \widehat{\mathrm{R}}_n(\hat{g}_n^*) + \widehat{\mathrm{R}}_n(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) \\ & \leq R(\hat{g}_n^*) - \widehat{\mathrm{R}}_n(\hat{g}_n^*) + \widehat{\mathrm{R}}_n(g_\varepsilon) - R(g_\varepsilon) + \varepsilon \,, \end{aligned}$$

où on a également utilisé que, par définition $\widehat{\mathrm{R}}_n(\hat{g}_n^*) \leq \widehat{\mathrm{R}}_n(g_\varepsilon)$. On en déduit finalement

$$R(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) \leq \; 2 \sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| + \varepsilon \,.$$

L'inégalité (III-2.3) en découle immédiatement car $\varepsilon > 0$ est arbitraire. L'inégalité (III-2.4) est trivialement satisfaite. □

Nous voyons que $\sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|$ fournit des bornes supérieures pour deux quantités d'intérêt simultanément :

1. Une borne supérieure pour l'excès de risque de $\hat{g}_n^*$:

$$R(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) \,.$$

2. Une borne supérieure pour l'erreur $|\widehat{\mathrm{R}}_n(\hat{g}_n^*) - R(\hat{g}_n^*)|$ commise en estimant le risque $R(\hat{g}_n^*)$ de $\hat{g}_n^*$ par son risque empirique $\widehat{\mathrm{R}}_n(\hat{g}_n^*)$.

Ainsi, l'estimateur du minimum du risque empirique $\widehat{\mathrm{R}}_n(\hat{g}_n^*)$ est biaisé de façon optimiste, mais son biais est borné par $\sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|$. Il est donc utile de disposer de bornes pour cette quantité.

Pour chaque $g \in \mathscr{C}$, la variable aléatoire $n\widehat{\mathrm{R}}_n(g)$ est distribuée suivant une loi binomiale de paramètres n et $R(g)$. Ainsi, pour obtenir des bornes pour le risque conditionnel de la règle de classification correspondant à la règle de décision minimisant le risque empirique, nous devons étudier les déviations uniformes de variables aléatoires binomiales par rapport à leurs moyennes.

III-2.2 Une borne PAC élémentaire

Nous établissons d'abord une borne PAC élémentaire dans un cadre très particulier. Cette borne est obtenue sous des hypothèses restrictives, que nous relâcherons dans la suite.

On suppose que la classe $\mathscr{C}$ est de cardinal fini et que $\min_{g\in\mathscr{C}} R(g) = 0$. Ceci implique en particulier que l'estimateur bayésien est un élément de la classe $\mathscr{C}$.

Théorème III-2.3. *Supposons que le cardinal de la classe $\mathscr{C}$, $|\mathscr{C}|$, est fini et que $\min_{g\in\mathscr{C}} R(g) = 0$. Notons $\hat{g}_n^* \in \arg\min_{g\in\mathscr{C}} \widehat{\mathrm{R}}_n(g)$ un minimiseur du risque empirique. Alors pour tout $n \in \mathbb{N}$ et $\varepsilon > 0$,*

$$\mathbb{P}\left(R(\hat{g}_n^*) \geq \varepsilon\right) \leq |\mathscr{C}|\,\mathrm{e}^{-n\varepsilon} \,. \tag{III-2.5}$$

Démonstration. Soit $g^* \in \arg\min_{g\in\mathscr{C}} R(g)$ le classifieur bayésien. Par hypothèse, $R(g^*) = 0$. Comme $\mathbb{E}[\widehat{\mathrm{R}}_n(g^*)] = 0$ et $\widehat{\mathrm{R}}_n(g^*) \geq 0$, nous avons $\mathbb{P}(\widehat{\mathrm{R}}_n(g^*) = 0) = 1$. Comme $\widehat{\mathrm{R}}_n(\hat{g}_n^*) \leq \widehat{\mathrm{R}}_n(g^*)$, nous avons donc aussi $\mathbb{P}(\widehat{\mathrm{R}}_n(\hat{g}_n^*) = 0) = 1$. Ainsi, $R(\hat{g}_n^*) \geq \varepsilon$ seulement s'il existe $g \in \mathscr{C}$ tel que $R(g) \geq \varepsilon$ et $\widehat{\mathrm{R}}_n(g) = 0$. Autrement dit

$$\mathbb{P}\left(R(\hat{g}_n^*) \geq \varepsilon\right) \leq \mathbb{P}\left(\bigcup_{g\in\mathscr{C}, R(g)\geq\varepsilon} \{\widehat{\mathrm{R}}_n(g) = 0\}\right) \,.$$

Par la borne d'union, on a donc

$$\mathbb{P}\left(R(\hat{g}_n^*) \geq \varepsilon\right) \leq \sum_{g\in\mathscr{C}, R(g)\geq\varepsilon} \mathbb{P}(\widehat{\mathrm{R}}_n(g) = 0) \,. \tag{III-2.6}$$

La variable aléatoire $n\widehat{\mathrm{R}}_n(g)$ est une variable binomiale de probabilité de succès $R(g)$. Nous avons donc $\mathbb{P}(\widehat{\mathrm{R}}_n(g) = 0) = (1 - R(g))^n$ et (III-2.6) implique donc que

$$\mathbb{P}\left(R(\hat{g}_n^*) \geq \varepsilon\right) \leq \sum_{g\in\mathscr{C}, R(g)\geq\varepsilon} (1-\varepsilon)^n \leq |\mathscr{C}|\,(1-\varepsilon)^n \,.$$

En remarquant que $(1-\varepsilon)^n \leq \mathrm{e}^{-n\varepsilon}$, on en déduit (III-2.5). □

Corollaire III-2.4. *Supposons que le cardinal de la classe $\mathscr{C}$, $|\mathscr{C}|$, est fini et que $\min_{g\in\mathscr{C}} R(g) = 0$. Alors, pour tout $\varepsilon > 0$ et $\delta \in]0,1[$, le minimiseur du risque empirique $\hat{g}_n^*$ est (ε,δ)-PAC pour tout $n \geq n(\varepsilon,\delta)$ où*

$$n(\varepsilon,\delta) = \frac{\log|\mathscr{C}| + \log(1/\delta)}{\varepsilon} \,.$$

Nous pouvons utiliser la borne (III-2.5) pour obtenir une borne du risque moyen. Pour cela, on utilise le lemme élémentaire suivant.

Lemme III-2.5. *Soit Z une variable aléatoire à valeurs dans $\mathbb{R}_+$ et $C, \alpha > 0$ des constantes telles que*

$$\forall \varepsilon > 0, \qquad \mathbb{P}(Z \geq \varepsilon) \leq C\mathrm{e}^{-\alpha\varepsilon} \,. \tag{III-2.7}$$

Alors,

$$\mathbb{E}[Z] \leq \frac{1 + \log(C)}{\alpha} \,.$$

Démonstration. Comme Z est à valeurs dans $\mathbb{R}_+$,

$$\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z > t)\mathrm{d}t \,.$$

Pour tout $u > 0$, en utilisant $\mathbb{P}(Z > t) \leq 1$, on a alors

$$\mathbb{E}[Z] \leq u + C\int_u^\infty \mathrm{e}^{-\alpha t}\mathrm{d}t \leq u + \frac{C\mathrm{e}^{-\alpha u}}{\alpha} \,.$$

En appliquant cette borne avec $u = (\log C)/\alpha$, nous obtenons le résultat désiré. □

Corollaire III-2.6. *Supposons que le cardinal de la classe $\mathscr{C}$, $|\mathscr{C}|$, est fini et que* $\min_{g\in\mathscr{C}} R(g) = 0$. *Alors*

$$\mathbb{E}[R(\hat{g}_n^*)] \leq \frac{1+\log(|\mathscr{C}|)}{n} .$$

III-2.3 Une borne PAC agnostique

La borne PAC obtenue dans la section précédente n'est valide que si le classificateur bayésien (qui dépend de la loi de (X,Y)) est un élément de la classe $\mathscr{C}$. Nous allons maintenant chercher des bornes PAC dites *agnostiques* : qui n'utilisent pas d'hypothèses sur la loi de l'observation. Le point de vue "agnostique" diffère ainsi du point de vue de la modélisation statistique classique, qui part elle de la connaissance a priori d'une classe de modèles (on ne connaît pas la loi de l'observation mais on formule l'hypothèse qu'elle appartient à une famille de lois de probabilités, qui traduit notre connaissance a priori du domaine).

Rappelons que, pour tout $g \in \mathscr{C}$, $n\widehat{\mathrm{R}}_n(g)$ suit la loi binomiale de paramètre de succès $R(g)$: on doit donc montrer une inégalité de déviation uniforme sur une classe de variables aléatoires binomiales. Le résultat clef est l'inégalité suivante, qui est une conséquence directe du théorème de Hoeffding (voir le Théorème IV-1.9 et le Corollaire IV-1.10 pour la preuve de ce résultat).

Si $Z_1,\ldots,Z_n$ sont des variables aléatoires de Bernoulli de paramètre $p \in [0,1]$ et si $\bar{Z}_n = n^{-1}\sum_{i=1}^n Z_i$, alors, pour tout $n \in \mathbb{N}$ et $t > 0$,

$$\mathbb{P}\left(\overline{Z}_n - p \geq t\right) \leq \exp(-2nt^2) .$$

$$\mathbb{P}\left(\overline{Z}_n - p \leq -t\right) \leq \exp(-2nt^2) .$$

Il est important de remarquer que le membre de droite *ne dépend pas* de la probabilité de succès $p \in [0,1]$. La *même* inégalité de déviation reste valable *quelque soit* la probabilité de succès p.

Théorème III-2.7. *Supposons que le cardinal de la classe $\mathscr{C}$, $|\mathscr{C}|$ soit fini. Alors nous avons pour tout $n \in \mathbb{N}$ et $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{g\in\mathscr{C}} \widehat{\mathrm{R}}_n(g) - R(g) > \varepsilon\right) \leq |\mathscr{C}|\,\mathrm{e}^{-2n\varepsilon^2} , \tag{III-2.8}$$

$$\mathbb{P}\left(\sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| > \varepsilon\right) \leq 2\,|\mathscr{C}|\,\mathrm{e}^{-2n\varepsilon^2} . \tag{III-2.9}$$

Démonstration. Notons $\mathscr{C} = \{g_i\}_{i=1}^{|\mathscr{C}|}$. La borne de l'union montre que

$$\mathbb{P}\left(\sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| > \varepsilon\right) \leq \sum_{i=1}^{|\mathscr{C}|} \mathbb{P}(|\widehat{\mathrm{R}}_n(g_i) - R(g_i)| > \varepsilon) .$$

Les inégalités (III-2.8) et (III-2.9) découlent du Corollaire IV-1.10 et du fait les variables $\{\mathbb{1}_{\{g(X_i)\neq Y_i\}}\}_{i=1}^n$ sont distribuées suivant une loi de Bernoulli. □

La différence entre (III-2.9) et l'inégalité de Hoeffding (Théorème IV-1.9) réside dans le terme $|\mathscr{C}|$ du membre de droite. Ce terme garantit que la borne soit valable simultanément pour toutes les règles de classification de $\mathscr{C}$.

Les bornes de déviation uniforme obtenues dans le Théorème III-2.7 ont de nombreuses conséquences. Elles montrent en particulier que, pour tout $\delta \in]0,1[$,

$$\mathbb{P}\left(\forall g \in \mathscr{C}, \quad R(g) \leq \widehat{\mathrm{R}}_n(g) + C(\mathscr{C}, n, \delta)\right) \geqslant 1-\delta \,, \tag{III-2.10}$$

où

$$C(\mathscr{C}, n, \delta) = \sqrt{\frac{\log|\mathscr{C}| + \log(1/\delta)}{2n}} \,.$$

Ainsi, le risque de chaque classifieur est contrôlé uniformément sur la classe $\mathscr{C}$ par le risque empirique du classificateur à un terme additif près, dépendant uniquement de la complexité de la classe (définie ici comme le logarithme de son cardinal). Comme (III-2.10) est vérifiée *uniformément*, elle est en particulier vérifiée par $\hat{g}_n^*$ et donc, avec une probabilité $1-\delta$,

$$R(\hat{g}_n^*) \leq \widehat{\mathrm{R}}_n(\hat{g}_n^*) + C(\mathscr{C}, n, \delta) \,.$$

La borne inférieure du risque empirique permet de contrôler le risque conditionnel de l'estimateur à un terme additif près. Par le Théorème III-2.2, pour tout $\varepsilon > 0$ et $n \in \mathbb{N}$, nous avons aussi

$$\mathbb{P}\left(R(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{g\in\mathscr{C}} \left|\widehat{\mathrm{R}}_n(g) - R(g)\right| \geq \varepsilon/2\right) \leq 2\,|\mathscr{C}|\,\mathrm{e}^{-n\varepsilon^2/2} \,. \tag{III-2.11}$$

On peut déduire de cette inégalité une borne PAC agnostique.

Corollaire III-2.8. *Supposons que le cardinal de la classe $\mathscr{C}$, $|\mathscr{C}|$ soit fini. Alors, pour tout $\varepsilon > 0$ et $\delta \in [0,1]$, l'estimateur $\hat{g}_n^*$ est (ε,δ)-PAC pour tout $n \geq n(\varepsilon,\delta)$ où*

$$n(\varepsilon,\delta) = \frac{2\{\log(1/\delta) + \log(2\,|\mathscr{C}|)\}}{\varepsilon^2} \,.$$

Notons que $\log(|\mathscr{C}|)$ est aussi le nombre d'éléments binaires nécessaires pour spécifier une fonction g particulière dans $\mathscr{C}$. On retrouve là un lien intéressant, et non fortuit, avec la théorie de l'information et du codage.

Notons aussi que, pour tout $\varepsilon > 0$ et $\delta \in]0,1[$,

$$\mathbb{P}\left(R(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) \leq \sqrt{\frac{\log(2\,|\mathscr{C}|) + \log(1/\delta)}{2n}}\right) \geq 1-\delta \,. \tag{III-2.12}$$

Il est aussi intéressant de déterminer une borne de

$$\mathbb{E}\left[\sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|\right].$$

Le Théorème III-2.7 peut être utilisé pour déduire une borne de cette quantité en utilisant le lemme élémentaire suivant.

Lemme III-2.9. *Soit Z une variable aléatoire à valeurs dans $\mathbb{R}_+$, $C, \alpha > 0$ des constantes telles que*

$$\forall \varepsilon > 0, \qquad \mathbb{P}(Z \geq \varepsilon) \leq C\mathrm{e}^{-\alpha\varepsilon^2} \,.$$

Alors

$$\mathbb{E}[Z] \leq \sqrt{\frac{1+\log(C)}{\alpha}} \,.$$

Démonstration. On a par hypothèse,

$$\forall \varepsilon > 0, \qquad \mathbb{P}(Z^2 \geq \varepsilon) \leq C\mathrm{e}^{-\alpha\varepsilon} .$$

Donc, d'après (III-2.7)

$$\mathbb{E}[Z^2] \leqslant \frac{1+\log(C)}{\alpha} .$$

On conclut par l'inégalité de Cauchy-Schwarz. □

Théorème III-2.10. *Supposons que le cardinal de l'ensemble des classifieurs $\mathscr{C}$ soit fini. Alors,*

$$\mathbb{E}\left[\sup_{g\in\mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|\right] \leq \sqrt{\frac{1+\log(2\,|\mathscr{C}|)}{2n}} , \tag{III-2.13}$$

$$\mathbb{E}[R(\hat{g}_n)] \leq \inf_{g\in\mathscr{C}} R(g) + 2\sqrt{\frac{1+\log(2\,|\mathscr{C}|)}{2n}} . \tag{III-2.14}$$

Il est possible d'améliorer les constantes numériques dans ce résultat en combinant le Lemme IV-1.8 et l'inégalité de Pisier (voir le Théorème IV-1.13)

III-2.4 Une application : classification par histogramme

Soit $m \in \mathbb{N}$ et $\{Q_1, \ldots, Q_m\}$ une partition finie de l'ensemble des attributs $\mathsf{X} \subset \mathbb{R}^d$. Par exemple, si $\mathsf{X} = [0,1]^d$, nous pouvons diviser chaque axe en q intervalles de longueur égales et considérer la famille des hypercubes

$$Q_{i_1,\ldots,i_d} = \prod_{j=1}^{d}](i_j - 1)/q, i_j/q] \ , (i_1, \ldots, i_d) \in \{1, \ldots, q\}^d .$$

On a alors $m = q^d$. Les histogrammes sur la partition $\{Q_1, \ldots, Q_m\}$ sont les classifieurs de la forme

$$\mathscr{C} = \left\{ g : \mathsf{X} \to \{0,1\} \ : \ g(x) = \sum_{i=1}^{m} d_i \mathbb{1}_{Q_i}(x), (d_1, \ldots, d_m) \in \{0,1\}^m \right\} .$$

Pour tout $i \in \{1, \ldots, m\}$, la décision de classification que nous prenons est la même pour tous les points $x \in Q_i$. Le nombre d'éléments de cette classe est 2^m. Le risque empirique de $g(x) = \sum_{i=1}^m d_i \mathbb{1}_{Q_i}(x)$ est donné par

$$\begin{aligned} \widehat{\mathrm{R}}_n(g) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g(X_i)\neq Y_i\}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{\{g(X_i)\neq Y_i, X_i \in Q_j\}} \\ &= \sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Q_j}(X_i)\left\{1 - \mathbb{1}_{\{d_j\}}(Y_i)\right\} . \end{aligned}$$

Ainsi, tout classifieur $g_n^* = \sum_{j=1}^m d_{n,j}^* \mathbb{1}_{Q_j}$, où

$$c_{n,j}^* = \begin{cases} 1 & \text{si } \sum_{i=1}^n Y_i \mathbb{1}_{\{Q_j\}}(X_i) > (1/2)\sum_{i=1}^n \mathbb{1}_{\{Q_j\}}(X_i) , \\ 0 & \text{sinon} , \end{cases}$$

minimise le risque empirique. Dans chaque élément de la partition, g_n^* prend la "décision majoritaire". En utilisant les résultats ci-dessus, nous pouvons calculer le risque d'approximation (voir (III-2.12)) et le

risque moyen (voir (III-2.14)) :

$$\mathbb{P}\left(R(\hat{g}_n^*) - \inf_{g\in\mathscr{C}} R(g) \leq \sqrt{\frac{(m+1)\log(2)+\log(1/\delta)}{2n}}\right) \geq 1-\delta\ ,$$

$$\mathbb{E}[R(\hat{g}_n)] - R^* \leq 2\sqrt{\frac{m\log(2)+\ln(2\mathrm{e}))}{2n}}\ .$$

Ainsi, l'erreur d'approximation et le risque moyen sont contrôlé par m/n : pour que l'erreur de généralisation soit faible il faut que le nombre d'éléments de la partition soit petit par rapport au nombre de données d'apprentissage (ce qui est intuitif, il faut disposer dans chaque élément de la partition d'un nombre suffisant d'exemples pour que le "vote majoritaire" ait réellement un sens).

III-2.5 Conclusion partielle

En particulier, (III-2.14) montre que, lorsque la classe $\mathscr{C}$ est de cardinal fini, la règle de classification minimisant le risque empirique est consistante (voir Définition III-1.12) : le risque moyen converge vers 0 à la vitesse $1/\sqrt{n}$.

On retiendra des résultats précédents trois idées importantes.

(i) D'abord, le cardinal de $\mathscr{C}$ (sa complexité en un certain sens), a un effet direct sur le contrôle de l'excès de risque. Cela confirme que le choix d'un ensemble de classifieurs $\mathscr{C}$ trop riche peut conduire à de mauvaises inductions.

(ii) Le contrôle de l'excès de risque se fait par une borne de la fluctuation du risque empirique autour de sa moyenne uniforme sur la classe $\mathscr{C}$. Nous verrons dans le chapitre suivant que la généralisation de ce raisonnement à des classes infinies de classifieurs fait aussi appel à un argument de convergence uniforme. Le contrôle du risque empirique est une analyse du "pire cas", pour lequel l'erreur d'approximation du risque par le risque empirique est maximale.

(iii) Ces bornes peuvent être améliorées si on abandonne le point de vue "agnostique" qui consiste à chercher à obtenir des bornes qui restent valables indépendamment de la loi des observations.

Chapitre III-3

Théorie de Vapnik-Chervonenkis

L'inégalité générale (III-2.3) montre l'intérêt d'un contrôle du suprémum de la différence entre le risque empirique et le risque sur la famille $\mathscr{C}$, valable indépendamment de la distribution des exemples (en supposant simplement que les exemples d'apprentissage sont indépendants). L'analyse PAC ("Probablement Approximativement Correct") a été menée au chapitre précédent dans le cas d'espaces de règles de classification de cardinal fini. Nous allons dans ce chapitre montrer qu'il est possible de généraliser cette étude à des familles de classificateurs de cardinal infini. Nous présentons de façon succincte l'approche due à Vladimir Vapnik, un des fondateurs de la théorie moderne de l'apprentissage ; voir [**?**]. Ces idées ont débouché non seulement sur une méthode d'évaluation de l'erreur $L(\hat{g}_n) - \inf_{g\in\mathscr{C}} R(g)$ (la dimension de Vapnik-Chervonenkis) mais aussi sur de nouveaux algorithmes généraux et performants (les séparateurs à vastes marges et leurs dérivés (SVM)).

III-3.1 Inégalité de McDiarmid

Dans ce paragraphe, nous donnons une extension de l'inégalité d'Hoeffding utilisée au chapitre précédent. Il s'agit une inégalité de concentration pour des fonctions de variables aléatoires indépendantes plus générales que la somme.

Soit X un ensemble borélien et $g : \mathsf{X}^n \to \mathbb{R}$ une fonction mesurable (de n variables). Nous allons maintenant obtenir des bornes de déviation de $g(X_1,\ldots,X_n)$ par rapport à sa moyenne lorsque les variables aléatoires $X_1,\ldots,X_n$ sont des variables aléatoires indépendantes prenant leurs valeurs dans l'ensemble A et g est à différences bornées, voir la Définition III-3.1. Par souci de concision, nous écrirons dans la suite, lorsqu'il n'y a pas d'ambiguité, pour tout $1 \le m \le n$, $X_{mn} = (X_m,\ldots,X_n)$.

Définition III-3.1 (Différences bornées). *La fonction mesurable $g : \mathsf{X}^n \to \mathbb{R}$ est à* différences bornées*, si pour tout $i \in \{1,\ldots,n\}$,*

$$\sup_{x_1,\ldots,x_n,x_i'\in\mathsf{X}^{n+1}} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \le c_i,$$

Pour de telles fonctions, il est possible de prouver l'inégalité exponentielle suivante, qui étend l'inégalité de Hoeffding pour les sommes.

Théorème III-3.2 (Inégalité de McDiarmid). *Soit $(X_1,\ldots,X_n)$ n variables aléatoires indépendantes à valeurs dans un ensemble $\mathsf{X} \in \mathscr{B}(\mathbb{R}^d)$. Soit $g : \mathsf{X}^n \to \mathbb{R}$ une fonction mesurable à différences bornées*

(voir Définition III-3.1). Alors pour tout $t > 0$,

$$\mathbb{P}\{g(X_{1n}) - \mathbb{E}[g(X_{1n})] \geq t\} \leq \exp\left(-2t^2 / \sum_{i=1}^{n} c_i^2\right) \tag{III-3.1}$$

et

$$\mathbb{P}\{g(X_{1n}) - \mathbb{E}[g(X_{1n})] \leq -t\} \leq \exp\left(-2t^2 / \sum_{i=1}^{n} c_i^2\right). \tag{III-3.2}$$

McDiarmid (1989) a prouvé cette inégalité à l'aide de techniques de martingales, que nous reproduisons ici. La preuve de Théorème III-3.2 utilise l'extension suivante de Lemme IV-1.8.

Lemme III-3.3. *Soit V et Z des variables aléatoires telles que $\mathbb{E}[V \,|\, Z] = 0$ presque-sûrement. Supposons de plus qu'il existe une fonction mesurable h et une constante $c \geq 0$ telle que.*

$$h(Z) \leq V \leq h(Z) + c.$$

Alors, pour tout $s > 0$,

$$\mathbb{E}\left[\mathrm{e}^{sV} \,\middle|\, Z\right] \leq \mathrm{e}^{s^2 c^2/8}.$$

Démonstration. (du Théorème III-3.2) Posons $G_n = g(X_{1n})$, $V = G_n - \mathbb{E}[G_n]$ et définissons

$$V_i = \mathbb{E}[G_n \,|\, X_{1i}] - \mathbb{E}[G_n \,|\, X_{1i-1}], \quad i = 1, \ldots, n;$$

avec la convention $\mathbb{E}[G_n \,|\, X_{10}] = \mathbb{E}[G_n]$. Par construction, $V = \sum_{i=1}^{n} V_i$. Introduisons

$$\forall i \in \{1, \ldots, n\}, \qquad H_i(X_{1i}) = \mathbb{E}[G_n \,|\, X_{1i}] = \int \cdots \int g(X_{1i}, x_{i+1n}) \prod_{j=i+1}^{n} \mu_j(\mathrm{d}x_j).$$

En notant μ_i la distribution de X_i et en utilisant l'indépendance de X_{1n}, nous obtenons

$$\forall i \in \{1, \ldots, n\}, \qquad V_i = H_i(X_{1i}) - \int H_i(X_{1i-1}, x_i)\mu_i(\mathrm{d}x_i).$$

Définissons les variables aléatoires

$$W_i = \sup_{u \in \mathsf{X}} \left(H_i(X_{1i-i}, u) - \int H_i(X_{1i-1}, x_i)\mu_i(\mathrm{d}x_i) \right),$$
$$Z_i = \inf_{u} \left(H_i(X_{1i-i}, u) - \int H_i(X_{1i-1}, x_i)\mu_i(\mathrm{d}x_i) \right).$$

Clairement, $Z_i \leq V_i \leq W_i$ presque-sûrement et, comme g est à différences bornées

$$\begin{aligned} W_i - Z_i &= \sup_{u \in \mathsf{X}} \sup_{v \in \mathsf{X}} \left(H_i(X_{1i-1}, u) - H_i(X_{1i-1}, v) \right) \\ &= \sup_{u \in \mathsf{X}} \sup_{v \in \mathsf{X}} \int \cdots \int \{ g(X_{1i-1}, u, x_{i+1n}) - g(X_{1i-1}, v, x_{i+1n}) \} \prod_{j=i+1}^{n} \mu_j(\mathrm{d}x_j) \leq c_i. \end{aligned}$$

D'après le Lemme III-3.3, pour tout $i \in \{1, \ldots, n\}$ et $s > 0$, nous avons

$$\mathbb{E}\left[\mathrm{e}^{sV_i} \,\middle|\, X_{1i-1}\right] \leq \mathrm{e}^{s^2 c_i^2/8}.$$

On utilise alors la propriété élémentaire de l'espérance conditionnelle : pour toutes variables aléatoires bornées, nous avons :

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY \,|\, Y]] = \mathbb{E}[Y\mathbb{E}[X \,|\, Y]].$$

On en déduit, pour tout $s > 0$,

$$\mathbb{E}\left[e^{s\Sigma_{i=1}^{n} V_i}\right] = \mathbb{E}\left[e^{s\Sigma_{i=1}^{n-1} V_i}\mathbb{E}\left[e^{sV_n} \,\middle|\, X_{1n-1}\right]\right] = e^{s^2 c_n^2/8}\,\mathbb{E}\left[e^{s\Sigma_{i=1}^{n-1} V_i}\right] .$$

Par une induction élémentaire, nous obtenons donc, pour tout $n \in \mathbb{N}$ et $s > 0$.

$$\mathbb{E}\left[e^{s\Sigma_{i=1}^{n} V_i}\right] \leq e^{s^2 \Sigma_{i=1}^{n} c_i^2/8} .$$

Finalement, en utilisant la borne de Chernoff (Corollaire IV-1.7), pour tout $s > 0$, nous avons

$$\mathbb{P}(G_n - \mathbb{E}[G_n] \geq t) \leq \mathbb{E}\left[e^{s\Sigma_{i=1}^{n} V_i}\right] e^{-st} \leq e^{-st} e^{s^2 \Sigma_{i=1}^{n} c_i^2/8} .$$

En prenant $s = 4t/\sum_{i=1}^{n} c_i^2$ nous obtenons (III-3.1). Nous obtenons (III-3.2) en appliquant (III-3.1) à la fonction $-g$. □

L'inégalité de McDiarmid a une conséquence immédiate mais essentielle.

Corollaire III-3.4. *Soit $\mathscr{C}$ une classe quelconque de règles de classification $g : \mathbb{R}^d \to \{0,1\}$. Alors*

$$\forall \varepsilon > 0, \qquad \mathbb{P}\left(\sup_{g\in\mathscr{C}} |\widehat{R}_n(g) - R(g)| - \mathbb{E}\left[\sup_{g\in\mathscr{C}} |\widehat{R}_n(g) - R(g)|\right] > \varepsilon\right) \leq e^{-2n\varepsilon^2} .$$

$$\forall \varepsilon > 0, \qquad \mathbb{P}\left(\left|\sup_{g\in\mathscr{C}} |\widehat{R}_n(g) - R(g)| - \mathbb{E}\left[\sup_{g\in\mathscr{C}} |\widehat{R}_n(g) - R(g)|\right]\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2} .$$

Démonstration. Considérons les fonctions

$$g((x_1,y_1),\ldots,(x_n,y_n)) = \sup_{g\in\mathscr{C}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g(x_i)\neq y_i\}} - R(g) ,$$

$$h((x_1,y_1),\ldots,(x_n,y_n)) = \sup_{g\in\mathscr{C}} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g(x_i)\neq y_i\}} - R(g)\right| .$$

g et h sont des fonctions à différences bornées sur $(\mathsf{X}\times\{0,1\})^n$ avec $c_i = 1/n$, pour tout $i \in \{1,\ldots,n\}$. Nous concluons en appliquant le Théorème III-3.2. □

La variable aléatoire $\sup_{g\in\mathscr{C}} |\widehat{R}_n(g) - R(g)|$ est concentrée autour de sa valeur moyenne, indépendamment de la distribution des données d'apprentissage. En particulier, pour tout $\delta \in]0,1[$, avec une probabilité supérieure $1-\delta$, nous avons

$$\sup_{g\in\mathscr{C}} |L_n(g) - R(g)| \leq \mathbb{E}\left[\sup_{g\in\mathscr{C}} |L_n(g) - R(g)|\right] + \sqrt{\frac{\log(1/\delta)}{2n}} . \tag{III-3.3}$$

Pour obtenir une borne de risque, il reste à borner l'espérance $\mathbb{E}[\sup_{g\in\mathscr{C}} |L_n(g) - R(g)|]$.

III-3.2 Inégalité de Vapnik-Chervonenkis

Rappelons que les Théorèmes III-2.7 and III-2.10 montrent que si le cardinal de la classe $\mathscr{C}$ des règles de classification est borné par N, alors pour tout $\varepsilon > 0$, nous avons

$$\mathbb{P}\left(\sup_{g\in\mathscr{C}} |\widehat{R}_n(g) - R(g)| > \varepsilon\right) \leq 2Ne^{-2n\varepsilon^2} ,$$

$$\mathbb{E}\left[\sup_{g\in\mathscr{C}} |\widehat{R}_n(g) - R(g)|\right] \leq \sqrt{\frac{\ln(2N)}{2n}} .$$

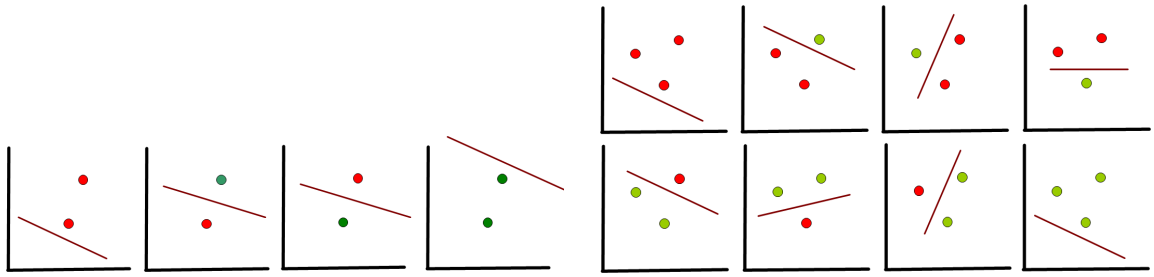

FIGURE III-3.1 – Pulvérisation une famille d'hyperplans en dimension 2 : panneau de gauche 2 points. Panneau de droite : 3 points

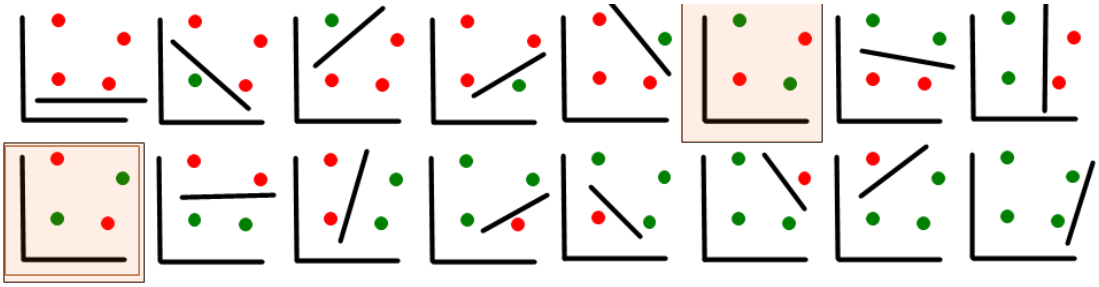

FIGURE III-3.2 – Pulvérisation d'un ensemble de quatre points par une famille d'hyperplans en dimension 2

Ces bornes élémentaires sont inutiles si N est infini (comme dans la classe des classifieurs linéaires par exemple). Soient $X_1,\dots,X_n$ n vecteurs aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$ de distribution μ. La *distribution empirique* de l'échantillon $(X_1,\dots,X_n)$ est définie pour tout $A \in \mathscr{B}(\mathbb{R}^d)$

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(X_i) \ .$$

Considérons une famille $\mathscr{A}$ d'ensembles de $\mathscr{B}(\mathbb{R}^d)$. L'objectif de ce paragraphe est l'étude de $\sup_{A\in A} |\mu_n(A) - \mu(A)|$. L'inégalité de McDiarmid implique que pour tout $n \in \mathbb{N}$ et $t > 0$,

$$\mathbb{P}\left(\left| \sup_{A\in\mathscr{A}} |\mu_n(A) - \mu(A)| - \mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mu_n(A) - \mu(A)| \right] \right| > t \right) \leq 2\mathrm{e}^{-2nt^2} \ .$$

Pour toute classe $\mathscr{A}$, l'écart maximal est concentré autour de sa moyenne. Nous allons maintenant montrer des inégalités pour

$$\mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mu_n(A) - \mu(A)| \right]$$

en fonction de certaines quantités décrivant la "complexité" combinatoire de la classe d'ensembles $\mathscr{A}$, ne dépendant pas de la distribution des observations. Etant donné n points $x_1,\dots,x_n$ de $\mathbb{R}^d$, définissons

$$N_{\mathscr{A}}(x_1,\dots,x_n) := |\{\{x_1,\dots,x_n\} \cap A : A \in \mathscr{A}\}|$$

$N_{\mathscr{A}}(x_1,\dots,x_n)$ est le nombre de sous-ensembles de $\{x_1,\dots,x_n\}$ que l'on peut obtenir par intersection de ces n points avec les ensembles $A \in \mathscr{A}$. On a toujours $N_{\mathscr{A}}(x_1,\dots,x_n) \leq 2^n$, et lorsque $N_{\mathscr{A}}(x_1,\dots,x_n) = 2^n$, on dit que la classe $\mathscr{A}$ *pulvérise* l'ensemble $\{x_1,\dots,x_n\}$.

Définition III-3.5 (Coefficient de pulvérisation de n points de la classe $\mathscr{A}$). *Soit $\mathscr{A}$ une classe d'ensembles boréliens. Le coefficient de pulvérisation de n points de la classe $\mathscr{A}$ est donné par*

$$\mathbb{S}_{\mathscr{A}}(n) = \max_{(x_1,\ldots,x_n)\in\mathbb{R}^d} N_{\mathscr{A}}(x_1,\ldots,x_n) .$$

Le théorème principal est la version suivante d'un résultat classique de Vapnik et Chervonenkis.

Théorème III-3.6 (Inégalité de Vapnik-Chervonenkis). *Soient $X_1,\ldots,X_n$ n variables aléatoires indépendantes, de même loi μ sur $\mathbb{R}^d$ et soit μ_n la mesure empirique correspondante. Alors, pour toute famille $\mathscr{A}$ d'ensembles boréliens de $\mathbb{R}^d$ et pour tout $\varepsilon > 0$, on a*

$$\mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mu_n(A) - \mu(A)|\right] \leq \sqrt{\frac{2\log 2\mathbb{S}_{\mathscr{A}}(2n)}{n}}.$$

Démonstration. Soit $(X'_1,\ldots,X'_n)$ une copie indépendante de $(X_1,\ldots,X_n)$ ($(X'_1,\ldots,X'_n)$ est indépendant de $(X_1,\ldots,X_n)$ et de même loi que $(X_1,\ldots,X_n)$). Définissons n variables indépendantes de Rademacher : $\sigma_1,\ldots,\sigma_n$ telles que $\mathbb{P}(\sigma_1 = -1) = \mathbb{P}(\sigma_1 = 1) = 1/2$ et indépendantes de (X_{1n}, X'_{1n}). Alors, en notant $A \in \mathscr{B}(\mathbb{R}^d)$,

$$\mu'_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A(X'_i),$$

nous avons

$$\begin{aligned}
\mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mu_n(A) - \mu(A)|\right] &= \mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mathbb{E}\left[\mu_n(A) - \mu'_n(A)\,\middle|\, X_{1n}\right]|\right] \\
&\leq \mathbb{E}\left[\sup_{A\in\mathscr{A}} \mathbb{E}\left[|\mu_n(A) - \mu'_n(A)|\,\middle|\, X_{1n}\right]\right] \leq \mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mu_n(A) - \mu'_n(A)|\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{A\in\mathscr{A}} \left|\sum_{i=1}^n \sigma_i(\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i))\right|\right] ,
\end{aligned}$$

où nous avons utilisé que, pour tout $A \in \mathscr{A}$, les variables aléatoires $\{\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i)\}_{i=1}^n$ et $\{\sigma_i[\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i)]\}_{i=1}^n$ ont même loi. Nous en déduisons donc que

$$\mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mu_n(A) - \mu(A)|\right] \leq \frac{1}{n}\mathbb{E}\left[\mathbb{E}\left[\sup_{A\in\mathscr{A}} \left|\sum_{i=1}^n \sigma_i(\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i))\right| \,\middle|\, X_{1n}, X'_{1n}\right]\right] .$$

Comme les variables aléatoires σ_{1n} sont indépendantes de X_{1n} et X'_{1n}, nous pouvons fixer $X_{1n} = x_{1n}$, $X'_{1n} = x'_{1n}$ et considérer

$$\mathbb{E}\left[\sup_{A\in\mathscr{A}} \left|\sum_{i=1}^n \sigma_i\{\mathbb{1}_A(x_i) - \mathbb{1}_A(x'_i)\}\right|\right] .$$

Pour tout (x_{1n}, x'_{1n}), nous notons par $\hat{\mathscr{A}}(x_{1n}, x'_{1n}) \subset \mathscr{A}$ la collection d'ensembles de $\mathscr{A}$ vérifiant les conditions suivantes

— Pour tout $B, C \in \hat{\mathscr{A}}(x_{1n}, x'_{1n})$, $B \neq C$, nous avons

$$B \cap \{x_{1n}, x'_{1n}\} \neq C \cap \{x_{1n}, x'_{1n}\} .$$

Deux ensembles distincts de $\hat{\mathscr{A}}(x_{1n}, x'_{1n})$ ont des intersections différentes avec (x_{1n}, x'_{1n})

— Pour tout $A \in \mathscr{A}$, il existe $B \in \hat{\mathscr{A}}(x_{1n}, x'_{1n})$ tel que

$$A \cap \{x_{1n}, x'_{1n}\} = B \cap \{x_{1n}, x'_{1n}\} .$$

Par construction, **chaque intersection possible de $\{x_{1n}, x'_{1n}\}$ avec un ensemble $A \in \mathscr{A}$ est présente dans $\hat{\mathscr{A}}(x_{1n}, x'_{1n})$ une fois et une seule**. Par définition du coefficient de pulvérisation de $2n$ points, nous avons $|\hat{\mathscr{A}}(x_{1n}, x'_{1n})| \leq \mathbb{S}_{\mathscr{A}}(2n)$, et

$$\mathbb{E}\left[\sup_{A\in\mathscr{A}}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{1}_A(x_i)-\mathbb{1}_A(x'_i))\right|\right] = \mathbb{E}\left[\max_{A\in\hat{\mathscr{A}}(x_{1n},x'_{1n})}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{1}_A(x_i)-\mathbb{1}_A(x'_i))\right|\right] .$$

En observant que $\sigma_i(\mathbb{1}_A(x_i) - \mathbb{1}_A(x'_i))$ est à moyenne nulle et à valeurs dans $[-1,1]$, nous obtenons en utilisant le Lemme IV-1.8 que $s > 0$,

$$\mathbb{E}[e^{s\sum_{i=1}^{n}\sigma_i(\mathbb{1}_A(x_i)-\mathbb{1}_A(x'_i))}] = \prod_{i=1}^{n}\mathbb{E}[e^{s\sigma_i(\mathbb{1}_A(x_i)-\mathbb{1}_A(x'_i))}] \leq e^{ns^2/2} .$$

Comme la distribution de la variable aléatoire $\sigma_i(\mathbb{1}_A(x_i) - \mathbb{1}_A(x'_i))$ est symétrique, le Théorème IV-1.13 (ou le Théorème III-2.7 et le Lemme III-2.9) implique immédiatement que

$$\mathbb{E}\left[\max_{A\in\hat{\mathscr{A}}}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{1}_A(x_i)-\mathbb{1}_A(x'_i))\right|\right] \leq \sqrt{2n\log 2\mathbb{S}_{\mathscr{A}}(2n)}.$$

□

Remarque III-3.7. La forme originale de l'inégalité Vapnik-Chervonenkis est la suivante

$$\mathbb{P}\left(\sup_{A\in\mathscr{A}}|\mu_n(A)-\mu(A)| > t\right) \leq 4\mathbb{S}_{\mathscr{A}}(2n)e^{-nt^2/8} .$$

En combinant le Théorème III-3.6 avec l'inégalité (III-3.3), on obtient une inégalité similaire. ♢

L'inégalité de Vapnik-Chervonenkis permet de convertir le contrôle uniforme des déviations des moyennes empiriques en un problème combinatoire, qui ne dépend pas de la loi des observations. L'étude du coefficient de pulvérisation $\mathbb{S}_{\mathscr{A}}(2n)$ est la clé de la compréhension du comportement de $\sup_{A\in\mathscr{A}}|\mu_n(A) - \mu(A)|$. Les classes pour lesquelles $\mathbb{S}_{\mathscr{A}}(2n)$ croît à un taux sous-géométrique avec le nombre d'observations n permettent de "généraliser" dans le sens où $\mathbb{E}[\sup_{A\in\mathscr{A}}|\mu_n(A) - \mu(A) - \mu(A)|]$ converge vers zéro.

III-3.2.1 Propriétés du coefficient de pulvérisation

Nous allons établir dans la Section III-3.2.1 quelques propriétés importantes du coefficient de pulvérisation. Soit $\mathscr{A}$ une classe d'ensembles boréliens de $\mathbb{R}^d$, et soit $x_{1n} \in \mathbb{R}^d$ des points arbitraires. Commençons par établir des propriétés élémentaires du coefficient de pulvérisation de n points.

Théorème III-3.8. *Soient $\mathscr{A}$ et $\mathscr{B}$ des classes d'ensembles boréliens de $\mathbb{R}^d$, et $n, m \geq 1$ des entiers. Alors*

(i) $\mathbb{S}_{\mathscr{A}}(n+m) \leq \mathbb{S}_{\mathscr{A}}(n)\mathbb{S}_{\mathscr{A}}(m)$;

(ii) *Si* $\mathscr{C} = \mathscr{A} \cup \mathscr{B}$, *alors* $\mathbb{S}_{\mathscr{C}}(n) \leq \mathbb{S}_{\mathscr{A}}(n) + \mathbb{S}_{\mathscr{B}}(n)$;

(iii) *Si* $\mathscr{C} = \{C = A^c : A \in \mathscr{A}\}$, *alors* $\mathbb{S}_{\mathscr{C}}(n) = \mathbb{S}_{\mathscr{A}}(n)$;

(iv) *Si* $\mathscr{C} = \{C = A \cap B : A \in \mathscr{A}\ et\ B \in \mathscr{B}\}$, *alors* $\mathbb{S}_{\mathscr{C}}(n) \leq \mathbb{S}_{\mathscr{A}}(n)\mathbb{S}_{\mathscr{B}}(n)$;

(v) *Si* $\mathscr{C} = \{C = A \cup B : A \in \mathscr{A}\ et B \in \mathscr{B}\}$, *alors* $\mathbb{S}_{\mathscr{C}}(n) \leq \mathbb{S}_{\mathscr{A}}(n)\mathbb{S}_{\mathscr{B}}(n)$;

(vi) *Si* $\mathscr{C} = \{C = A \times B : A \in \mathscr{A}\ et\ B \in \mathscr{B}\}$, *alors* $\mathbb{S}_{\mathscr{C}}(n) \leq \mathbb{S}_{\mathscr{A}}(n)\mathbb{S}_{\mathscr{B}}(n)$.

Démonstration. Les assertions (i), (ii), (iii), et (vi) découlent de façon immédiate de la définition.

Pour montrer (iv), prenons $x_{1n} = \{x_1, \ldots, x_n\}$ et posons $N = |\mathscr{A}(x_{1n})| \leq \mathbb{S}_{\mathscr{A}}(n)$. Notons $A_1, A_2, \ldots, A_N$ les N ensembles (deux à deux différents) de la forme $\{x_1, \ldots, x_n\} \cap A$, $A \in \mathscr{A}$. Pour tout $i \in \{1, \ldots, N\}$, nous avons $\mathbb{S}_{\mathscr{B}}(|A_i|) \leq \mathbb{S}_{\mathscr{B}}(n)$ et,

$$|\mathscr{C}(x_{1n})| \leq \sum_{i=1}^{N} \mathbb{S}_B(|A_i|) \leq \mathbb{S}_{\mathscr{A}}(n)\mathbb{S}_{\mathscr{B}}(n) \; .$$

(v) découle de (iv) et (iii). □

Définition III-3.9 (Dimension de Vapnik-Chervonenkis (dim-VC) **d'une classe d'ensembles).** *La dimension de Vapnik-Chervonenkis* dim-VC($\mathscr{A}$) *de la classe $\mathscr{A}$ d'ensembles boréliens est le plus grand entier n tel que*

$$\mathbb{S}_{\mathscr{A}}(n) = 2^n.$$

Si $\mathbb{S}_{\mathscr{A}}(n) = 2^n$ pour tout n, nous posons dim-VC($\mathscr{A}$) $= \infty$.

Clairement, si $\mathbb{S}_{\mathscr{A}}(n) < 2^n$ pour certains n, alors pour tout $m > n$, $\mathbb{S}_{\mathscr{A}}(m) < 2^m$ (voir Théorème III-3.8-(i)), et donc la dimension dim-VC est toujours bien définie.

Exemple III-3.10 (Dim-VC de la classe des hyperplans linéaires). Considérons la classe $\mathscr{L}_d$ des sous-ensembles de $\mathbb{R}^d$ de la forme $\left\{x \in \mathbb{R}^d \,:\, a^T x \geq 0\right\}$ pour tout $a \in \mathbb{R}^d$. Nous allons montrer que $V \leq d$. Considérons tout d'abord l'ensemble $S = \{e_1, \ldots, e_d\}$ des vecteurs canoniques de $\mathbb{R}^d$: pour $i \in \{1, \ldots, d\}$, $e_{i,i} = 1$ et $e_{i,j} = 0$, si $1 \leq j \neq i \leq d$. Nous allons montrer que cet ensemble est pulvérisé par $\mathscr{L}_d$. En effet, pour tout $(y_1, \ldots, y_d) \in \{-1, +1\}^d$, considérons l'ensemble

$$A_{y_1,\ldots,y_d} = \left\{ x \in \mathbb{R}^d \,:\, \sum_{i=1}^{d} y_i e_i^T x \geq 0 \right\} .$$

Un point e_i appartient à l'intersection $A_{y_1,\ldots,y_d} \cap S$ si et seulement si $y_i = +1$. Par conséquent,

$$N_{\mathscr{L}_d}(e_1, \ldots, e_d) \geq |\left\{ S \cap A_{y_1,\ldots,y_d} \,:\, (y_1, \ldots, y_d) \in \{-1, +1\}^d \right\}| = 2^d.$$

Supposons maintenant qu'il existe un ensemble $\{x_1, \ldots, x_{d+1}\}$ de $d+1$ pulvérisé par $\mathscr{L}_d$. Ceci implique que pour les 2^{d+1} sous ensembles possibles $\{S_k\}_{k=1}^{2^{d+1}}$ de $\{x_1, \ldots, x_{d+1}\}$, il est possible de trouver un vecteur $\{w_k\}_{k=1}^{2^{d+1}}$ tels que

$$S_k = \{x_1, \ldots, x_{d+1}\} \cap \left\{ x \in \mathbb{R}^d \,:\, w_k^T x \geq 0 \right\} .$$

Considérons la matrice obtenue en concaténant les sorties réelles de tous ces classifieurs

$$\mathbf{H} = \begin{bmatrix} w_1^T x_1 & w_2^T x_2 & \ldots & w_{2^{d+1}}^T x_1 \\ \vdots & \ddots & & \vdots \\ w_1^T x_{d+1} & \ldots & & w_{2^{d+1}}^T x_{d+1} \end{bmatrix} = \mathbf{X}\mathbf{W} \, , \tag{III-3.4}$$

où nous avons introduit $\mathbf{X} = [x_1, \ldots, x_{d+1}]^T$ et $\mathbf{W} = [w_1, \ldots, w_{2^{d+1}}]$. Considérons la matrice $\mathbf{G} = \text{signe}(\mathbf{H})$ où $\mathbf{G}_{i,j} = \text{signe}(\mathbf{H}_{i,j})$, $i \in \{1, \ldots, d+1\}$, $j \in \{1, \ldots, 2^{d+1}\}$. Les colonnes de la matrices $\mathbf{G}$ contiennent les 2^{d+1} combinaisons de signes possibles dans un ensemble de $d+1$ éléments.

Soit $a \in \mathbb{R}^{d+1}$ un vecteur. La k-ème coordonnée du vecteur $a^T\mathbf{H}$ est égale à $a^T\mathbf{X}w_k$. Il existe k tel que $\text{signe}(\mathbf{X}w_k) = \text{signe}(a)$. Par conséquent, $a^T\mathbf{X}w_k > 0$ (somme de termes tous positifs, dont un au moins est strictement positif). Ceci montre que $a^T\mathbf{H} \neq \mathbf{0}_{1\times 2^{d+1}}$ et donc que les lignes de matrice $\mathbf{H}$ sont linéairement indépendantes. Ceci implique que $\text{Rang}(\mathbf{H}) = d+1$. Mais comme $\mathbf{H} = \mathbf{X}\mathbf{W}$, nous avons

$$\text{Rang}(\mathbf{H}) \leq \min\left(\text{Rang}(\mathbf{X}), \text{Rang}(\mathbf{W})\right) \leq d$$

car $\mathbf{X}$ est une matrice $(d+1) \times d$. Nous aboutissons à une contradiction. Il n'y a pas d'ensemble de $d+1$ points qui puissent être pulvérisés : ceci implique que dim-VC($\mathscr{L}_d$) $\leq d$ et comme nous avions déjà établi que dim-VC($\mathscr{L}_d$) $\geq d$, nous en déduisons dim-VC($\mathscr{L}_d$) $= d$. ♢

Exemple III-3.11 (Dimension VC de la classe des hyperplans affines). Considérons maintenant la classe $\mathscr{L}_f^a$ des sous-ensembles $\left\{x \in \mathbb{R}^d \,:\, a^T x + b \geq 0\right\}$ où $a \in \mathbb{R}^d$ et $b \in \mathbb{R}$. Par rapport à la classe des hyperplans, nous avons rajouté un paramètre qui est le décalage de l'hyperplan par rapport à l'origine de l'espace. Notons que $\mathscr{L}_d \subset \mathscr{L}_d^a$ et donc $d = \text{dim-VC}(\mathscr{L}_d) \leq \text{dim-VC}(\mathscr{L}_d^a)$. En fait, comme nous allons le montrer ci-dessous, $\text{dim-VC}(\mathscr{L}_d^a) = d+1$.

Nous établissons tout d'abord que $\text{dim-VC}(\mathscr{L}_d^a) \geq d+1$. Il suffit pour cela de construire un ensemble de $d+1$ éléments de $\mathbb{R}^d$ qui soit pulvérisé par $\mathscr{L}_d^a$. Considérons à cet effet l'ensemble $S = \{e_1, \ldots, e_d\}$ des vecteurs de la base canonique de $\mathbb{R}^d$ auquel nous adjoignons l'origine 0 de l'espace affine. $(y_1, \ldots, y_{d+1}) \in \{-1,+1\}^{d+1}$, considérons l'ensemble

$$B_{y_1,\ldots,y_{d+1}} = \left\{ x \in \mathbb{R}^d \,:\, f_y(x) := \sum_{i=1}^{d} (y_i - y_{d+1}) e_i^T x + y_{d+1} \geq 0 \right\} .$$

Notons que $f_{y_1,\ldots,y_{d+1}}(e_j) = y_j$ pour $j \in \{1,\ldots,d\}$ et que $f_{y_1,\ldots,y_{d+1}}(0) = y_{d+1}$, nous avons donc

$$N_{S\cup\{0\}}(\mathscr{L}_d^a) \geq |\left\{ (S\cup\{0\}) \cap B_{y_1,\ldots,y_{d+1}} \,:\, (y_1,\ldots,y_{d+1}) \in \{-1,+1\}^{d+1} \right\}| \geq 2^{d+1} .$$

La borne supérieure de la dimension VC est une conséquence directe du fait qu'un hyperplan affine est un hyperplan linéaire d'un espace $\mathbb{R}^{d+1}$. ♢

Exemple III-3.12 (Dimension VC de la classe des hyper-rectangles). Nous nous intéressons maintenant à la classe $\mathscr{R}_d$ de tous les hyper-rectangles de $\mathbb{R}^d$ dont les côtés sont parallèles aux axes de coordonnées.

$$\mathscr{R}_d = \{[a_1,b_1] \times \cdots \times [a_d,b_d] \,:\, -\infty < a_i \leq b_i < \infty, i \in \{1,\ldots,d\}\} . \tag{III-3.5}$$

Nous allons tout d'abord montrer que $\mathbb{S}_{\mathscr{R}_d}(2d) \geq 2^{2d}$. Nous notons comme précédemment l'ensemble $S = \{e_1, -e_1, e_2, -e_2, \ldots, e_d, -e_d\}$ où e_i est le i-ème vecteur canonique de $\mathbb{R}^d$. Le cardinal de l'ensemble S est égal à $2d$ vecteurs ; les composantes d'un vecteur quelconque de S sont nulles à l'exception d'une qui est égale à $+1$ ou -1. Nous considérons l'ensemble des rectangles $\tilde{\mathscr{R}}_d$ Le cardinal de cet ensemble est égal à 2^{2d} et il est élémentaire d'établir que cet ensemble pulvérise S. Par consequent, $\mathbb{S}_{\mathscr{R}_d}(2d) \geq 2^{2d}$.

Considérons maintenant un ensemble quelconque de $2d+1$ éléments de $\mathbb{R}^d$, $S = \{x_1, \ldots, x_{2d+1}\}$. Pour $i \in \{1,\ldots,d+1\}$ et $j \in \{1,\ldots,d\}$, nous notons $x_{i,j}$ la j-ème coordonnée du vecteur x_i. Nous notons

$$k_1 = \arg\min(x_{i,1}, i \in \{1,\ldots,d+1\}) \quad \text{et} \quad \ell_1 = \arg\max(x_{i,1}, i \in \{1,\ldots,d+1\} \setminus \{k_1\})$$

et par récurrence, pour $j \in \{2,\ldots,d\}$

$$\begin{aligned} k_j &= \arg\min(x_{i,j}, i \in \{1,\ldots,d+1\} \setminus \{k_1,\ell_1,\ldots,k_{j-1},\ell_{j-1}\}) \\ \ell_j &= \arg\max(x_{i,j}, i \in \{1,\ldots,d+1\} \setminus \{k_1,\ell_1,\ldots,k_{j-1},\ell_{j-1},k_j\}) \end{aligned}$$

Finalement, nous appelons $\{k_{d+1}\} = \{1,\ldots,d+1\} \setminus \{k_1,\ell_1,\ldots,k_d,\ell_d\}$. Nous notons $y_i = x_{k_i}$ et $z_i = x_{\ell_i}$ pour $i \in \{1,\ldots,d\}$. Par construction, pour $i \in \{1,\ldots,d\}$,

$$y_{i,i} \leq x_{k_{d+1},i} \leq z_{i,i} .$$

ou, de façon équivalente, $x_{k_{d+1}} \in \left[y_{i,1}, z_{i,1}\right] \times \cdots \times \left[y_{d,d}, z_{d,d}\right]$ Il n'est donc pas possible de construire un hyper-rectangle contenant les $2d$ points $\{y_1, z_1, \ldots, y_d, z_d\}$ et qui ne contienne pas $x_{k_{d+1}}$ On en déduit que : $\mathbb{S}_{\mathscr{A}}(2d+1) < 2^{2d+1}$. Par conséquent, $\text{dim-VC}(\mathscr{R}_d) = 2d$. ♢

Comme le montre le résultat de base suivant, la dimension dim-VC fournit une borne utile pour le coefficient d'éclatement d'une classe.

Théorème III-3.13 (Lemme de Sauer). *Soit $\mathscr{A}$ une classe d'ensembles boréliens de dimension* dim-VC $V = \text{dim-VC}(\mathscr{A}) < \infty$. *Alors, pour tout n,*

$$\mathbb{S}_{\mathscr{A}}(n) \leq \sum_{i=0}^{V} \binom{n}{i} .$$

Démonstration. On dit que $B \subset \{0,\ 1\}^n$ pulvérise $S = \{s_1, \ldots, s_m\} \subset \{1, \ldots, n\}$ si

$$\{\ (b_{s_1},\ b_{s_m}) :\ \ (b_1,\ b_n)\ \in B\} = \{0,\ 1\}^m\ .$$

De façon équivalente, B pulvérise S si la restriction de B aux coordonnées données par S est $\{0,\ 1\}^m$ tout entier. On fixe maintenant $(x_1, \ldots, x_n)$ tels que $|\mathscr{A}(x_{1n})| = \mathbb{S}_{\mathscr{A}}(n)$, et on pose $B_0 = \mathscr{A}(x_{1n})$.

Pour montrer le résultat, on constate tout d'abord que B_0 ne peut par définition de la dimension de Vapnik-Chervonenkis V pulvériser aucun ensemble de taille $m > V$. En effet, dans le cas contraire, on aurait $\mathbb{S}_{\mathscr{A}}(m) = 2^m$.

On va maintenant effectuer une transformation de l'ensemble B_0 qui préserve son cardinal. Pour $b = (b_1, \ldots, b_n) \in B_0$, on appelle b' l'élément suivant :

— Si $b_1 = 1$, on pose $b' = (0,\ b_2,\ b_n)$ si ce vecteur n'est pas déjà dans B_0 et $b' = b$ sinon.

— Si $b_1 = 0$, on pose $b' = b$.

On définit ainsi un ensemble B_1 constitué des éléments b' qui est alors clairement de même cardinal que B_0. On remarque alors que si B_1 pulvérise $S = \{s_1, \ldots, s_m\} \subset \{1,\ n\}$ alors il en est de même de B_0. Cette assertion est triviale si $1 \notin S$. Vérifions maintenant qu'elle est satisfaite dans le cas où $1 \in S$. Sans perte de généralité, on suppose que $s_1 = 1$. Alors, quel que soit $v \in \{0,\ 1\}^{m-1}$, il existe $b \in B_1$ tel que $b_1 = 1$ et $(b_{s_2},\ b_{s_m}) = v$. Par construction de B_1, on a $b \in B_0$ et $(0,\ b_2,\ b_n) \in B_0$. Donc B_0 pulvérise aussi S.

On recommence cette transformation en agissant maintenant sur les deuxièmes coordonnées des éléments de B_1. On obtient ainsi B_2, puis on réitère le procédé pour arriver à B_n. Clairement, si B_n pulvérise un ensemble de taille $m > V$, alors B_0 aussi. Comme on l'a vu pour B_1, cela revient à dire que si $b \in B_n$ alors les vecteurs de la forme $c = (c_1, \ldots, c_n)$, où $c_i = b_i$ ou 0 sont aussi dans B_n, ce qui implique que tout élément de B_n a au plus V coordonnées valant 1 (sinon B_n pulvérise un ensemble dont le cardinal est trop grand). Par conséquent :

$$|B_0|\ =\ |B_n|\ \leq \sum_{k=0}^{V} \binom{n}{k}\ .$$

□

Une conséquence importante du Lemme de Sauer, que nous utiliserons de façon fréquente dans la suite, est donnée dans le corollaire suivant.

Corollaire III-3.14. *Soit $\mathscr{A}$ une classe d'ensembles boréliens de dimension* dim-VC $V < \infty$. *Alors pour tout $n \in \mathbb{N}$,*

$$\mathbb{S}_{\mathscr{A}}(n) \leq (n+1)^V,$$

et pour tout $n \geq V$,

$$\mathbb{S}_{\mathscr{A}}(n) \leq \left(\frac{ne}{V}\right)^V\ .$$

Démonstration. Nous avons en utilisant la formule de binôme

$$(n+1)^V = \sum_{i=0}^{V} n^i \binom{V}{i} = \sum_{i=0}^{V} \frac{n^i V!}{i!(V-i)!} \geq \sum_{i=0}^{V} \frac{n^i}{i!} \geq \sum_{i=0}^{V} \binom{n}{i}\ .$$

Si $V/n \leq 1$, alors

$$\left(\frac{V}{n}\right)^V \sum_{i=0}^{V} \binom{n}{i} \leq \sum_{i=0}^{V} \left(\frac{V}{n}\right)^i \binom{n}{i} \leq \sum_{i=0}^{n} \left(\frac{V}{n}\right)^i \binom{n}{i} = \left(1+\frac{V}{n}\right)^n \leq \mathrm{e}^V\ .$$

□

L'inégalité de Vapnik-Chervonenkis (Théorème III-3.6) montre que si $\mathscr{A}$ est une classe d'ensembles boréliens de dimension $V =$ dim-VC$(\mathscr{A})$, alors

$$\mathbb{E}\left[\sup_{A \in \mathscr{A}} |\mu_n(A) - \mu(A)|\right] \leq 2\sqrt{\frac{V \log(n+1) + \log 2}{n}},$$

et donc, si la classe d'ensembles boréliens $\mathscr{A}$ a une dimension dim-VC finie, $\mathbb{E}\left[\sup_{A\in\mathscr{A}} |\mu_n(A)-\mu(A)|\right]$ converge vers 0 à un taux $O(\sqrt{\log n}/n)$.

Lemme III-3.15. *Soit $\mathscr{G}$ un sous-espace vectoriel de dimension m de l'espace vectoriel des fonctions mesurables de $\mathbb{R}^d$ à valeurs réelles. La dimension VC de l'ensemble*

$$\mathscr{A} = \left\{ \left\{ x \in \mathbb{R}^d \,:\, g(x) \geq 0 \right\} \,:\, g \in \mathscr{G} \right\}$$

est inférieure à m : dim-VC$(\mathscr{A}) \leq m$.

Démonstration. Il suffit de montrer qu'un ensemble de cardinal $m+1$ ne peut être pulvérisé par les ensembles de $\mathscr{A}$. Soit $x_{1n+1} = (x_1, \ldots, x_{m+1}) \in \mathbb{R}^d$. On définit $\phi : \mathscr{G} \to \mathbb{R}^{m+1}$ par :

$$\phi(g) = (g(x_1), \ldots, g(x_{m+1}))$$

L'image de ϕ est un sous-espace vectoriel de $\mathbb{R}^{m+1}$ contenu dans un hyperplan, donc il existe un vecteur non nul $a = (a_1, \ldots, a_{m+1})$ tel que, pour tout $g \in \mathscr{G}$

$$a_1 g(x_1) + \cdots + a_{m+1} g(x_{m+1}) = 0$$

Si tous les coefficients $a_1, \ldots, a_{m+1}$ sont positifs, alors il est clair que les ensembles de $\mathscr{A}$ ne peuvent pulvériser $\{x_1,\ x_{m+1}\}$, sinon on choisit $g \in \mathscr{G}$ telle que $g(x_i) < 0$ pour tout $i \in \{1, \cdots, m+1\}$, et on aurait alors $a_1 g(x_1) + \cdots + a_{m+1} g(x_{m+1}) < 0$.

Supposons maintenant que les coefficients $\{a_i\}_{i=1}^{m+1}$ ne sont pas tous de même signe. Nous avons alors :

$$\sum_{i, a_i \geq 0} a_i g(x_i) = \sum_{i, a_i < 0} -a_i g(x_i)$$

Supposons qu'il existe $g \in \mathscr{G}$ qui soit telle que $\left\{ x \in \mathbb{R}^d \,:\, g(x) > 0 \right\}$ contienne exactement les coordonnées x_i telles que $a_i \geq 0$. Dans ce cas, tous les termes du membre de gauche sont positifs alors que les termes du membre de droite sont négatifs. On aboutit ainsi à une contradiction. □

Corollaire III-3.16. *(i) Soit $\mathscr{A}$ la classe de toutes les boules de $\mathbb{R}^d$, i.e. les ensembles de la forme*

$$\left\{ x \in \mathbb{R}^d \,:\, \|x-a\|^2 \leq b \right\} \quad \textit{avec } a \in \mathbb{R}^d \textit{ et } b \in \mathbb{R}^d.$$

Alors dim-VC$(\mathscr{A}) \leq d+2$.

(ii) Soit $\mathscr{A}$ la classe de tous les ellipsoïdes de $\mathbb{R}^d$, i.e. les ensembles de la forme

$$\left\{ x \in \mathbb{R}^d \,:\, x^T \Sigma^{-1} x \leq 1 \right\}, \quad \textit{où } \Sigma \textit{ est une matrice définie positive}.$$

Alors dim-VC$(\mathscr{A}) \leq d(d+1)/2+1$.

III-3.2.2 Applications à la minimisation du risque empirique

Dans cette section, nous appliquons les principaux résultats des sections précédentes pour obtenir les limites supérieures de la performance des règles de classification minimisant le risque empirique.

Rappelons brièvement le contexte : $\mathscr{C}$ est une classe de classifieurs contenant des fonctions de décision de la forme $g : \mathbb{R}^d \to \{0,1\}$. Nous utilisons les données d'apprentissage (X_1, Y_1), …, (X_n, Y_n) pour calculer l'erreur empirique $\widehat{\mathrm{R}}_n(g)$ pour tout $g \in \mathscr{C}$. Un minimiseur du risque empirique g_n^* est une règle de classification vérifiant

$$\widehat{\mathrm{R}}_n(\hat{g}_n^*) \leq \widehat{\mathrm{R}}_n(g) \quad \text{for all } g \in \mathscr{C}.$$

Nous notons par $L_{\mathscr{C}}$ le minimum du risque sur la classe $\mathscr{C}$, c'est-à-dire

$$L_{\mathscr{C}} = \inf_{g \in \mathscr{C}} R(g) \ .$$

(Ici, nous supposons implicitement que l'infimum est atteint. Cette hypothèse est essentiellement motivée par la commodité de la notation, mais elle n'est pas essentielle). L'inégalité de base du Théorème III-2.2 montre que

$$L(\hat{g}_n^*) - L_{\mathscr{C}} \leq 2 \sup_{g \in \mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|.$$

Ainsi, la quantité d'intérêt est le supremum sur la classe de classifieurs de la différence entre les probabilités empiriques d'erreur et leur espérance sur la classe. Ces quantités sont estimées par l'inégalité de Vapnik-Chervonenkis (Théorème III-3.6). En effet, la variable aléatoire $\sup_{g \in \mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|$ est de la forme $\sup_{A \in \mathscr{A}} |\mu_n(A) - \mu(A)|$, où le rôle de la classe des ensembles $\mathscr{A}$ est maintenant joué par la classe $\mathscr{A}^{\bar{}}$ des ensembles d'erreurs, définis par

$$\left\{ \left\{ (x,y) \in \mathbb{R}^d \times \{0,1\} \ : \ g(x) \neq y \right\} \ : \ g \in \mathscr{C} \right\} \ .$$

L'inégalité de Vapnik-Chervonenkis borne $\sup_{g \in \mathscr{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|$ en fonction des coefficients d'éclatement (ou de la dimension dim-VC) de la classe $\mathscr{A}^{\bar{}}$ des ensembles d'erreurs.

Au lieu des ensembles d'erreurs, il est plus pratique de travailler avec des classes d'ensembles de la forme

$$\left\{ \left\{ x \in \mathbb{R}^d \ : \ g(x) = 1 \right\} \ : \ g \in \mathscr{C} \right\} \ .$$

Nous désignons la classe d'ensembles ci-dessus par $\mathscr{A}$. Le fait simple suivant montre que les classes $\mathscr{A}^{\bar{}}$ et $\mathscr{A}$ sont équivalentes d'un point de vue combinatoire.

Lemme III-3.17. *Pour chaque $n \in \mathbb{N}$ nous avons $\mathbb{S}_{\mathscr{A}^{\bar{}}}(n) = \mathbb{S}_{\mathscr{A}}(n)$, et donc les dimensions correspondantes* dim-VC *sont aussi égales : $V_{\mathscr{A}^{\bar{}}} = V_{\mathscr{A}}$.*

Démonstration. Soit N un entier positif. Nous allons montrer que pour tout ensemble de n points $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ de $\mathbb{R}^d \times \{0,1\}$, si N ensembles de $\mathscr{A}^{\bar{}}$ sélectionnent N sous-ensembles de n paires, alors il existe N ensembles de $\mathscr{A}$ qui sélectionnent N sous-ensembles de $\{x_1,\ldots,x_n\}$, et vice versa.

Plus précisément, considérons n paires $(x_1,0)$, …, $(x_m,0)$, $(x_{m+1},1),\ldots,(x_n,1)$. Puisque l'ordre des paires ne joue aucun rôle, nous pouvons représenter n'importe quel ensemble de n paires cette manière. Supposons que pour un certain ensemble $A \in \mathscr{A}$, l'ensemble correspondant

$$\overline{A} = A \times \{0\} \cup A^c \times \{1\} \in \mathscr{A}^{\bar{}}$$

choisit les paires $(x_1,0),\ldots,(x_k,0)$, $(x_{m+1},1)$, …, $(x_{m+l},1)$, c'est-à-dire que l'ensemble de ces paires est l'intersection de $\mathscr{A}^{\bar{}}$ et des paires n. Encore une fois, nous pouvons supposer sans perte de généralité que les paires sont ordonnées de cette façon. Cela signifie que l'ensemble A sélectionne dans l'ensemble $\{x_1,\ldots,x_n\}$ le sous-ensemble $\{x_1,\ldots,x_k,x_{m+l+1},\ldots,x_n\}$, et les deux sous-ensembles sont associés de façon biunivoque. Ceci prouve que $\mathbb{S}_{\mathscr{A}^{\bar{}}}(n) \leq \mathbb{S}_{\mathscr{A}}(n)$. Pour prouver la réciproque, remarquez que si A choisit un sous-ensemble de k points $\{x_1,\ldots,x_k\} \subset \{x_1,\ldots,x_n\}$, alors l'ensemble correspondant $\overline{A} \in \mathscr{A}^{\bar{}}$ choisit les paires avec les mêmes indices $\{(x_1,0),\ldots,(x_k,0)\}$.

L'égalité des dimensions dim-VC découle de l'égalité des coefficients d'éclatement pour tout n. □

Nous noterons dans la suite de l'exposé la valeur commune de $\mathbb{S}_{\mathscr{A}^{\bar{}}}(n)$ et $\mathbb{S}_{\mathscr{A}}(n)$ par $\mathbb{S}_{\mathscr{C}}(n)$, que nous appellerons le coefficient de pulvérisation de n-points de la famille de règles de classification $\mathscr{C}$. Il s'agit simplement du nombre maximum de façons différentes dont les points n peuvent être classés par les classifieurs dans la famille $\mathscr{C}$. De même, nous appellerons $V_{\mathscr{A}^{\bar{}}} = V_{\mathscr{A}}$ la dimension dim-VC de la famille de classifieurs $\mathscr{C}$, que nous noterons $V_{\mathscr{C}}$.

Nous disposons de tous les éléments nécessaires pour présenter les principales performances du minimiseur du risque empirique.

Corollaire III-3.18.

$$\mathbb{E}[L(\hat{g}_n^*) - L_{\mathscr{C}}] \leq \sqrt{2\frac{\log(2\mathbb{S}_{\mathscr{C}}(2n))}{n}}\,.$$

Nous pouvons aisément déduire de ce résultat une borne pour $\mathbb{P}(L(\hat{g}_n^*) - L_{\mathscr{C}} > \varepsilon)$ en utilisant le Théorème III-3.2.

Quatrième partie

Outils probabilistes

Chapitre IV-1

Inégalités de déviations

IV-1.1 Inégalités de Markov et de Bienayme-Tchebychev

Lemme IV-1.1 (Markov). *Soit Z une v.a. positive et $p > 0$ telle que $\mathbb{E}(Z^p) < \infty$. Alors, pour tout $\delta > 0$, nous avons :*

$$\mathbb{P}(Z \geq \delta) \leq \delta^{-p}\, \mathbb{E}(Z^p)\ .$$

Démonstration. On écrit

$$\mathbb{E}[Z^p] = \mathbb{E}[Z^p \mathbb{1}_{[\delta,\infty[}(Z)] + \mathbb{E}[Z^p \mathbb{1}_{[0,\delta[}(Z)] \geq \mathbb{E}[Z^p \mathbb{1}_{[\delta,\infty]}(Z)] \geq \delta^p \mathbb{P}(Z \geq \delta)\ ,$$

où $\mathbb{1}_A$ désigne la fonction indicatrice de l'ensemble A. □

Cette inégalité conduit, lorsque $p = 2$, à l'inégalité bien connue de Bienayme-Tchebychev (nous noterons dans la littérature les nombreuses retranscriptions possibles de ce nom !).

Lemme IV-1.2 (Inégalité de Bienayme-Tchebychev). *Soit Z une variable aléatoire réelle vérifiant $\mathbb{E}[Z^2] < \infty$. Notons $\mu = \mathbb{E}[Z]$. Alors, pour tout $\delta > 0$:*

$$\mathbb{P}(|Z - \mu| \geq \delta) \leq \mathrm{Var}(Z)/\delta^2\ . \tag{IV-1.1}$$

Notons au passage que l'inégalité de Bienayme-Tchebychev n'est pas précise. En particulier, la borne apparaissant dans le membre de droite peut éventuellement être plus grande que 1 (en fait dès que $\delta < \sqrt{\mathrm{Var}(Z)}$). Il est possible, à moindres frais, de proposer un énoncé de cette inégalité ne souffrant pas de ce problème :

Lemme IV-1.3 (Inégalité de Tchebychev-Cantelli). *Soit Z une variable aléatoire réelle vérifiant $\mathbb{E}[Z^2] < \infty$. Alors, pour tout $\delta > 0$,*

$$\mathbb{P}(Z - \mathbb{E}(Z) \geq \delta) \leq \frac{\mathrm{Var}(Z)}{\mathrm{Var}(Z) + \delta^2}.$$

Démonstration. On peut supposer sans perte de généralité que $\mathbb{E}[Z] = 0$. Par suite, pour tout $\delta > 0$, nous avons :

$$\delta = \mathbb{E}[(\delta - Z)] \leq \mathbb{E}[(\delta - Z)\mathbb{1}_{]-\infty,\delta]}(Z)]\ .$$

L'inégalité de Cauchy-Schwarz montre que :

$$\delta^2 \leq \mathbb{E}[(\delta - Z)^2]\mathbb{P}(Z \leq \delta) = (\delta^2 + \mathrm{Var}(Z))\mathbb{P}(Z \leq \delta),$$

ce qui donne le résultat désiré. □

IV-1.2 Inégalité de Jensen

Lemme IV-1.4 (Inégalité de Jensen). *Soit* I *un intervalle ouvert de* $\mathbb{R}$, f *une fonction convexe sur I, Y une variable aléatoire réelle telle que* $\mathbb{P}(Y \in \mathsf{I}) = 1$ *et* $\mathbb{E}[|Y|] < \infty$. *Alors*

$$f(\mathbb{E}[Y]) \leq \mathbb{E}[f(Y)] \,. \tag{IV-1.2}$$

Si la fonction f est strictement convexe, alors l'inégalité est stricte dès que la variable aléatoire Y n'est pas presque sûrement constante (i.e. $\mathbb{P}(Y \neq \mathbb{E}[Y]) > 0$).

Démonstration. Si f est une fonction convexe sur un intervalle ouvert I alors pour tout $t \in \mathsf{I}$, il existe une constante c_t telle que

$$f(t) + c_t(y-t) \leq f(y) \,, \quad \text{pour tout } y \in \mathsf{I}. \tag{IV-1.3}$$

L'inégalité est stricte si f est strictement convexe et $y \neq t$. En appliquant cette inégalité avec $t = \mathbb{E}[Y]$, nous avons donc

$$f(\mathbb{E}[Y]) + c(y - \mathbb{E}[Y]) \leq f(y) \,, \quad \text{pour tout } y \in \mathsf{I}$$

et par conséquent (IV-1.2) découle de

$$f(\mathbb{E}[Y]) + c(Y - \mathbb{E}[Y]) \leq f(Y) \,, \mathbb{P} - \text{p.s.} \,.$$

Si f est strictement convexe, et Y n'est pas presque-sûrement constante, alors (IV-1.3) est stricte sur l'évènement $\{Y \neq \mathbb{E}[Y]\}$ qui est de probabilité strictement positive. □

IV-1.3 Inégalités de Chernoff

L'inégalité la plus simple pour borner la probabilité de la différence entre une variable aléatoire et son espérance est l'inégalité de Bienayme-Tchebychev (voir Lemme IV-1.2). Supposons que $X_1, \ldots, X_n$ sont des variables aléatoires réelles indépendantes. Nous cherchons à borner la probabilité de queue $\mathbb{P}(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| \geq \varepsilon)$ avec $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. L'inégalité de Bienayme-Tchebychev et l'indépendance des variables aléatoires $\{X_i\}_{i=1}^n$ impliquent immédiatement

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sum_{i=1}^n \mathrm{Var}(X_i)}{n^2\varepsilon^2}.$$

La précision de cette inégalité est peut-être plus facile à appréhender si nous supposons que les variables aléatoires $\{X_i\}_{i=1}^n$ sont i.i.d. et distribuées suivant une loi de Bernoulli de paramètre p - (i.e., $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p$).

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - p\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2}. \tag{IV-1.4}$$

Cette borne est très imprécise, comme nous allons maintenant le voir. Soit $\Phi(y) = \int_{-\infty}^{y} \mathrm{e}^{-t^2/2}/\sqrt{2\pi}\mathrm{d}t$ la fonction de répartition de la loi normale centrée réduite. Le théorème de la limite centrale montrer que

$$n^{-1/2}\sum_{i=1}^n \{X_i - p\} \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0, \mathrm{Var}(X_1)) \,. \tag{IV-1.5}$$

Comme les variables aléatoires $\{X_i,\ i \in \mathbb{N}\}$ sont binômiales, nous avons $\mathrm{Var}(X_1) = p(1-p)$. Nous avons par conséquent,

$$\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{i=1}^n X_i - p\right) \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0,1) \,,$$

ce qui implique que, pour tout $y > 0$,

$$\mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p\right) \geq y\right) \to \int_y^\infty \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\mathrm{d}t$$
$$\leq \frac{1}{y}\int_y^\infty \frac{t}{\sqrt{2\pi}} \exp(-t^2/2)\mathrm{d}t \leq \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-y^2/2}}{y},$$

Cette propriété suggère une inégalité de la forme, pour $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \varepsilon\right) \approx n^{-1/2}\,\mathrm{e}^{-n\varepsilon^2/(2p(1-p))}. \tag{IV-1.6}$$

Même s'il l'argument précédent est heuristique (le théorème de limite centrale ne permet pas de contrôler des "queues" de distribution, il donne clairement à penser que l'inégalité de Chebyshev (Equation (IV-1.4)) est très pessimiste.

Une amélioration significative de cette borne peut être obtenue en utilisant une technique due à Chernoff. On va encore utiliser l'inégalité de Markov, mais de façon plus astucieuse.

Une application directe de l'inégalité de Markov montre que, si λ est un nombre positif arbitraire, alors pour toute variable aléatoire X, et tout $t > 0$,

$$\mathbb{P}(X \geq t) = \mathbb{P}(\mathrm{e}^{\lambda X} \geq \mathrm{e}^{\lambda t}) \leq \frac{\mathbb{E}[\mathrm{e}^{\lambda X}]}{\mathrm{e}^{\lambda t}}\,.$$

L'idée de la méthode de Chernoff, est de choisir le nombre $\lambda > 0$ qui minimise le membre de droite de l'inégalité précédente.

Théorème IV-1.5 (Chernoff). *Soit X une variable aléatoire réelle vérifiant, pour tout $\lambda \in \mathbb{R}$:*

$$\phi_X(\lambda) = \mathbb{E}[\mathrm{e}^{\lambda X}] < \infty\,. \tag{IV-1.7}$$

On a, pour tout $a \geq 0$,

$$\mathbb{P}(X \geq a) \leq \mathrm{e}^{-h_X(a)} \qquad \textit{pour tout } a \geq 0 \tag{IV-1.8}$$

et

$$\mathbb{P}(X \leq -a) \leq \mathrm{e}^{-h_X(-a)} \qquad \textit{pour tout } a \geq 0 \tag{IV-1.9}$$

où la fonction h_X est la transformée de Crámer de X, donnée par :

$$h_X(a) = \begin{cases} \sup_{\lambda \geq 0}\{a\lambda - \log\phi_X(\lambda)\} & \textit{si } a \geq 0 \\ \sup_{\lambda \leq 0}\{-a\lambda - \log\phi_X(\lambda)\} & \textit{si } a \leq 0. \end{cases} \tag{IV-1.10}$$

Démonstration. Pour $\lambda \geq 0$ et $a \geq 0$, la majoration $\mathrm{e}^{\lambda(X-a)} \geq \mathbb{1}_{\{X \geq a\}}$ implique

$$\mathbb{P}(X \geq a) \leq \mathbb{E}\left[\mathrm{e}^{\lambda(X-a)}\right] \leq \mathbb{E}[\mathrm{e}^{\lambda X}]\mathrm{e}^{-a\lambda} = \mathrm{e}^{-(a\lambda - \psi_X(\lambda))}\,.$$

où $\psi_X(\lambda) = \log\phi_X(\lambda)$. L'Eq (IV-1.8) découle de l'optimisation de cette inégalité en $\lambda > 0$.

$$\mathbb{P}(X \geq a) \leq \mathrm{e}^{-\sup_{\lambda \geq 0}(a\lambda - \psi_X(\lambda))} = \mathrm{e}^{-h_X(a)}\,.$$

La preuve de (IV-1.9) est analogue. Pour $a \geq 0$ et $\lambda \leq 0$, nous avons

$$\mathbb{P}(X \leq -a) = \mathbb{E}[\mathbb{1}_{\{X \leq -a\}}] \leq \mathbb{E}[\mathrm{e}^{\lambda(X+a)}] = \mathrm{e}^{a\lambda + \psi_X(\lambda)} = \mathrm{e}^{-(-a\lambda - \psi_X(\lambda))}\,,$$

et on conclut de la même manière. □

Si $X_1,\dots,X_n$ sont des variables aléatoires indépendantes vérifiant $\mathbb{E}[e^{\lambda X_i}] < \infty$ pour tout $\lambda \in \mathbb{R}$, alors il est facile de relier les transformées de Cramér des variables $X_1,\dots,X_n$ à la transformée de Cramér de la moyenne empirique $\bar{X}_n$.

Théorème IV-1.6. *Soit $X, X_1,\dots,X_n$ n variables aléatoires réelles i.i.d. . On suppose que, pour tout $\lambda \in \mathbb{R}$, $\phi_X(\lambda) = \mathbb{E}[e^{\lambda X}] < \infty$ et on note $h_X(a)$ pour $a \in \mathbb{R}$, la transformée de Cramér de X. On note $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ la moyenne empirique de l'échantillon. Alors pour tout $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \bar{X}_n}] < \infty$ et la transformée de Cramér de $\bar{X}_n$, notée $h_{\bar{X}_n}(a)$, vérifie, pour tout $a \in \mathbb{R}$:*

$$h_{\bar{X}_n}(a) = n h_X(a)$$

Démonstration. On note $\phi_{\bar{X}_n}(\lambda)$ la fonction génératrice de $\bar{X}_n$, alors en utilisant l'hypothèse d'indépendance

$$\phi_{\bar{X}_n}(\lambda) = \mathbb{E}\left[e^{\lambda n^{-1}\Sigma_{i=1}^n X_i}\right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i/n}] = \{\phi_X(\lambda/n)\}^n \ . \tag{IV-1.11}$$

et donc $\psi_{\bar{X}_n}(\lambda) = n\psi_X(\lambda/n)$ où $\psi_{\bar{X}_n}(\lambda) = \log \phi_{\bar{X}_n}(\lambda)$ et $\psi_X(\lambda) = \log \phi_X(\lambda)$. Pour $a \geq 0$

$$h_{\bar{X}_n}(a) = \sup_{\lambda \geq 0}(a\lambda - n\psi_X(\lambda/n)) = n \sup_{\lambda \geq 0}(a\lambda/n - \psi_X(\lambda/n)) = n h_X(a) \ .$$

La preuve pour $a \leq 0$ est identique. □

Corollaire IV-1.7. *Sous les conditions du théorème IV-1.6, nous avons pour tout $a > 0$.*

$$\mathbb{P}(\bar{X}_n \geq a) \leq e^{-nh_X(a)} \ ,$$
$$\mathbb{P}(\bar{X}_n \leq -a) \leq e^{-nh_X(-a)} \ .$$

IV-1.4 Inégalité de Hoeffding

L'obtention d'une inégalité plus précise que Bienayme-Tchebychev revient à trouver une "bonne" borne de la fonction génératrice des moments $s \mapsto \mathbb{E}[e^{s(X_i - \mathbb{E}[X_i])}]$ (la transformée de Laplace de la distribution de la variable aléatoire $X_i - \mathbb{E}[X_i]$). Cet objectif peut être atteint de différentes façons, qui dépendent des hypothèses dont on dispose sur la distribution de la variable aléatoire X_i.

Si la variable aléatoire X_i est presque-sûrement bornée, Hoeffding a établi une borne de la fonction génératrice des moments qui ne dépend que du support de la loi de X. [**?**, **?**] :

Lemme IV-1.8. *Soit Y est une variable aléatoire réelle de loi $\mathbb{Q}$, où $\mathbb{Q}([a,b]) = 1$ et $\int y\mathbb{Q}(dy) = 0$. Posons*

$$\psi_Y(\lambda) = \log \int \exp(\lambda y)\mathbb{Q}(dy) \qquad \lambda > 0 \ .$$

Nous avons

$$\psi_Y(\lambda) \leq \lambda^2 (b-a)^2/8 \ .$$

Démonstration. La convexité de la fonction exponentielle implique que pour tout $a \leq x \leq b$,

$$e^{\lambda x} \leq \frac{x-a}{b-a} e^{\lambda b} + \frac{b-x}{b-a} e^{\lambda a}.$$

En utilisant $\mathbb{E}[Y]=0$, et en introduisant la notation $p=-a/(b-a)$, nous obtenons

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\lambda Y}] &\leq \frac{b}{b-a}\,\mathrm{e}^{\lambda a}-\frac{a}{b-a}\,\mathrm{e}^{\lambda b}\\ &= (1-p+p\mathrm{e}^{\lambda(b-a)})\mathrm{e}^{-p\lambda(b-a)} := \mathrm{e}^{\rho(u)},\end{aligned}$$

où $u=\lambda(b-a)$ et $\rho(u)=-pu+\log(1-p+p\mathrm{e}^{u})$. Un calcul élémentaire montre que la dérivée de ρ est donnée par

$$\rho'(u)=-p+\frac{p}{p+(1-p)\mathrm{e}^{-u}}$$

et donc que $\rho(0)=\rho'(0)=0$. En notant que, pour tous réels a et b, $4ab\leqslant(a+b)^2$ (remarquons en effet que $2ab\leq a^2+b^2$),

$$\rho''(u)=\frac{p(1-p)\mathrm{e}^{-u}}{(p+(1-p)\mathrm{e}^{-u})^2}\leq\frac{1}{4}.$$

En utilisant le théorème de Taylor-Lagrange, nous obtenons qu'il existe $\theta\in[0,u]$ tel que

$$\rho(u)=\rho(0)+u\rho'(0)+\frac{u^2}{2}\rho''(\theta)\leq\frac{u^2}{8}=\frac{\lambda^2(b-a)^2}{8}.$$

Nous donnons maintenant une autre preuve plus probabiliste. La fonction $\lambda\to\psi_Y(\lambda)$ est deux fois dérivable et,

$$\psi_Y''(\lambda)=\mathrm{e}^{-\psi_Y(\lambda)}\int y^2\exp(\lambda y)\,\mathbb{Q}(\mathrm{d}y)-\mathrm{e}^{-2\psi_Y(\lambda)}\left(\int y\exp(\lambda y)\,\mathbb{Q}(\mathrm{d}y)\right)^2. \tag{IV-1.12}$$

Posons,

$$\mathbb{Q}_\lambda=\mathrm{e}^{-\psi_Y(\lambda)}\mathrm{e}^{\lambda y}\cdot\mathbb{Q}\,.$$

Par construction, $\mathbb{Q}_\lambda$ est une mesure de probabilité sur $\mathscr{B}(\mathbb{R})$. A l'aide de cette probabilité $\mathbb{Q}_\lambda$, on peut interpréter (IV-1.12) de la manière suivante :

$$\psi_Y''(\lambda)=\int\left(z-\int z\mathbb{Q}_\lambda(\mathrm{d}z)\right)^2\mathbb{Q}_\lambda(\mathrm{d}z)=\mathrm{Var}_\lambda(Z)\,,$$

où Z est une variable aléatoire à valeurs dans $[a,b]$ de loi $\mathbb{Q}_\lambda$. Maintenant, pour toute variable aléatoire Z sur l'intervalle $[a,b]$, on a toujours

$$\left|Z-\frac{b+a}{2}\right|\leq\frac{b-a}{2}\,,$$

et donc

$$\mathrm{Var}_\lambda(Z)=\mathrm{Var}_\lambda\left(Z-(b+a)/2\right)\leq\mathbb{E}_\lambda\left[(Z-(b+a)/2)^2\right]\leq\frac{(b-a)^2}{4}\,,$$

ce qui implique

$$\psi_Y''(\lambda)\leq(b-a)^2/4\,. \tag{IV-1.13}$$

En intégrant (IV-1.13) et en utilisant $\psi_Y(0)=\psi_Y'(0)=0$, on déduit

$$\psi_Y(\lambda)\leq\lambda^2\frac{(b-a)^2}{8}\,. \tag{IV-1.14}$$

□

Nous pouvons maintenant utiliser directement ce résultat dans la borne obtenue par la méthode de Chernoff (voir Théorème IV-1.6) dès que les variables aléatoires X_i sont presque-sûre bornées. On obtient ainsi l'*inégalité de Hoeffding* : elle joue un rôle important dans les développements théoriques de l'apprentissage statistique.

Théorème IV-1.9 (Inégalité de Hoeffding). *Soient $X_1,\dots,X_n$ des variables aléatoires réelles indépendantes telles que, pour tout $i\in\{1,\dots,n\}$, $\mathbb{E}[X_i]=0$ et $a_i\le X_i\le b_i$. Pour tout $t>0$, nous avons*

$$\mathbb{P}\left(\bar{X}_n\ge t\right)\le \mathrm{e}^{-2n^2t^2/\sum_{i=1}^n (b_i-a_i)^2}\,,$$
$$\mathbb{P}\left(\bar{X}_n\le -t\right)\le \mathrm{e}^{-2n^2t^2/\sum_{i=1}^n (b_i-a_i)^2}\,.$$

Démonstration. En utilisant le théorème IV-1.6, nous avons pour tout $t>0$,

$$\mathbb{P}(\bar{X}_n\ge t)\le \mathrm{e}^{-h_{\bar{X}_n}(t)} \quad \text{où } h_{\bar{X}_n}(t)=\sup_{\lambda\ge 0}\{t\lambda-\sum_{i=1}^n \psi_{X_i}(\lambda/n)\}\,.$$

Le Lemme IV-1.8 implique que

$$\begin{aligned} h_{\bar{X}_n}(t) &= \sup_{\lambda\ge 0}\left\{t\lambda-\left(\frac{\lambda}{n}\right)^2\frac{1}{8}\sum_{i=1}^n (b_i-a_i)^2\right\}\\ &= -2n^2t^2/\sum_{i=1}^n (b_i-a_i)^2 \end{aligned}$$

□

Corollaire IV-1.10. *Si $X_1,\dots,X_n$ sont des variables aléatoires de Bernoulli de paramètre p et si $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, alors, pour tout $t>0$,*

$$\mathbb{P}\left(|\overline{X}_n-p|\ge t\right)\le 2\exp(-2nt^2)\,.$$

Démonstration. Appliquons l'inégalité de Hoeffding à $Y_i=X_i-p$. Les conditions du théorème IV-1.9 sont vérifiées avec $b_i-a_i=1$. On conclut en écrivant

$$\mathbb{P}\left(|\bar{X}_n-p|\ge t\right)=\mathbb{P}\left(\bar{X}_n-p\ge t\right)+\mathbb{P}\left(\bar{X}_n-p\le -t\right)\,.$$

IV-1.5 Inégalité de Pisier

Dans la preuve de Théorème IV-1.9, l'élément important n'est pas tant que les variables aléatoires soient bornées presque-sûrement, mais plutôt que le logarithme de la fonction génératrice des moments soit borné par une fonction quadratique.

Nous allons maintenant formaliser cette propriété en introduisant la notion de loi sous-gaussienne. Il y a plusieurs façons de le faire et nous vous proposons la définition suivante, basée sur la fonction génératrice des moments de X.

Définition IV-1.11 (Loi sous-gaussienne, variable sous-gaussienne). *Soit $\mathbb{Q}$ une probabilité sur $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ vérifiant $\int_{-\infty}^{\infty} x\mathbb{Q}(\mathrm{d}x)=0$. La probabilité $\mathbb{Q}$ est dite sous-gaussienne de facteur de variance σ^2 si pour tout $\lambda\in\mathbb{R}$,*

$$\psi(\lambda)=\log\int \mathrm{e}^{\lambda x}\mathbb{Q}(\mathrm{d}x)\le \lambda^2\sigma^2/2\,.$$

Nous notons $\mathscr{SG}(\sigma^2)$ ces lois. Nous dirons qu'une variable aléatoire Y est sous-gaussienne de facteur de variance σ^2 si la loi de $Y-\mathbb{E}[Y]$ est un élément de $\mathscr{SG}(\sigma^2)$.

Cette définition est naturelle parce que nous savons que la fonction génératrice des moments d'une variable gaussienne centrée de variance σ^2 est égale à $e^{\sigma^2\lambda^2/2}$. Cette définition est naturelle car elle est stable par convolution. Si $X_1,\dots,X_n$ sont des variables aléatoires indépendantes de loi sous-gaussienne de facteurs de variance $\sigma_1^2,\dots,\sigma_n^2$, alors la loi de $\sum_{i=1}^n X_i$ est elle aussi sous-gaussienne de facteur de variance $\sum_{i=1}^n \sigma_i^2$.

Remarque IV-1.12. Notons que la variance d'une variable aléatoire de loi sous-gaussienne de facteur de variance σ^2 n'est pas nécessairement égale à σ^2. On peut toutefois montrer que $\int x^2\mu(\mathrm{d}x) \le \sigma^2$. ♢

Théorème IV-1.13. *Soit $\sigma > 0$, $n \ge 2$, et $\{Y_i\}_{i=1}^n$ n variables aléatoires réelles sous-gaussiennes de facteur de variance σ^2. Alors,*

$$\mathbb{E}\left[\max_{1\le i\le n} Y_i\right] \le \sigma\sqrt{2\ln n}. \tag{IV-1.15}$$

Si, de plus, $\mathbb{E}[e^{s(-Y_i)}] \le e^{s^2\sigma^2/2}$ pour tout $s > 0$ et $i \in \{1,\dots,n\}$, alors pour tout $n \ge 1$,

$$\mathbb{E}\left[\max_{1\le i\le n} |Y_i|\right] \le \sigma\sqrt{2\ln(2n)}. \tag{IV-1.16}$$

Démonstration. En utilisant l'inégalité de Jensen, pour tout $s > 0$, nous avons

$$e^{s\mathbb{E}[\max_{1\le i\le n} Y_i]} \le \mathbb{E}\left[e^{s\max_{1\le i\le n} Y_i}\right] = \mathbb{E}\left[\max_{1\le i\le n} e^{sY_i}\right] \le \sum_{i=1}^n \mathbb{E}\left[e^{sY_i}\right] \le n e^{s^2\sigma^2/2}.$$

Par conséquent, en utilisant de nouveau l'inégalité de Jensen,

$$\mathbb{E}\left[\max_{1\le i\le n} Y_i\right] \le \frac{\ln n}{s} + \frac{s\sigma^2}{2},$$

et nous obtenons (IV-1.15) en posant $s = \sqrt{2\ln n}/\sigma$.

Notons que

$$\max_{i\le n} |Y_i| = \max(Y_1, -Y_1, \dots, Y_n, -Y_n)\ .$$

En appliquant (IV-1.15) nous obtenons (IV-1.16). □

Chapitre IV-2

Fonction de répartition, quantiles et statistiques d'ordre

IV-2.1 Fonction de répartition

Définition IV-2.1 (Fonction de répartition). *Soit μ une mesure de probabilité sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. La fonction de répartition de la mesure de probabilité μ est la fonction $F : x \mapsto \mu(]-\infty, x])$.*

Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et X une variable aléatoire à valeurs réelles. La fonction de répartition *F de la variable X est la fonction de répartition de la mesure image de $\mathbb{P}$ par X, i.e. la fonction $F : x \mapsto \mathbb{P}(X \leq x)$.*

Proposition IV-2.2. *Soit $F : \mathbb{R} \to [0,1]$ la fonction de répartition d'une mesure de probabilité μ sur $\mathbb{R}$. Alors :*

(i) $0 \leq F(x) \leq 1$ *pour tout* $x \in \mathbb{R}$.

(ii) F *est croissante.*

(iii) F *est continue à droite, pour tout* $a \in \mathbb{R}$, $\lim_{x \downarrow a} F(x) = F(a)$.

(iv) En tout point $a \in \mathbb{R}$, F *admet une limite à gauche réelle, égale* $\mu(]-\infty, a[$.

(v) $\lim_{x \to -\infty} F(x) = 0$ *et* $\lim_{x \to \infty} F(x) = 1$.

Démonstration. (i) Par définition d'une probabilité, $0 \leq \mu(A) \leq 1$ pour tout $A \in \mathscr{B}(\mathbb{R})$.

(ii) Soient a et b deux réels tels que $a \leq b$. Nous avons

$$]-\infty, b] =]-\infty, a] \cup]a, b]$$

Par conséquent, $F(b) = F(a) + \mu(]a, b]) \geq F(a)$.

(iii) Soit $a \in \mathbb{R}$. Pour toute suite $\{h_n,\ n \in \mathbb{N}\}$ décroissante et positive, on a $]-\infty, a] = \bigcap_n]-\infty, a + h_n]$; les ensembles $]-\infty, a + h_n]$ étant décroissants, il vient

$$\mu\left(\bigcap_n]-\infty, a + h_n]\right) = \lim \downarrow \mu\left(]-\infty, a + h_n]\right)$$

dont on déduit que $F(a) = \lim \downarrow F(a + h_n)$. Par suite, F est continue à droite en a.

(iv) Soit $a \in \mathbb{R}$. Pour toute suite $\{h_n,\, n \in \mathbb{N}\}$ décroissante de réels positifs, on écrit $]-\infty, a[= \bigcup_n]-\infty, a-h_n]$. On en déduit que $\lim_{x \underset{<}{\to} a} F(x)$ existe et vaut $\mu(]-\infty, a[)$.

(v) Soit $\{x_n,\, n \in \mathbb{N}\}$ une suite décroissante de réels telle que $\lim_{n\to\infty} x_n = -\infty$ et posons $A_n :=]-\infty, x_n]$. La suite $\{A_n,\, n \in \mathbb{N}\}$ est une suite décroissante et $\bigcap_{n=0}^{\infty} A_n = \emptyset$. La Proposition A.26-(v) montre que

$$\lim_{x\to-\infty} F(x) = \lim_{n\to\infty} F(x_n) = \lim \downarrow \mu\left(]-\infty, x_n]\right) = \mu(\emptyset) = 0\,.$$

On prouve de même que $\lim_{x\to+\infty} F(x) = 1$. □

Définition IV-2.3 (Fonction de répartition empirique). *La* fonction de répartition empirique *de l'échantillon $(X_1, \ldots, X_n)$, notée $\widehat{F}_n$, est définie par*

$$x \mapsto \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{]-\infty, x]}(X_k)\,.$$

La fonction de répartition empirique $\widehat{F}_n$ est la fonction de répartition de la mesure empirique

$$n^{-1} \sum_{i=1}^{n} \delta_{X_i}\,.$$

La proposition suivante montre que $\widehat{F}_n$ est un estimateur (fortement consistant) de la fonction de répartition de la loi de X_1 ; et elle établit quelques propriétés de cet estimateur.

Théorème IV-2.4. *Soit $\{X_k,\, k \in \mathbb{N}^*\}$ une suite de variables aléatoires réelles i.i.d. définies sur $(\Omega, \mathscr{F}, \mathbb{P})$ et de fonction de répartition F. Soit $\widehat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1, \cdots, X_n)$.*

Pour tout $x_0 \in \mathbb{R}$ et $n \in \mathbb{N}^*$ on a

$$\mathbb{E}\left[\widehat{F}_n(x_0)\right] = F(x_0)\,, \tag{IV-2.1}$$

$$\mathrm{Var}\left(\widehat{F}_n(x_0)\right) = n^{-1}\, F(x_0)\big(1 - F(x_0)\big)\,, \tag{IV-2.2}$$

$$\mathbb{P}\left(\left|\widehat{F}_n(x_0) - F(x_0)\right| \geq \varepsilon\right) \leq 2\mathrm{e}^{-2n\varepsilon^2}\,, \quad \textit{pour tout } \varepsilon > 0. \tag{IV-2.3}$$

De plus, nous avons pour tout $x_0 \in \mathbb{R}$,

$$\widehat{F}_n(x_0) \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} F(x_0)\,, \tag{IV-2.4}$$

et

$$\sqrt{n}(\widehat{F}_n(x_0) - F(x_0)) \Longrightarrow \mathrm{N}(0, F(x_0)(1 - F(x_0))). \tag{IV-2.5}$$

Démonstration. Les variables aléatoires $\{\mathbb{1}_{]-\infty, x_0]}(X_k), k \geq 1\}$ sont indépendantes, de loi de Bernoulli de paramètre $F(x_0)$. Nous avons donc

$$\mathbb{E}[\hat{F}_n(x_0)] = n^{-1} \sum_{k=1}^{n} F(x_0) = F(x_0)$$

$$\mathrm{Var}(\hat{F}_n(x_0)) = n^{-2} \sum_{k=1}^{n} F(x_0)(1 - F(x_0)) = n^{-1} F(x_0)(1 - F(x_0))\,.$$

La preuve de (IV-2.3) découle de l'inégalité d'Hoeffding (Théorème IV-1.9). □

Un raffinement de la preuve montre que la convergence (IV-2.4) de la fonction de répartition empirique est en fait uniforme en $x \in \mathbb{R}$.

Théorème IV-2.5 (Glivenko-Cantelli). *Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de variables aléatoires réelles i.i.d. définies sur $(\Omega, \mathscr{F}, \mathbb{P})$, de fonction de répartition F. Soit $\widehat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1, \cdots, X_n)$. Alors,*

$$\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} 0\ . \tag{IV-2.6}$$

$$\lim_{n\to\infty} \mathbb{E}\left[\left\{\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)|\right\}^2\right] = 0\ . \tag{IV-2.7}$$

Démonstration. Nous allons nous ramener à la loi forte des grands nombres, en discrétisant le problème. Pour $m \in \mathbb{N}^*$ construisons $a_0, a_1, \ldots, a_m, a_{m+1}$ de telle sorte que

$$[a_i,\ a_{i+1}[= \left\{x \in \mathbb{R} :\ \frac{i}{m} \le F(x) < \frac{i+1}{m}\right\}\ , \quad \text{pour } i = 1, 2, \ldots, m-1,$$

et $a_0 = -\infty$,$a_{m+1} = +\infty$. Pour $i \in \{0, \ldots, m\}$ et $x \in [a_i, a_{i+1}[$, nous avons

$$\widehat{F}_n(x) - F(x) \le \widehat{F}_n(a_{i+1}-) - F(a_i) \le \widehat{F}_n(a_{i+1}-) - F(a_{i+1}-) + \frac{1}{m}\ ,$$
$$F(x) - \widehat{F}_n(x) \le F(a_i) - \widehat{F}_n(a_i) + \frac{1}{m}\ .$$

Par conséquent, comme $\mathbb{R} =]a_0, a_1[\cup (\bigcup_{i=0}^m [a_i, a_{i+1}[)$ et $\lim_{x\to a_0} \widehat{F}_n(x) = \lim_{x\to a_0} F(x) = 0$ et $\lim_{x\to a_{m+1}} \widehat{F}_n(x) = \lim_{x\to a_{m+1}} F(x) = 1$, nous avons

$$\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| = \max_{1\le i\le m} \left\{|F(a_i) - \widehat{F}_n(a_i)| + |F(a_i-) - \widehat{F}_n(a_i-)| + \frac{1}{m}\right\}\ , \tag{IV-2.8}$$

Pour tout m, lorsque $n \to \infty$, (IV-2.4) entraine

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \le \frac{1}{m}\ .$$

Comme m peut être choisi arbitrairement grand, ceci montre (IV-2.6). La preuve de (IV-2.7) découle du théorème de convergence dominé (Théorème A.40) en remarquant que $\sup_{x\in\mathbb{R}} |F(x) - \widehat{F}_n(x)| \le 2$. □

Il est possible d'obtenir un résultat beaucoup plus précis que celui donné par l'équation (IV-2.3) sur la déviation de la fonction de répartition empirique.

Théorème IV-2.6. *[Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW).] Soit $X_1, \ldots, X_n$ des variables i.i.d. définies sur $(\Omega, \mathscr{F}, \mathbb{P})$ et de fonction de répartition F. Soit $\widehat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1, \cdots, X_n)$. Pour tout $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| > \varepsilon\right) \le 2\mathrm{e}^{-2n\varepsilon^2}\ . \tag{IV-2.9}$$

On remarque que la borne est la même que celle qui apparaît dans la borne exponentielle (IV-2.3). La preuve de l'inégalité (IV-2.9) est délicate. En utilisant le fait que

$$\mathbb{E}[U] = \int_0^\infty \mathbb{P}(U > t)\mathrm{d}t$$

pour une variable aléatoire positive U, on déduit du Théorème IV-2.6 que

$$\begin{aligned}\left(\mathbb{E}\left[\sup_{x\in\mathbb{R}}|\widehat{F}_n(x)-F(x)|\right]\right)^2 &\leq \mathbb{E}\left[\sup_{x\in\mathbb{R}}|\widehat{F}_n(x)-F(x)|^2\right] \\ &\leq \int_0^\infty \mathbb{P}\left(\sup_{x\in\mathbb{R}}|\widehat{F}_n(x)-F(x)|^2 > t\right)\mathrm{d}t \leq \frac{1}{n}\end{aligned}$$

ce qui complète le résultat donné par le Théorème IV-2.5.

Démonstration. (du théorème IV-2.6) Nous allons montrer un résultat plus faible : pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}|\widehat{F}_n(x)-F(x)| > \varepsilon\right) \leq 4\left(\frac{2}{\varepsilon}+1\right)\mathrm{e}^{-n\frac{\varepsilon^2}{2}} \ .$$

Soit ε et N_ε la partie entière supérieure de $2/\varepsilon$. Pour tout $t \in \mathbb{R}$, on définit

$$\widehat{F}_n(t^-) := \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{]-\infty,t[}(X_k) \ .$$

Pour tout $k = 0, \cdots, N_\varepsilon$, notons

$$a_k := \inf\{t \in \mathbb{R} : \ F(t) \geq k/N_\varepsilon\} \ .$$

On a toujours $a_k \leq a_{k+1}$. Par continuité à droite de F et par définition des points a_j, on a

$$F(a_k) \geq \frac{k}{N_\varepsilon} \geq F(a_k^-) \ ,$$

et donc

$$F(a_{k+1}^-) - F(a_k) \leq \frac{1}{N_\varepsilon} \leq \frac{\varepsilon}{2}$$

ou de façon équivalente

$$F(a_k) + \frac{\varepsilon}{2} \geq F(a_{k+1}^-) \ .$$

Soit $t \in [a_k, a_{k+1}[$. Puisque $\widehat{F}_n$ et F sont croissantes,

$$\widehat{F}_n(a_k) - F(a_{k+1}^-) \leq \widehat{F}_n(t) - F(t) \leq \widehat{F}_n(a_{k+1}^-) - F(a_k) \ ,$$

et donc

$$\widehat{F}_n(a_k) - F(a_k) - \frac{\varepsilon}{2} \leq \widehat{F}_n(t) - F(t) \leq \widehat{F}_n(a_{k+1}^-) - F(a_{k+1}^-) + \frac{\varepsilon}{2} \ .$$

Ainsi,

$$\sup_{t\in\mathbb{R}}\left|\widehat{F}_n(t) - F(t)\right| \leq \max_{k=1,\ldots,N_\varepsilon}\left(\left|\widehat{F}_n(a_k) - F(a_k)\right| \vee \left|\widehat{F}_n(a_k^-) - F(a_k^-)\right|\right) + \frac{\varepsilon}{2} \ .$$

Or, par le théorème IV-2.4, pour tout $k = 1, \ldots, N_\varepsilon$, on a,

$$\mathbb{P}\left(\left|\widehat{F}_n(a_k) - F(a_k)\right| > \frac{\varepsilon}{2}\right) \leq 2\mathrm{e}^{-n\frac{\varepsilon^2}{2}} \ , \qquad \mathbb{P}\left(\left|\widehat{F}_n(a_k^-) - F(a_k^-)\right| > \frac{\varepsilon}{2}\right) \leq 2\mathrm{e}^{-n\frac{\varepsilon^2}{2}} \ .$$

On en déduit donc

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_n(t) - F(t)\right| > \varepsilon\right) &\leq \mathbb{P}\left(\bigcup_{k=1}^{N_\varepsilon}\left\{\left|\widehat{F}_n(a_k) - F(a_k)\right| > \frac{\varepsilon}{2}\right\} \cup \left\{\left|\widehat{F}_n(a_k^-) - F(a_k^-)\right| > \frac{\varepsilon}{2}\right\}\right) \\ &\leq \sum_{k=1}^{N_\varepsilon}\left(\mathbb{P}\left(\left|\widehat{F}_n(a_k) - F(a_k)\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\left|\widehat{F}_n(a_k^-) - F(a_k^-)\right| > \frac{\varepsilon}{2}\right)\right) \\ &\leq 4N_\varepsilon\mathrm{e}^{-n\frac{\varepsilon^2}{2}} \leq 4\left(\frac{2}{\varepsilon}+1\right)\mathrm{e}^{-n\frac{\varepsilon^2}{2}} \ .\end{aligned}$$

□

IV-2.2 Quantiles

La fonction F ne définit pas nécessairement une bijection de $\mathbb{R}$ sur $[0,1]$. On peut néanmoins définir une inverse généralisée : pour tout $p \in [0,1]$, posons

$$F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\} , \tag{IV-2.10}$$

où, par convention,

$$\inf \mathbb{R} = -\infty \quad \text{et} \quad \inf \emptyset = +\infty .$$

La proposition suivante établit les propriétés de cette inverse généralisée.

Proposition IV-2.7. *Soit F la fonction de répartition d'une mesure de probabilité sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. La fonction F^{-1} définie sur $]0,1[$ par* (IV-2.10) *vérifie les propriétés suivantes :*

(i) F^{-1} *est croissante sur* $[0,1]$.

(ii) $F^{-1}[F(x)] \leq x$ *pour tout* $x \in \mathbb{R}$ *tel que* $0 < F(x) < 1$.

(iii) $F[F^{-1}(p)] \geq p$ *pour tout* $p \in]0,1[$ *avec égalité si F est continue au point* $F^{-1}(p)$.

(iv) F^{-1} *est continue à gauche sur* $]0,1[$ *:* $F^{-1}(p-) = F^{-1}(p)$ *pour tout* $p \in]0,1[$.

(v) F^{-1} *admet des limites à droite sur* $]0,1[$ *:* $F^{-1}(p+) = \inf\{x \in \mathbb{R} : F(x) > p\}$.

Démonstration. (i) Si $0 \leq p \leq q \leq 1$, alors $\{x \in \mathbb{R} : F(x) \geq q\} \subseteq \{x \in \mathbb{R} : F(x) \geq p\}$. Donc $F^{-1}(p) \leq F^{-1}(q)$ et F^{-1} est croissante.

(ii) La définition de l'inverse généralisée montre que $F^{-1}[F(x)]$ est le plus petit $y \in \mathbb{R}$ tel que $F(y) \geq F(x)$.

(iii) Par définition, $F^{-1}(p)$ est une valeur de y telle que $F(y) \geq p$.

(iv) La fonction F est continue à droite

(v) La fonction F a des limites à gauche. □

Définition IV-2.8 (Quantiles). *Soit F la fonction de répartition d'une mesure de probabilité μ sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Pour $0 \leq p \leq 1$, le p-*ième quantile de *F est défini par*

$$Q(p) := F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\} , \tag{IV-2.11}$$

Si la fonction $x \mapsto F(x)$ est continue et strictement croissante, alors F définit une bijection de $\mathbb{R}$ sur $[0,1]$ et la fonction des quantiles $p \mapsto Q(p)$ est simplement l'inverse de la fonction F. On a dans ce cas

$$F \circ F^{-1}(p) = F(Q(p)) = p . \tag{IV-2.12}$$

Si la fonction F est strictement croissante mais est discontinue, la relation $F \circ F^{-1}(p) = p$ n'est pas nécessairement vérifiée.

Proposition IV-2.9. *Soit X une variable aléatoire réelle de fonction de répartition F. Si F est continue et strictement croissante, alors la variables aléatoires $F(X)$ suit une loi uniforme sur* $[0,1]$.

Démonstration. Soit $u \in]0,1[$. Par hypothèse, on peut appliquer l'équation (IV-2.12), donc

$$\mathbb{P}(F(X) \leq u) = \mathbb{P}(F(X) \leq F(F^{-1}(u))) = \mathbb{P}(X \leq F^{-1}(u)) = F \circ F^{-1}(u) = u. \qquad \square$$

Définition IV-2.10 (Quantile empirique). *Soient $(X_1,\dots,X_n)$ n variables aléatoires réelles définies sur $(\Omega,\mathscr{F},\mathbb{P})$. Le* p-ième quantile empirique *est le p-ième quantile de la fonction de répartition empirique $\widehat{F}_n : x \mapsto \widehat{F}_n(x) = n^{-1}\sum_{i=1}^n \mathbb{1}_{]-\infty,x]}(X_i)$, i.e.*

$$\widehat{Q}_n(p) := \widehat{F}_n^{-1}(p) = \inf\left\{x \in \mathbb{R} : \widehat{F}_n(x) \geq p\right\} .$$

Théorème IV-2.11. *Soit $(X_1,\dots,X_n)$ des variables aléatoires réelles i.i.d. de fonction de répartition F et de fonction quantile associée Q. Soit $\widehat{Q}_n$ la fonction quantile empirique associée à la fonction de répartition empirique de l'échantillon $(X_1,\dots,X_n)$.*

Pour tout $p \in]0,1[$ et $\varepsilon > 0$,

$$\mathbb{P}(|\widehat{Q}_n(p) - Q(p)| > \varepsilon) \leq 2\mathrm{e}^{-2n\delta_\varepsilon^2} ,$$

où

$$\delta_\varepsilon := \min\{F(Q(p)+\varepsilon) - p, p - F(Q(p)-\varepsilon)\} . \tag{IV-2.13}$$

Démonstration. Soit $\varepsilon > 0$. Nous avons

$$\mathbb{P}(|\widehat{Q}_n(p) - Q(p)| > \varepsilon) = \mathbb{P}(\widehat{Q}_n(p) > Q(p)+\varepsilon) + \mathbb{P}(\widehat{Q}_n(p) < Q(p)-\varepsilon) .$$

Nous allons borner ces deux termes séparément. Nous avons tout d'abord, en utilisant que $\widehat{Q}_n(p)$ est par définition, l'infimum des réels vérifiant $\widehat{F}_n(x) \geq p$,

$$\begin{aligned}
\mathbb{P}\left(\widehat{Q}_n(p) > Q(p)+\varepsilon\right) &= \mathbb{P}\left(p \geq \widehat{F}_n(Q(p)+\varepsilon)\right) = \mathbb{P}\left(np \geq \sum_{i=1}^n \mathbb{1}_{\{X_i \leq Q(p)+\varepsilon\}}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{\{X_i > Q(p)+\varepsilon\}} \geq n(1-p)\right) \\
&= \mathbb{P}\left(\sum_{i=1}^n V_i - n\mathbb{E}[V_1] \geq n(1-p) - n\{1 - F(Q(p)+\varepsilon)\}\right) ,
\end{aligned}$$

où l'on a posé $V_i := \mathbb{1}_{\{X_i > Q(p)+\varepsilon\}}$. Les variables aléatoires $V_1,\cdots,V_n$ sont i.i.d. de loi de Bernoulli de paramètre $1 - F(Q(p)+\varepsilon)$. Donc $\mathbb{E}[V_1] = 1 - F(Q(p)+\varepsilon)$. En notant $\delta_1 := F(Q(p)+\varepsilon) - p$, nous avons donc, en utilisant le théorème IV-1.9,

$$\mathbb{P}(\widehat{Q}_n(p) > Q(p)+\varepsilon) = \mathbb{P}_F\left(\sum_{i=1}^n \{V_i - \mathbb{E}[V_i]\} > n\delta_1\right) \leq \mathrm{e}^{-2n\delta_1^2} . \tag{IV-2.14}$$

En procédant de même, nous avons

$$\mathbb{P}\left(\widehat{Q}_n(p) < Q(p)-\varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^n \{W_i - \mathbb{E}[W_i]\} \geq n\delta_2\right) \leq \mathrm{e}^{-2n\delta_2^2} , \tag{IV-2.15}$$

où $W_i := \mathbb{1}_{\{X_i \leq Q(p)-\varepsilon\}}$ et $\delta_2 := p - F(Q(p)-\varepsilon)$. $\square$

Le théorème suivant établit la convergence presque-sûre du quantile empirique.

Théorème IV-2.12. *Soit $\{X_n,\, n\in\mathbb{N}\}$ une suite de variables aléatoires réelles i.i.d. définies sur $(\Omega,\mathscr{F},\mathbb{P})$, de fonction de répartition F et de fonction quantile associée Q. Soit $\widehat{Q}_n$ la fonction quantile empirique associée à la fonction de répartition empirique de l'échantillon $(X_1,\cdots,X_n)$. Soit $p\in]0,1[$. Supposons que pour tout $\varepsilon>0$, $F(Q(p)+\varepsilon)>p$. Alors,*

$$\widehat{Q}_n(p) \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} Q(p)\,.$$

Démonstration. D'après le théorème IV-2.11, pour tout $\varepsilon>0$, il existe $\delta_\varepsilon>0$ tel que

$$\sum_{n=1}^{\infty}\mathbb{P}(|\widehat{Q}_n(p)-Q(p)|>\varepsilon)\le 2\sum_{n=1}^{\infty}\mathrm{e}^{-2n\delta_\varepsilon^2}<\infty\,.$$

Par conséquent, le Lemme de Borel-Cantelli (Lemme IV-5.11) montre que $\widehat{Q}_n(p)-Q(p)\overset{\mathbb{P}-\text{p.s.}}{\longrightarrow}0$ lorsque $n\to\infty$. □

Proposition IV-2.13. *Soit $(X_1,\cdots,X_n)$ des variables aléatoires réelles i.i.d. définies sur le même espace $(\Omega,\mathscr{F},\mathbb{P})$, et de fonction de répartition F. Soit $\widehat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1,\cdots,X_n)$. Pour tout $\varepsilon>0$,*

$$\mathbb{P}\left(\sup_{p\in]0,1[}\left|F(\widehat{F}_n^{-1}(p))-p\right|>\varepsilon+\frac{1}{n}\right)\le 2\mathrm{e}^{-2n\varepsilon^2}\quad.$$

Démonstration. Soit $p\in]0,1[$. On écrit

$$F(\widehat{F}_n^{-1}(p))-p=\widehat{F}_n(\widehat{F}_n^{-1}(p))-p+F(\widehat{F}_n^{-1}(p))-\widehat{F}_n(\widehat{F}_n^{-1}(p))\,,$$

dont on déduit la majoration

$$\left|F(\widehat{F}_n^{-1}(p))-p\right|\le\left|\widehat{F}_n(\widehat{F}_n^{-1}(p))-p\right|+\sup_{x\in\mathbb{R}}\left|F(x)-\widehat{F}_n(x)\right|\,.$$

Notons K_n la partie entière supérieure de np. On a par définition de la fonction de répartition empirique et de la fonction quantile empirique $\widehat{F}_n(\widehat{F}_n^{-1}(p))=K_n/n$, donc $p\le\widehat{F}_n(\widehat{F}_n^{-1}(p))\le p+1/n$. Par suite,

$$\left|\widehat{F}_n(\widehat{F}_n^{-1}(p))-p\right|\le\frac{1}{n}.$$

On a finalement

$$\sup_{p\in]0,1[}\left|F(\widehat{F}_n^{-1}(p))-p\right|\le\frac{1}{n}+\sup_{x\in\mathbb{R}}\left|F(x)-\widehat{F}_n(x)\right|\quad.$$

On conclut avec l'inégalité de Dvoretsky-Kiefer-Wolfowitz (Théorème IV-2.6). □

IV-2.3 Statistiques d'ordre

Définition IV-2.14 (Statistiques d'ordre). *Soient $X_1, X_2, \ldots, X_n$ n variables aléatoires réelles. Les valeurs ordonnées $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ de $X_1, X_2, \cdots, X_n$ sont appelées* les statistiques d'ordre.

Le minimum $X_{1:n}$ de $X_1, X_2, \ldots, X_n$ est la 1-ère statistique d'ordre, le maximum $X_{n:n}$ est la n-ème statistique d'ordre. Notons que, pour tout $i \in \{1, \ldots, n\}$,

$$X_{i:n} = \widehat{F}_n^{-1}(i/n) \; . \tag{IV-2.16}$$

On appelle *médiane* de l'échantillon $(X_1, \cdots, X_n)$, le quantile empirique d'ordre $1/2$; on le note M_n. On a

$$M_n := \widehat{F}_n^{-1}(1/2) = \begin{cases} X_{m+1:n} & \text{si } n = 2m+1 \\ X_{m:n} & \text{si } n = 2m \end{cases} \; .$$

Dans le cas où $n = 2m$, on trouve aussi parfois la définition symétrique suivante

$$M_{2m} = \frac{1}{2} \left(X_{m:2m} + X_{m+1:2m} \right) \; .$$

Théorème IV-2.15 (Loi jointe des statistiques d'ordre). *Soient $X_1, X_2, \ldots, X_n$ des v.a. réelles i.i.d. définies sur $(\Omega, \mathscr{F}, \mathbb{P})$ et de loi admettant une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$. La loi jointe des statistiques d'ordre $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, donnée par*

$$(y_1, \, y_2, \, \ldots, \, y_n) \mapsto n! f(y_1) f(y_2) \cdots f(y_n) \mathbb{1}_{\{y_1 < y_2 < \cdots < y_n\}} \; .$$

Démonstration. Comme les variables $(X_1, \ldots, X_n)$ sont i.i.d. et que leur loi possède une densité, nous avons $\mathbb{P}(X_i = X_j) = 0$ pour tout $i \neq j$. Par suite,

$$\mathbb{P}(X_{1:n} < X_{2:n} < \cdots < X_{n:n}) = 1 \; .$$

Nous allons tout d'abord montrer que la loi jointe des statistiques d'ordre admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$. Soit $B \in \mathscr{B}(\mathbb{R}^n)$ un ensemble négligeable sous la mesure de Lebesgue. Nous avons

$$\begin{aligned} \mathbb{P}((X_{1:n}, \ldots, X_{n:n}) \in B) &= \mathbb{P}((X_{1:n}, \ldots, X_{n:n}) \in B, X_{1:n} < X_{2:n} < \cdots < X_{n:n}) \\ &= \sum_{\pi} \mathbb{P}((X_{\pi(1)}, \ldots, X_{\pi(n)}) \in B, X_{\pi(1)} < X_{\pi(2)} < \cdots < X_{\pi(n)}) \end{aligned}$$

où la somme porte sur l'ensemble des $n!$ permutations π de l'ensemble $\{1, \ldots, n\}$. Comme pour tout π, $\mathbb{P}((X_{\pi(1)}, \ldots, X_{\pi(n)}) \in B, X_{\pi(1)} < X_{\pi(2)} < \cdots < X_{\pi(n)}) = 0$ puisque B est négligeable et que $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ possède une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, nous avons $\mathbb{P}((X_{1:n}, \ldots, X_{n:n}) \in B) = 0$. Ainsi, la loi jointe des statistiques d'ordre est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ (voir définition A.46) et par le théorème de Radon-Nikodyn (voir théorème A.47), elle admet une densité.

Soit $-\infty \leq x_1, \ldots, x_n \leq \infty$. L'événement $A := \{X_{1:n} \leq x_1, X_{2:n} \leq x_2, \ldots, X_{n:n} \leq x_n\}$ est l'union des $n!$ événements disjoints

$$A_\pi := \{X_{\pi(1)} \leq x_1, X_{\pi(2)} \leq x_2, \ldots, X_{\pi(n)} \leq x_n\} \cap \{X_{\pi(1)} < X_{\pi(2)} < \cdots < X_{\pi(n)}\}$$

où π parcourt l'ensemble des permutations de $\{1, \ldots, n\}$. Notons de plus que

$$\mathbb{P}(A_\pi) = \int \cdots \int_{y_1 < \cdots < y_n} \prod_{i=1}^{n} \mathbb{1}_{\{y_i \leq x_i\}} f(y_i) \mathrm{d}y_i \; ,$$

qui ne dépend pas du choix de π. Par conséquent, nous avons $\mathbb{P}(A) = n! \mathbb{P}(A_\pi)$. On en déduit l'expression de la densité. □

On obtient facilement la loi marginale des statistiques d'ordre.

Corollaire IV-2.16. *Soit $X_1,\dots,X_n$ des v.a. réelles i.i.d. définies sur $(\Omega,\mathscr{F},\mathbb{P})$, et de fonction de répartition F. Supposons que la loi de X_1 admet une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Alors la loi de la r-ième statistique d'ordre $X_{r:n}$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par*

$$x \mapsto \frac{n!}{(r-1)!(n-r)!} F^{r-1}(x) f(x) \{1-F(x)\}^{n-r} . \tag{IV-2.17}$$

Démonstration. Soit $1 \le r \le n$. Calculons la r-ième marginale de la loi jointe des statistiques d'ordre. D'après le théorème IV-2.15, nous avons

$$\begin{aligned}\mathbb{P}(X_{r:n} \le x_\star) &= n! \int_{-\infty}^{x_\star} \left(\int \mathbb{1}_{y_1<\dots<y_r} \prod_{i=1}^{r-1} f(y_i)\mathrm{d}y_i \right) \left(\int \mathbb{1}_{y_r<\dots<y_n} \prod_{i=r+1}^{n} f(y_i)\mathrm{d}y_i \right) f(y_r)\mathrm{d}y_r , \\ &= n! \int_{-\infty}^{x_\star} I_r(y_r) J_r(y_r)\, f(y_r)\mathrm{d}y_r\end{aligned}$$

en ayant posé

$$\begin{aligned} I_r(x) &:= \int \mathbb{1}_{\{y_1<\dots<y_{r-1}<x\}} \left(\prod_{i=1}^{r-1} f(y_i) \right) \mathrm{d}y_1 \dots \mathrm{d}y_{r-1} , \\ J_r(x) &:= \int \mathbb{1}_{\{x<y_{r+1}<\dots<y_n\}} \left(\prod_{i=r+1}^{n} f(y_i) \right) \mathrm{d}y_{r+1} \dots \mathrm{d}y_n , \end{aligned}$$

avec la convention $I_1(x) = J_n(x) = 1$. Notons que pour $r \ge 2$ et $\ell \le n-1$, nous avons

$$I_r(x) = \int_{-\infty}^{x} f(x_{r-1}) I_{r-1}(x_{r-1}) \mathrm{d}x_{r-1} , \qquad J_\ell(x) = \int_x^{+\infty} J_{\ell+1}(y_{\ell+1})\, f(y_{\ell+1})\, \mathrm{d}y_{\ell+1} .$$

On montre par une récurrence élémentaire que pour $r \ge 1$ et $\ell \le n$

$$I_r(x) = \frac{F^{r-1}(x)}{(r-1)!}, \qquad J_\ell(x) = \frac{(1-F(x))^{n-\ell}}{(n-\ell)!} ,$$

ce qui conclut la preuve. □

Nous terminons en établissant la distribution asymptotique des quantiles centraux.

Théorème IV-2.17. *Soit $\{X_n,\, n \in \mathbb{N}\}$ une suite de v.a. réelles i.i.d. définies sur le même espace $(\Omega,\mathscr{F},\mathbb{P})$, de fonction de répartition F et de fonction quantile associée Q. Soit $p \in]0,1[$ et $\{k_n,\, n \in \mathbb{N}\}$ une suite d'entiers tels que*

$$\sqrt{n}(k_n/n - p) \to 0 . \tag{IV-2.18}$$

Si F est dérivable au point $Q(p)$ et $F'(Q(p)) > 0$, alors

$$\sqrt{n}(X_{k_n:n} - Q(p)) \Longrightarrow \mathrm{N}\left(0,\ \frac{p(1-p)}{[F'(Q(p))]^2}\right) .$$

Démonstration. Soit $p \in]0,1[$. Par hypothèse, la fonction F est continue au point $Q(p)$ et donc

$$F(Q(p)) = p . \tag{IV-2.19}$$

Nous étudions la limite quand $n \to \infty$ de la quantité

$$\mathbb{P}(\sqrt{n}(X_{k_n:n} - Q(p)) \le a) = \mathbb{P}(X_{k_n:n} \le Q(p) + an^{-1/2}) \,,$$

et ce pour tout $a \in \mathbb{R}$. Définissons

$$K_n := \sum_{i=1}^{n} \mathbb{1}\Big\{X_i \le Q(p) + an^{-1/2}\Big\} \,.$$

La clef de la preuve est d'observer que

$$\{X_{k_n:n} \le Q(p) + an^{-1/2}\} = \{K_n \ge k_n\} \,. \tag{IV-2.20}$$

Posons, pour $i \in \{1, \ldots, n\}$,

$$Y_{n,i}(a) := n^{-1/2} \left\{ \mathbb{1}\Big\{X_i \le Q(p) + an^{-1/2}\Big\} - F(Q(p) + an^{-1/2}) \right\} \,.$$

Nous définissons ainsi un tableau triangulaire de variables aléatoires réelles, $\{\{Y_{n,i}(a)\}_{i=1}^{n}, n \in \mathbb{N}\}$ centrées. Nous allons montrer que, pour tout $a \in \mathbb{R}$,

$$Z_n(a) := \sum_{i=1}^{n} Y_{n,i}(a) \Longrightarrow \mathrm{N}(0, p(1-p)) \,, \tag{IV-2.21}$$

par application du théorème de Lindeberg-Feller (Théorème IV-5.41). Notons tout d'abord que, pour tout $\varepsilon > 0$,

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}[Y_{n,i}^2(a) \mathbb{1}\{|Y_{n,i}(a)| \ge \varepsilon\}] &= n\mathbb{E}[Y_{n,1}^2(a) \mathbb{1}\{|Y_{n,1}(a)| \ge \varepsilon\}] \le \mathbb{P}\left(|Y_{n,1}(a)| \ge \varepsilon\right) \\ &\le \mathbb{P}(|\,\mathbb{1}\Big\{X_1 \le Q(p) + an^{-1/2}\Big\} - F(Q(p) + an^{-1/2})| \ge \varepsilon\sqrt{n}) \,.\end{aligned}$$

On en déduit que

$$\lim_n \sum_{i=1}^{n} \mathbb{E}[Y_{n,i}^2(a) \mathbb{1}\{|Y_{n,i}(a)| \ge \varepsilon\}] = 0,$$

ce qui montre la condition (IV-5.17) du théorème IV-5.41. Remarquons d'autre part, que

$$\sum_{i=1}^{n} \mathbb{E}[Y_{n,i}^2(a)] = \mathrm{Var}\left(\mathbb{1}\Big\{X_1 \le Q(p) + an^{-1/2}\Big\}\right) = F(Q(p) + an^{-1/2})(1 - F(Q(p) + an^{-1/2}))$$

et comme par (IV-2.19), $F(Q(p)) = p$, nous avons

$$\lim_n \sum_{i=1}^{n} \mathbb{E}[Y_{n,i}^2(a)] = \sigma^2(p) \,, \quad \text{où } \sigma^2(p) := p(1-p);$$

ce qui établit la condition (IV-5.18) du théorème IV-5.41. La convergence (IV-2.21) découle directement du théorème de Lindeberg-Feller (théorème IV-5.41). Le théorème de Polya (Théorème IV-5.44) montre que la convergence (IV-2.21) est uniforme, i.e.

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}(Z_n < x) - \Phi(\sigma^{-1}(p)x)\right| = 0 \,, \tag{IV-2.22}$$

où Φ est la fonction de répartition d'une loi normale centrée réduite (et donc $\Phi(x/\sigma)$ est la fonction de répartition d'une loi normale centrée de variance σ^2). Comme $\mathbb{P}(Z_n \ge x) = 1 - \mathbb{P}(Z_n < x)$ et $1 - \Phi(u) = \Phi(-u)$ pour tout $u \in \mathbb{R}$, (IV-2.22) implique

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}(Z_n \ge x) - \Phi(-\sigma^{-1}(p)x)\right| = 0 \,. \tag{IV-2.23}$$

Il découle maintenant de (IV-2.23) que

$$\begin{aligned}\mathbb{P}(\sqrt{n}(X_{k_n:n}-Q(p))\le a) &= \mathbb{P}(K_n\ge k_n)\\ &= \mathbb{P}\left(\frac{K_n-nF(Q(p)+an^{-1/2})}{\sqrt{n}}\ge\frac{k_n-nF(Q(p)+an^{-1/2})}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(Z_n(a)\ge\frac{k_n-nF(Q(p)+an^{-1/2})}{\sqrt{n}}\right)\\ &= \Phi(\sigma^{-1}(p)\{nF(Q(p)+an^{-1/2})-k_n\}/\sqrt{n})+o(1)\,.\end{aligned}$$

Comme la fonction F est dérivable au point $Q(p)$, nous avons en utilisant (IV-2.18),

$$\begin{aligned}\frac{nF(Q(p)+an^{-1/2})-k_n}{\sqrt{n}} &= a\frac{F(Q(p)+an^{-1/2})-p}{an^{-1/2}}+\frac{np-k_n}{\sqrt{n}}\\ &= a\frac{F(Q(p)+an^{-1/2})-p}{an^{-1/2}}-\sqrt{n}\left(\frac{k_n}{n}-p\right)\to aF'(Q(p))\,.\end{aligned}$$

Nous avons donc établi que, pour tout $a\in\mathbb{R}$,

$$\mathbb{P}(\sqrt{n}(X_{k_n:n}-Q(p))\le a)\to\Phi(\sigma^{-1}(p)aF'(Q(p)))$$

ou de façon équivalente

$$\sqrt{n}(X_{k_n:n}-Q(p))\Longrightarrow \mathrm{N}\left(0,\ \frac{\sigma^2(p)}{[F'(Q(p))]^2}\right)\,. \tag{IV-2.24}$$

□

Chapitre IV-3

Famille de distributions

IV-3.1 Loi gaussienne

Définition IV-3.1 (Loi Gaussienne réduite). *Une variable aléatoire X à valeur dans $\mathbb{R}$ est dite gaussienne réduite si sa loi admet pour densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$:*

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

La fonction caractéristique de la loi gaussienne réduite a pour expression :

$$\varphi(t) = \mathbb{E}\left[\exp(itX)\right] = \exp(-t^2/2)$$

Les moments de la loi gaussienne réduite se déduisent du développement de Taylor de $\varphi(t)$ en 0 : les moments d'ordre impair sont nuls et les moments d'ordre pair sont donnés par

$$\mu_{2n} = \mathbb{E}\left[X^{2n}\right] = \frac{(2n)!}{n!\ 2^n} = 1 \times 3 \times 5 \ldots \times (2n-1)$$

Définition IV-3.2 (Loi gaussienne). *Une variable aléatoire X à valeur dans $\mathbb{R}$ est dite gaussienne si elle peut s'écrire sous la forme $X = \sigma X_r + \mu$ où X_r est une variable aléatoire gaussienne réduite. On note $X \sim \mathrm{N}(\mu, \sigma^2)$. μ est l'espérance de X et σ^2 sa variance. Lorsque $\sigma^2 > 0$, la loi $\mathrm{N}(\mu, \sigma^2)$ a une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par :*

$$\mathrm{g}_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

La fonction caractéristique d'une variable gaussienne de moyenne μ et de variance σ^2 est donnée par

$$\varphi_{\mu,\sigma^2}(t) = \exp\left(\mathrm{i}\mu t - \frac{\sigma^2}{2} t^2\right) . \tag{IV-3.1}$$

IV-3.2 Loi gaussienne multivariée

Définition IV-3.3 (Loi gaussienne multivariée). *Un vecteur aléatoire* $\mathbf{X} = (X_1, \dots, X_n)$ *est dit* gaussien *si, pour tout* $\mathbf{t} = (t_1, \dots, t_n) \in \mathbb{R}^n$, $\sum_{j=1}^n t_j X_j = \mathbf{t}^T \mathbf{X}$ *est une variable aléatoire gaussienne.*

Soit $\mathbf{X} = (X_1, \dots, X_n)$ un vecteur aléatoire gaussien. Pour $\mathbf{t} = (t_1, \dots, t_n) \in \mathbb{R}^n$, $Y = \mathbf{t}^T \mathbf{X}$ est une variable gaussienne, dont l'espérance et la variance sont données respectivement par :

$$\mathbb{E}[Y] = \sum_{i=1}^n t_i \mathbb{E}[X_i] = \mathbf{t}^T \mathbb{E}[\mathbf{X}]$$
$$\mathbb{E}[(Y - \mathbb{E}[Y])^2] = \sum_{i,j=1}^n t_i t_j \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbf{t}^T \Gamma \mathbf{t}$$

où $\Gamma = (\mathrm{Cov}(X_i, X_j))_{1 \le i,j \le n}$ est la matrice de covariance du vecteur $\mathbf{X}$.

Définition IV-3.4 (loi $\mathrm{N}(\boldsymbol{\mu}, \Gamma)$). *Soit* $\boldsymbol{\mu} \in \mathbb{R}^n$ *et* Γ *une matrice* $n \times n$ *symétrique et semi-définie positive. Nous dirons que* $\mathbf{X} = (X_1, \dots, X_n)$ *suit une loi multivariée gaussienne de moyenne* $\boldsymbol{\mu}$ *et de covariance* Γ *et nous écrirons* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$*, si, pour tout* $\mathbf{t} = (t_1, \dots, t_n) \in \mathbb{R}^n$*, nous avons* $\mathbf{t}^T \mathbf{X} \sim \mathrm{N}(\mathbf{t}^T \boldsymbol{\mu}, \mathbf{t}^T \Gamma \mathbf{t})$.

Cette définition implique de façon immédiate :

Proposition IV-3.5. *Soit* $\mathbf{A}$ *une matrice* $m \times n$*,* $\mathbf{b} \in \mathbb{R}^n$ *et soit* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$*. Alors,* $\mathbf{AX} + \mathbf{b} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}^T \Gamma \mathbf{A})$.

Démonstration. Posons $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ et notons que, pour tout $\mathbf{s} \in \mathbb{R}^m$, nous avons :

$$\mathbf{s}^T \mathbf{Y} = (\mathbf{A}^T \mathbf{s})^T \mathbf{X} + s^T \mathbf{b} \sim \mathrm{N}(\mathbf{s}^T \mathbf{A}\boldsymbol{\mu} + s^T \mathbf{b}, \mathbf{s}^T \mathbf{A} \Gamma \mathbf{A}^T \mathbf{s}). \qquad \square$$

Proposition IV-3.6. *Soit* $\Gamma \in \mathrm{Mat}_n(\mathbb{R})$ *une matrice semi-définie positive,* $\mathrm{Rang}(\Gamma) = k \le n$ *et soit* $\boldsymbol{\mu} \in \mathbb{R}^n$*.* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$ *si et seulement si, pour tout* $\mathbf{A} \in \mathrm{Mat}_{n,k}(\mathbb{R})$ *tel que* $\mathbf{A}\mathbf{A}^T = \Gamma$*, il existe* $\mathbf{Z} \sim \mathrm{N}(0, \mathrm{I}_n)$ *tel que* $\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}$.

Démonstration. Si $\mathbf{Z} \sim \mathrm{N}(0, \mathrm{I}_n)$, nous déduisons de la proposition IV-3.5 $\mathbf{AZ} + \boldsymbol{\mu} \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$.

Réciproquement, soit $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$. Comme $\mathbf{A}$ est de rang k, la matrice $\mathbf{A}^T \mathbf{A} \in \mathrm{Mat}_k(\mathbb{R})$ est inversible et $\mathbf{A}$ est inversible à gauche. Notons $\mathbf{A}^\# := (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ son inverse à gauche. Nous avons : $\mathbf{A}^\# \mathbf{A} = \mathrm{I}_k$ et $\mathbf{A}\mathbf{A}^\#$ est le projecteur orthogonal sur l'image de $\mathbf{A}$ (par construction, $\mathrm{Im}(\mathbf{A}) = \mathrm{Im}(\Gamma)$). Soit $\mathbf{Z} = \mathbf{A}^\#(\mathbf{X} - \boldsymbol{\mu})$. La proposition IV-3.5 implique que $\mathbf{Z} \sim \mathrm{N}(0, \mathrm{I}_k)$. $\square$

On pourrait choisir cette caractérisation comme définition de la loi gaussienne $\mathrm{N}(\boldsymbol{\mu}, \Gamma)$.

La fonction caractéristique de $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$ se déduit directement de (IV-3.1). Comme la fonction caractéristique caractérise la loi, nous avons :

Proposition IV-3.7. $\mathbf{X} = (X_1, \dots, X_n) \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$ *si et seulement si sa fonction caractéristique* $\varphi_{\boldsymbol{\mu}, \Gamma}(\mathbf{t}) := \mathbb{E}[\mathrm{e}^{\mathrm{i}\mathbf{t}^T \mathbf{X}}]$ *est donnée par :*

$$\varphi_{\boldsymbol{\mu}, \Gamma}(\mathbf{t}) = \exp\left(\mathrm{i}\mathbf{t}^T \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right) \qquad \text{(IV-3.2)}$$

Soit $n \in \mathbb{N}$ et soient n_1, n_2 tels que $n_1 + n_2 = n$. Pour tout $\mathbf{x} \in \mathbb{R}^n$, on définit $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ et $\mathbf{x}_2 \in \mathbb{R}^{n_2}$ tels que $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$. De façon similaire, toute matrice $\Gamma \in \mathbb{R}_n^n$ se décompose par bloc :

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}.$$

Nous avons :

Proposition IV-3.8. *Soit* $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T \sim \mathrm{N}(\boldsymbol{\mu}, \Gamma)$. $\mathbf{X}_1$ *est indépendant de* $\mathbf{X}_2$ *si et seulement si* $\Gamma_{12} = 0$.

Démonstration. Si $\mathbf{X}_1$ et $\mathbf{X}_2$ sont indépendants, alors $\Gamma_{12} = \mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$. Réciproquement, supposons que $\Gamma_{12} = 0$. Alors $\mathbf{t}^T \Gamma \mathbf{t} = \mathbf{t}_1^T \Gamma_{1,1} \mathbf{t}_1 + \mathbf{t}_2^T \Gamma_{2,2} \mathbf{t}_2$, donc, d'après la proposition IV-3.6

$$\mathbb{E}\left[\mathrm{e}^{i\boldsymbol{t}^T(\mathbf{X}_1^T, \mathbf{X}_2^T)^T}\right] = \varphi_{\boldsymbol{\mu},\Gamma}(\mathbf{t}) = \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\Gamma\mathbf{t}\right) = \mathbb{E}\left[\mathrm{e}^{i\boldsymbol{t_1}^T\mathbf{X}_1}\right]\mathbb{E}\left[e^{i\boldsymbol{t_2}^T\mathbf{X}_2}\right].$$

Par suite, les variables aléatoires $\mathbf{X}_1$ et $\mathbf{X}_2$ sont indépendantes, ce qui conclut la preuve. □

Corollaire IV-3.9. *Soient* $\mathbf{A}_1 \in \mathrm{Mat}_{n,n_1}(\mathbb{R})$ *et* $\mathbf{A}_2 \in \mathrm{Mat}_{n,n_2}(\mathbb{R})$ *deux matrices telles que* $\mathbf{A}_1^T\mathbf{A}_2 = \mathbf{0}_{n_1 \times n_2}$ *et soit* $\mathbf{Z} \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n)$. *Alors, le vecteur* $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T$ *avec* $\mathbf{Y}_1 := \mathbf{A}_1\mathbf{Z}$ *et* $\mathbf{Y}_2 := \mathbf{A}_2\mathbf{Z}$ *est gaussien et les vecteurs aléatoires* $\mathbf{Y}_1$ *et* $\mathbf{Y}_2$ *sont indépendants.*

Remarque IV-3.10. La *décorrélation* des composantes d'un vecteur aléatoire n'implique *l'indépendance* de ses composantes que dans le cas où le vecteur est gaussien. Nous donnons un contre-exemple pour illustrer l'importance de cette hypothèse. Soit X une variable aléatoire de loi $\mathrm{N}(0,1)$; $Y = \varepsilon X$, où ε est variable aléatoire indépendante de X telle que $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$. On démontre aisément que $Y \sim \mathrm{N}(0,1)$. De plus,

$$\mathbb{E}[XY] = \mathbb{E}[\varepsilon X^2] = \mathbb{E}[\varepsilon]\mathbb{E}[X^2] = 0,$$

et donc $\mathrm{Cov}(X,Y) = 0$, ces variables aléatoires sont décorrélées. Pourtant, elles ne sont pas indépendantes puisque $\mathbb{P}(X \in [0,1], Y \in [1,2]) = 0 \neq \mathbb{P}(X \in [0,1])\mathbb{P}(Y \in [1,2])$. Pour voir que le vecteur (X,Y) n'est pas un vecteur aléatoire gaussien, on peut remarquer que $\mathbb{P}(X+Y=0) = \frac{1}{2}$ donc $X+Y$, qui n'est ni constante p.s. ni à densité, n'est pas gaussienne. ♢

Proposition IV-3.11. *Soit* Γ *une matrice définie positive* $n \times n$. *Une loi gaussienne de moyenne* $\boldsymbol{\mu}$ *et de covariance* Γ *admet une densité par rapport à la mesure de Lebesgue sur* $\mathbb{R}^n$ *de la forme :*

$$\mathrm{g}_{\boldsymbol{\mu},\Gamma}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sqrt{\det(\Gamma)}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Gamma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^n. \tag{IV-3.3}$$

Démonstration. Si $\Gamma = \mathrm{I}_n$ la proposition IV-3.8 montre que les variables aléatoires $X_1, \ldots, X_n$ sont i.i.d. et donc leur densité jointe est égale au produit des densités marginales, ce qui conduit au résultat dans ce cas particulier. Si Γ est une matrice définie positive quelconque, nous utilisons la proposition IV-3.6 : il existe $\mathbf{A}$ *inversible* et $\mathbf{Z} \sim \mathrm{N}(0, \mathrm{I}_n)$ tel que $\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$ et l'expression IV-3.3 découle de la formule du changement de variable. □

IV-3.3 Loi Gamma

La loi Gamma permet de construire de nombreuses autres distributions. La loi Gamma est elle-même liée à la fonction Gamma, définie sur le demi plan complexe $\mathrm{Re}(z) > 0$ par :

$$\Gamma(z) = \int_0^\infty \exp(-t)\, t^{z-1}\mathrm{d}t = 2\int_0^\infty \exp(-t^2)\, t^{2z-1}\mathrm{d}t. \tag{IV-3.4}$$

En intégrant par partie pour $x > 0$ *réel positif* l'expression précédente, nous avons :

$$\Gamma(x) = [-t^{x-1}\mathrm{e}^{-t}]_0^\infty + (x-1)\int_0^\infty t^{x-2}\mathrm{e}^{-t}\mathrm{d}t = (x-1)\Gamma(x-1)\,.$$

Donc, pour tout un entier naturel $n \geq 1$, $\Gamma(n) = (n-1)\Gamma(n-1) = \ldots = (n-1)(n-2)\ldots 1 = (n-1)!$.

Définition IV-3.12. *(i) Pour tout réel $p > 0$, on appelle loi Gamma réduite à p degrés de liberté (et l'on note* Gamma(p)*) la loi définie sur l'ensemble des réels positifs par la densité*

$$f_p(x) = \frac{1}{\Gamma(p)}\exp(-x)\,x^{p-1}, \quad x > 0\,.$$

(ii) Pour $\lambda > 0$, on appelle loi Gamma Gamma(p,λ)*, la loi de la variable aléatoire $X = Z/\lambda$ où Z est une loi Gamma à p degrés de liberté et λ est le paramètre d'intensité. La densité de la loi* Gamma(p,λ) *est donnée par*

$$f_{p,\lambda}(x) = \frac{\lambda^p}{\Gamma(p)}\exp(-\lambda x)\,x^{p-1}\,, \quad x > 0\,.$$

Un cas particulier important est fourni par la loi exponentielle d'intensité $\lambda > 0$, de densité sur $\mathbb{R}_+$ donnée par $x \mapsto \lambda\mathrm{e}^{-\lambda x}$ qui coïncide avec la loi $\Gamma(1,\lambda)$.

Si Z est une loi Gamma(p), (IV-3.4) implique que, pour tout $r > -p$, nous avons pour tout $r > -p$,

$$\mathbb{E}[Z^r] = \frac{\Gamma(p+r)}{\Gamma(p)}\,. \tag{IV-3.5}$$

La fonction caractéristique de la loi Gamma(p,λ) est donnée par :

$$\phi_{\lambda,p}(t) = \int_0^\infty \frac{\lambda^p}{\Gamma(p)}x^{p-1}\mathrm{e}^{\lambda(\mathrm{i}t/\lambda-1)x}\mathrm{d}x = (1-\mathrm{i}t/\lambda)^{-p}\,. \tag{IV-3.6}$$

Cette expression particulière de la fonction caractéristique a pour conséquence immédiate le lemme de convolution suivant pour les lois Gammas.

Lemme IV-3.13. *Soit $(X_1,\ldots,X_n)$ n variables aléatoires indépendantes distribuées suivant des lois* Gamma(p_i,θ) *avec $\theta > 0$ et $p_i > 0$, $i \in \{1,\ldots,n\}$. Alors, $\sum_{i=1}^n X_i$ est distribuée suivant une loi* Gamma$(\sum_{i=1}^n p_i,\theta)$.

IV-3.4 La loi du χ^2 à k degrés de liberté

Définition IV-3.14 (Loi du χ^2-centrée ou Khi-deux). *Soient $(X_1,\cdots,X_\nu)$, ν variables aléatoires gaussiennes de moyenne nulle et de variance unité indépendantes. La variable aléatoire $U = \sum_{i=1}^\nu X_i^2$ suit une loi du χ^2 centrée à ν degrés de liberté, notée χ^2_ν.*

Lemme IV-3.15. *(i) Soit X une variable aléatoire gaussienne centrée réduite. X^2 suit une loi* Gamma$(1/2,1/2)$.

(ii) Soit $\nu \in \mathbb{N}$ et $X_1,X_2,\ldots,X_\nu$ ν variables aléatoires gaussiennes centrées réduites. $\sum_{i=1}^\nu X_i^2$ suit une loi Gamma$(\nu/2,1/2)$.

Démonstration. (i) $\mathbb{P}\left(X^2 < z\right) = 0$ si $z < 0$. Pour $z > 0$, nous avons :

$$\mathbb{P}\left(X^2 < z\right) = \mathbb{P}\left(-\sqrt{z} < X < \sqrt{z}\right) = \int_{-\sqrt{z}}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \mathrm{d}x$$
$$= 2\int_0^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \mathrm{d}x = \frac{2}{\sqrt{2\pi}} \int_0^z \exp(-\frac{u}{2}) \frac{1}{2\sqrt{u}}\, \mathrm{d}u$$

Ceci conduit au résultat, en utilisant le résultat élémentaire $\Gamma(1/2) = \sqrt{\pi}$.

(ii) Le résultat est une conséquence élémentaire de Lemme IV-3.13. □

Proposition IV-3.16. *La densité de la loi du χ^2 centrée à ν degrés de liberté est :*

$$f_\nu(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} \mathrm{e}^{-\frac{x}{2}} \mathbb{1}_{\{x>0\}} \tag{IV-3.7}$$

Démonstration. La preuve découle de Lemme IV-3.15 et de l'expression de la densité d'une loi Gamma$(\nu/2, 1/2)$.□

Proposition IV-3.17. *La moyenne de la loi du χ^2_ν centrée vaut ν, sa variance 2ν.*

Démonstration. Repartons de la définition de la loi χ^2_ν . Comme $(Z_1, \ldots, Z_\nu)$ sont ν variables gaussiennes centrées réduites, on a

$$\mathbb{E}[Z_i^2] = \mathrm{Var}(Z_i) = 1 \quad \text{donc} \quad \mathbb{E}\left[\sum_{i=1}^{\nu} Z_i^2\right] = \nu\,,$$
$$\mathrm{Var}(Z_i^2) = \mathbb{E}[Z_i^4] - (\mathbb{E}[Z_i^2])^2 = 2 \quad \text{donc} \quad \mathrm{Var}\left(\sum_{i=1}^{\nu} Z_i^2\right) = 2\nu\,. \qquad \square$$

Proposition IV-3.18. *Soient $(X_1, \cdots, X_k)$, k variables aléatoires gaussiennes de moyenne $(\mu_1, \ldots, \mu_k)$ de variance unité et indépendantes. Posons $U = \sum_{i=1}^k X_i^2$ et $\gamma = \sum_{i=1}^k \mu_i^2$.*

(i) La fonction caractéristique de la loi de la variable U est donnée par :

$$\left(\frac{1}{\sqrt{1-2\mathrm{i}t}}\right)^k \exp\left(\frac{\mathrm{i}t\gamma}{1-2\mathrm{i}t}\right)$$

(ii) La loi de la variable aléatoire U admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnéee :

$$f_{k,\gamma}(x) = \sum_{i=0}^{\infty} \frac{\mathrm{e}^{-\gamma/2}(\gamma/2)^i}{i!} f_{k+2i}(x)$$

où f_ν est la densité d'une loi du χ^2 à q degrés de liberté.

Démonstration. (i) Par définition, si $Z_1, \ldots, Z_{k-1}, X$ sont des variables aléatoires indépendantes, $Z_i \sim \mathrm{N}(0,1)$ et $X \sim \mathrm{N}(\mu, 1)$ alors la variable aléatoire $U = \sum_{i=1}^{k-1} Z_i^2 + X^2 \sim \chi^2_k(\gamma)$ avec $\gamma = \mu^2/2$. Notons que $\sum_{i=1}^{k-1} Z_i^2$ et X^2 sont indépendantes et que $\sum_{i=1}^{k-1} Z_i^2 \sim \chi^2_{k-1}$. Par conséquent,

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}tU}] = (1-2\mathrm{i}t)^{-(k-1)/2} \mathbb{E}[\mathrm{e}^{\mathrm{i}tX^2}].$$

Un calcul direct montre que :

$$\begin{aligned}\mathbb{E}[e^{itX^2}] &= \int_{-\infty}^{\infty} e^{itx^2}(2\pi)^{-1/2}e^{-(x-\mu)^2/2}dx \\ &= \exp\left[\frac{\mu^2(it)}{1-2it}\right]\int_{-\infty}^{\infty}(2\pi)^{-1/2}\exp\left(-\frac{1-2it}{2}\left[x-\frac{\mu}{1-2it}\right]^2\right)dx \\ &= (1-2it)^{-1/2}\exp(\gamma it/(1-2it)).\end{aligned}$$

(ii) La preuve est délicate et est omise. □

Définition IV-3.19 (Loi du χ^2 non centrée). *Soient $(X_1,\cdots,X_k)$, k variables aléatoires gaussiennes de moyenne $(\mu_1,\ldots,\mu_k)$ de variance unité et indépendantes, $U=\sum_{i=1}^k X_i^2$ et $\gamma=\sum_{i=1}^k \mu_i^2$. La loi de U, notée $\chi_k^2(\gamma)$, est appelée loi du χ^2 à k degrés de liberté non-centrée, de paramètre de non-centralité γ.*

Proposition IV-3.18. La loi du khi-2 non-central peut donc être vue comme un loi mélange de lois du khi-2 centrées. Supposons que la variable J suit une loi de Poisson de moyenne $\gamma/2$ et que la loi conditionnelle de Z sachant $J=i$ soit la loi du khi-2 centré à $k+2i$ degrés de liberté. Alors la loi (non conditionnelle) de Z est la loi du khi-2 non centrée à k degrés de liberté, de paramètre de non-centralité γ. Un calcul élémentaire montre que la moyenne et la variance d'une loi $\chi_k^2(\gamma)$ sont respectivement données par $k+\gamma$ et $2k+4\gamma$.

Le résultat suivant joue un rôle important dans la théorie de l'inférence dans les modèles de régression linéaire multiple.

Proposition IV-3.20. *Soit $\mathbf{Z}\sim N_n(\boldsymbol{\mu},\sigma^2 I)$ et soit Π un projecteur orthogonal de rang $k<n$. La variable aléatoire $\|\Pi\mathbf{Z}\|^2/\sigma^2$ est distribuée suivant une loi de χ^2 non-centrée à k degrés de liberté de paramètre de non-centralité $\|\Pi\boldsymbol{\mu}\|^2$.*

Démonstration. Soit $H=[\mathbf{h}_1,\ldots,\mathbf{h}_k]$ une base orthonormale de l'image de Π. Nous avons donc $\Pi = HH^T$ et $H^TH=I_k$ où I_k est la matrice identité $(k\times k)$. Par conséquent, $\mathbf{Z}^T\Pi\mathbf{Z}=\|\mathbf{E}\|^2$, où $\mathbf{E}=H^T\mathbf{Z}$. La proposition IV-3.5 implique que $\mathbf{E}\sim N_k(H^T\boldsymbol{\mu},I_k)$. Par conséquent, $\|\mathbf{E}\|^2\sim\chi_k^2(\delta)$ avec $\delta=\left\|H^T\boldsymbol{\mu}\right\|^2=\boldsymbol{\mu}^T\Pi\boldsymbol{\mu}$, ce qui conclut la preuve. □

IV-3.5 Loi de Student

Définition IV-3.21. *Soient X et Y deux variables aléatoires indépendantes telles que :*

- *X suit une loi gaussienne centrée réduite,*
- *Y suit une loi du χ^2 centrée à r degrés de liberté,*

Alors $T=X/\sqrt{Y/r}$ suit une loi de Student à r degrés de liberté, notée $t(r)$.

Proposition IV-3.22. *La densité d'une loi de Student à r-degrés de liberté est donnée par :*

$$f_r(t)=\frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)}\frac{1}{(r\pi)^{1/2}}\left(1+\frac{t^2}{r}\right)^{-\frac{r+1}{2}}$$

Démonstration. La distribution conjointe des variable aléatoire X et Y est donnée par

$$f_{XY}(x,y) \propto \mathrm{e}^{-x^2/2} y^{(r/2)-1} \mathrm{e}^{-y/2}, \quad x \in \mathbb{R}, y > 0.$$

En appliquant la transformation $\phi : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R} \times \mathbb{R}^+$, $(x,y) \mapsto (x(y/r)^{-1/2}, y)$, la loi jointe de T et de Y est donnée par :

$$f_{TY}(t,y) = f_{XY}(t(y/r)^{1/2}, y)(y/r)^{1/2}, \quad x \in \mathbb{R}, y > 0,$$

car le Jacobien de la transformation est égal à $(y/r)^{1/2}$. La distribution de T est obtenue en intégrant la loi jointe f_{TY} par rapport à y,

$$f_T(t) \propto \int_0^\infty \mathrm{e}^{-y(1+t^2/r)/2} y^{((r+1)/2)-1} \mathrm{d}y.$$

et on obtient la formule désirée en faisant le changement de variable $u = y(1+t^2/r)/2$. □

Lorsque $r = 1$, la densité de la loi de Student se réduit à

$$f_1(t) = \frac{1}{\pi(1+t^2)}, \quad t \in \mathbb{R}$$

qui est la densité d'une loi de Cauchy qui n'admet pas de moments d'ordre 1 (voir Section IV-3.7).

Proposition IV-3.23. *La suite de loi de Student* $\{\mathrm{t}(r), r \in \mathbb{N}\}$ *converge faiblement vers une loi normale centrée réduite.*

Démonstration. Nous allons donner deux preuves de cette propriété, la première analytique et la deuxième probabiliste. Lorsque $r \to \infty$, on démontre aisément que pour tout $t \in \mathbb{R}$,

$$\lim_{r\to\infty} f_r(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \, .$$

En effet, pour tout $t \in \mathbb{R}$ nous avons

$$\lim_{r\to\infty} \left(1 + \frac{t^2}{r}\right)^{-(r+1)/2} = \mathrm{e}^{-t^2/2} \, .$$

En utilisant le développement de Stirling de la fonction Γ

$$\Gamma(z) = \sqrt{2\pi} z^{z-\frac{1}{2}} \mathrm{e}^{-z} \left[1 + \frac{1}{12z} + O\left(\frac{1}{z^{-2}}\right)\right] \, .$$

valide pour tout $z \in \mathbb{C} \setminus \mathbb{Z}_-$, on démontre que

$$\lim_{r\to\infty} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \frac{1}{(r)^{1/2}} = \frac{1}{\sqrt{2}} \, .$$

D'autre part, il existe une constante $C < \infty$ telle que pour tout $r \geq 1$ et $t \in \mathbb{R}$,

$$f_r(t) \leq C(1+t^2)^{-1} \, .$$

En utilisant le théorème de la convergence dominée, nous avons donc, pour toute fonction h continue bornée

$$\lim_{r\to\infty} \int h(t) f_r(t) \mathrm{d}t = \int h(t) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \mathrm{d}t \, ,$$

ce qui montre que la suite de lois $\mathrm{t}(r)$ converge vers une loi normale centrée réduite.

On peut donner une preuve probabiliste de ce résultat. En effet, soit $\{Z_k,\ k \in \mathbb{N}\}$ une suite de variables aléatoires gaussiennes centrées réduites. Nous savoir (voir Définition IV-3.14) que pour tout entier r, $\sum_{i=1}^{r} Z_i^2$ suit une loi du χ^2 à r degrés de liberté. Soit X une variable aléatoire centrée réduite indépendante de la suite $\{Z_k,\ k \in \mathbb{N}\}$. Pour tout entier r la variable aléatoire

$$T_r = \left(r^{-1} \sum_{i=1}^{r} Z_i^2 \right)^{-1/2} X \ ,$$

est distribuée suivant une loi de student à r-degrés de liberté. Par la loi des grands nombres (Théorème IV-5.18), $r^{-1}\sum_{i=1}^{r} Z_i^2 \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 1$ et par le lemme de Slutsky (Lemme IV-5.33) nous avons donc que $T_r \Longrightarrow \mathrm{N}(0,1)$. □

Le résultat suivant, du à Gosset (1907), fait partie des "classiques favoris" des statistiques élémentaires et justifie à lui seul l'intérêt porté à la distribution de Student.

Théorème IV-3.24. *Soient $(X_1,\ldots,X_n)$ n variables aléatoires gaussiennes indépendantes de moyenne μ et de variance $\sigma^2 > 0$.*

(i) La moyenne empirique $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ et la variance de l'échantillon

$$S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

sont indépendantes.

(ii) La moyenne empirique $\bar{X}_n$ suit une loi normale $\mathrm{N}(\mu, \sigma^2/n)$.

(iii) $(n-1)S_n^2/\sigma^2$ suit une loi du χ^2 centrée à $(n-1)$ degrés de liberté.

(iv) La variable T_n définie par :

$$T_n = \sqrt{n}\frac{\bar{X}_n - \mu}{S_n}$$

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

Démonstration. Posons $\mathbf{X} = (X_1,\ldots,X_n) \sim \mathrm{N}(\mu \mathbf{1}_n, \sigma^2 \mathrm{I}_n)$ où $\mathbf{1}_n = [1,\ldots,1]^T$ Notons que $\bar{X}_n = n^{-1}\mathbf{1}_n^T\mathbf{X}$ et donc que :

$$(n-1)S_n^2 = \left\| \mathbf{X} - n^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{X} \right\|^2 = \left\| (\mathrm{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X} \right\|^2 \ .$$

Remarquons que

$$\Pi := \mathrm{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$$

est un projecteur orthogonal de rang $(n-1)$ et $\Pi\mathbf{1}_n = 0$. La proposition IV-3.20 montre que $(n-1)S_n^2/\sigma^2$ est distribuée suivant une loi du χ^2 centré à $(n-1)$ degrés de liberté. Le corollaire IV-3.9 montre que $\bar{X}_n = n^{-1}\mathbf{1}_n^T\mathbf{X}$ et $\Pi\mathbf{X}$ sont indépendants et le résultat découle de : $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim \mathrm{N}(0,1)$. □

Remarque IV-3.25. On peut montrer que la propriété d'indépendance de $\bar{X}_n$ et S_n^2 est *caractéristique* du cas Gaussien : si cette propriété est vérifiée, alors, $\mathbf{X}$ est Gaussien. ♢

IV-3.6 Loi de Fisher

Définition IV-3.26 (Loi de Fisher). *Soient X et Y deux variables aléatoires indépendantes telles que :*

— *X suit une loi du χ^2 centré à q-degrés de liberté,*

— *Y suit une loi du χ^2 centré à r degrés de liberté,*
Alors $W = (X/q)/(Y/r)$ suit une loi de Fisher à (q,r)-degrés de liberté, ce que l'on note $\mathrm{F}(q,r)$.

Proposition IV-3.27. *La loi de Fisher à (q,r)-degrés de liberté a une densité donnée par*

$$f(w) = \frac{\Gamma\left(\frac{q+r}{2}\right)}{\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{r}{2}\right)} \left(\frac{q}{r}\right)^{q/2} \frac{w^{q/2-1}}{(1+(q/r)w)^{(q+r)/2}}, \quad w > 0\,.$$

Démonstration. La preuve est similaire à la preuve de la IV-3.22 et est omise. □

Remarquons que, par définition, si W est distribuée suivant la loi de Fisher $\mathrm{F}_{q,r}$ alors $1/W$ est distribuée suivant la loi de Fisher $\mathrm{F}_{r,q}$. Notons aussi que si T est distribuée suivant une loi de Student à r degrés de liberté, alors T^2 est distribuée suivant une loi de Fisher à $(1,r)$-degrés de liberté.

Les applications de la loi de Fisher sont nombreuses en statistique. L'application la plus directe est la suivante. Soient $(X_1,\ldots,X_n)$ des variables aléatoires gaussiennes indépendantes de loi $\mathrm{N}(\mu_1,\sigma^2)$ et $(Y_1,\ldots,Y_m)$ des variables aléatoires gaussiennes indépendantes de loi $\mathrm{N}(\mu_2,\sigma^2)$, indépendantes de $(X_1,\ldots,X_n)$. On note $S_1^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ et $S_2^2 = (m-1)^{-1}\sum_{i=1}^m (Y_i - \bar{Y}_m)^2$ les variances empiriques de ces deux échantillons. Dans ce cas,

$$\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2(n-1)\,,\ \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(m-1)\,,$$

d'où immédiatement :

$$\frac{S_1^2}{S_2^2} \sim F(n-1,\ m-1)\,.$$

IV-3.7 Loi de Cauchy

La loi de Cauchy doit son nom au mathématicien Augustin Louis Cauchy. Une variable aléatoire X suit une loi de Cauchy de paramètre de position $\mu \in \mathbb{R}$ et de paramètre d'échelle σ si elle admet une densité $x \mapsto p(\theta,x)$ avec $\theta = (\mu,\sigma)$ par rapport à la mesure de Lebesgue définie par :

$$p(\theta,x) = p(\mu,\sigma,x) = \frac{1}{\pi\sigma\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]} = \frac{1}{\pi}\left[\frac{\sigma}{(x-\mu)^2+\sigma^2}\right] \tag{IV-3.8}$$

La fonction ainsi définie s'appelle très classiquement une lorentzienne. Cette distribution est symétrique par rapport au paramètre de position μ, le paramètre d'échelle σ donnant une information sur l'étalement de la distribution. La fonction de répartition d'une loi de Cauchy est donnée par

$$F_\theta(x) = \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{2}\,.$$

La loi de Cauchy n'admet ni espérance ni écart-type car la fonction

$$x \mapsto \frac{\sigma}{\mu}\frac{1}{\pi}\left[\frac{x}{(x-\mu)^2+\sigma^2}\right]$$

n'est pas intégrable au sens de Lebesgue.

La loi de Cauchy (avec notamment la loi normale et la loi de Lévy) est un cas particulier de loi stable. La fonction caractéristique de la loi de Cauchy est donnée par

$$\Phi_\theta(t) = \exp(\mathrm{i}\mu t - \sigma|t|) \, .$$

La fonction caractéristique de la somme de deux variables de Cauchy X_1, X_2 indépendantes de paramètres (μ_1, σ_1) et (μ_2, σ_2) est égale à

$$\exp(\mathrm{i}\mu_1 t - \sigma_1|t|)\exp(\mathrm{i}\mu_1 t - \sigma_1|t|) = \exp(\mathrm{i}(\mu_1+\mu_2)t - (\sigma_1+\sigma_2)|t|)$$

et est donc distribuée suivant une loi de Cauchy de paramètre $(\mu_1+\mu_2, \sigma_1+\sigma_2)$. En particulier, si $X_1, \ldots, X_n$ sont n variables aléatoires de Cauchy de paramètre de position μ et d'échelle σ, la moyenne empirique $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ est distribuée suivant une loi de Cauchy de paramètre (μ, σ). La moyenne empirique $\bar{X}_n$ et X_1 ont donc la même distribution ! En particulier, la loi des grands nombres n'est pas satisfaite par la loi de Cauchy. Ceci ne contredit pas Théorème IV-5.18 car les variables de Cauchy ne sont pas intégrables.

Chapitre IV-4

Famille Exponentielle

Définition IV-4.1 (Famille exponentielle canonique). *Soit* $(\mathsf{X}, \mathscr{X})$ *un espace mesurable et* $\mu \in \mathbb{M}_+(\mathsf{X}, \mathscr{X})$ *une mesure* σ*-finie. Nous appelons famille exponentielle canonique de dimension d engendrée par* $\mathbf{T} = (T^{(1)}, \ldots, T^{(d)})$ *et la fonction mesurable h la famille de lois indexées par* $\eta = (\eta_1, \ldots, \eta_d)$ *de densité*

$$q(\eta, x) = h(x) \exp\left(\eta^T \mathbf{T}(x) - A(\eta)\right) , \quad x \in \mathsf{X} , \tag{IV-4.1}$$

par rapport à une mesure μ *où* $A(\eta)$ *est défini par :*

$$A(\eta) = \log \int h(x) \exp\left(\eta^T \mathbf{T}(x)\right) \mu(\mathrm{d}x) . \tag{IV-4.2}$$

*L'*espace des paramètres naturels *de la famille canonique associée à* $(\mathbf{T}, h)$ *est l'ensemble*

$$\Xi = \left\{ \eta = (\eta_1, \ldots, \eta_d) \in \mathbb{R}^d \,:\, |A(\eta)| < \infty \right\} . \tag{IV-4.3}$$

Exemple IV-4.2. Supposons que μ soit la mesure de Lebesgue sur $\mathbb{R}$, $h = \mathbb{1}_{\{0,\infty\}}$, $d = 1$ et $T_1(x) = x$. Nous avons alors

$$A(\eta) = \log \int_0^\infty \mathrm{e}^{\eta x} \mathrm{d}x = \begin{cases} \log(-1/\eta) , & \eta < 0 \\ \infty, & \eta \geq 0. \end{cases}$$

Donc, $p(\eta, x) = \exp(\eta x - \log(-1/\eta)) \mathbb{1}_{\{0,\infty\}}(x)$ est une densité pour tout $\eta \in]-\infty, 0[$. Ceci correspond à la paramétrisation canonique de la loi exponentielle. ♢

Il est parfois souhaitable de ne pas se limiter à la paramétrisation canonique.

Définition IV-4.3 (Famille Exponentielle). *Soit* $(\mathsf{X}, \mathscr{X})$ *un espace mesurable. Soient* μ *une mesure* σ*-finie sur* $(\mathsf{X}, \mathscr{X})$*,* Θ *un ouvert de* $\mathbb{R}^d$*,* $\mathbf{T} = (T^{(1)}, \ldots, T^{(d)})$ *des fonctions mesurables de* $\mathsf{X} \to \mathbb{R}$*,* $\varphi = (\varphi^{(1)}, \ldots, \varphi^{(d)})$ *des fonctions de* $\Theta \to \mathbb{R}$*, et h une fonction positive measurable de* $\mathsf{X} \to \mathbb{R}^+$*.*

La famille de lois $(\mathbb{P}_\theta, \theta \in \Theta)$ *sur* X *est une famille exponentielle de dimension d si pour tout* $\theta \in \Theta$*,* $\mathbb{P}_\theta$ *admet une densité par rapport à la mesure* μ *donnée par*

$$p(\theta, x) = h(x) \exp\left[\varphi(\theta)^T \mathbf{T} - B(\theta)\right] , \quad x \in \mathsf{X}. \tag{IV-4.4}$$

Si nous disposons d'un n-échantillon $X_1,\ldots,X_n$ i.i.d. de cette famille exponentielle, la loi jointe est donnée par

$$p(\theta,x_1,\ldots,x_n)=\prod_{i=1}^{n}h(x_i)\exp\left[\sum_{j=1}^{d}\varphi_j(\theta)T_{n,j}(\mathbf{x})-nB(\theta)\right]$$

où $T_{n,j}(x_1,\ldots,x_n)=\sum_{i=1}^{n}T_j(x_i)$, et donc la loi du n-échantillon appartient encore à une famille exponentielle de dimension d, indépendamment du nombre n d'échantillons.

Exemple IV-4.4 (Loi de Poisson). Soit $(\mathbb{P}_\theta,\theta\in\Theta=\mathbb{R}_+^*)$ la distribution de la loi de Poisson de moyenne θ. Nous avons pour $x\in\mathsf{X}=\mathbb{N}$, l'ensemble des entiers naturels,

$$p(\theta,x)=\frac{\theta^x\mathrm{e}^{-\theta}}{x!}=\frac{1}{x!}\exp\left[x\log(\theta)-\theta\right],\quad \theta\in\Theta,$$

et donc le modèle de Poisson est une famille exponentielle de dimension $d=1$ avec

$$\varphi_1(\theta)=\log(\theta),\quad B(\theta)=\theta,\quad T(x)=x,\quad h(x)=1/x!.$$

La paramétrisation canonique est donnée par

$$q(\eta;x)=\frac{1}{x!}\exp\left[x\eta-\exp(\eta)\right]$$

et l'ensemble des paramètres naturels est donné par

$$\left|\log\left(\sum_{x=0}^{\infty}\frac{1}{x!}\exp(x\eta)\right)\right|=\exp(\eta)<\infty$$

♢

Exemple IV-4.5 (Loi binômiale). Soit $\{\mathbb{P}_\theta,\theta\in\Theta=]0,1[\}$ la distribution de la loi binômiale sur $\mathsf{X}=\{0,\ldots,n\}$. Pour $\theta\in\Theta$ et $x\in\{0,\ldots,n\}$, nous avons

$$p(\theta,x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}=\binom{n}{x}\exp\left[x\log\left(\frac{\theta}{1-\theta}\right)+n\log(1-\theta)\right].$$

Par conséquent, la loi binômiale est une famille exponentielle de dimension $d=1$ et

$$\varphi_1(\theta)=\log\left(\frac{\theta}{1-\theta}\right),\quad B(\theta)=-n\log(1-\theta),\quad T(x)=x,\quad h(x)=\binom{n}{x}.$$

♢

Exemple IV-4.6 (Famille gaussienne). Soit $(\mathbb{P}_\theta,\theta=(\mu,\sigma^2)\in\Theta=\mathbb{R}\times\mathbb{R}_+^*)$ la distribution de la loi gaussienne de moyenne μ et de variance σ^2. Nous avons,

$$p(\theta,x)=\exp\left(\frac{\mu}{\sigma^2}x-\frac{x^2}{2\sigma^2}-\frac{1}{2}\left\{\frac{\mu^2}{\sigma^2}+\log(2\pi\sigma^2)\right\}\right),$$

et $\mathbb{P}_\theta$ est une famille exponentielle spécifiée par

$$\begin{aligned}&\varphi_1(\theta)=\frac{\mu}{\sigma^2},\quad T_1(x)=x,\quad \varphi_2(\theta)=-\frac{1}{2\sigma^2},\quad T_2(x)=x^2\\&B(\theta)=\frac{1}{2}\left(\frac{\mu^2}{\sigma^2}+\log(2\pi\sigma^2)\right),\quad h(x)=1.\end{aligned}$$

♢

Exemple IV-4.7 (IV-4.6, suite). Dans ce cas, $d=2$, $\mathbf{T}(x)=(x,x^2)$, $\eta_1=\mu/\sigma^2$ et $\eta_2=-1/2\sigma^2$, $A(\eta)=(1/2)((-\eta_1^2/2\eta_2)+\log(\pi/-\eta_2))$. Par suite, $\mathscr{E}=\mathbb{R}\times(\mathbb{R}^-\setminus\{0\})$. ♢

On appelle *sous-modèle* de la famille canonique de dimension d associée aux statistiques $\mathbf{T}$ et à la fonction h des familles paramétriques de loi de la forme

$$p(\theta,x)=q(\varphi(\theta),x),$$

où $\varphi:\Theta\subset\mathbb{R}^m\mapsto\mathbb{R}^d$, $m\leq d$. Le modèle binomial ou le modèle gaussien sont des exemples de sous-modèles. Cette famille est appelée *courbe* si $\phi(\Theta)$ est inclus dans un sous-espace de dimension $l\leq(d-1)$.

IV-4.0.1 Quelques propriétés de la famille exponentielle

Lemme IV-4.8. *Soit $\{q(\eta,\cdot),\eta\in\Xi\}$ la famille exponentielle canonique de dimension d engendrée par les statistiques $\mathbf{T}$ et la fonction h (voir Définition IV-4.1). Alors*

(i) L'espace des paramètres naturel Ξ est convexe.

(ii) La fonction $A:\Xi\to\mathbb{R}$ définie par (IV-4.2) est convexe.

Démonstration. Soient $\eta=(\eta_1,\dots,\eta_d)$ et $\tilde\eta=(\eta_1,\dots,\eta_d)\in\Xi$. L'inégalité de Hölder montre que, pour toutes fonctions positives u, v, h mesurables et $\lambda\in]0,1[$, nous avons :

$$\int u^\lambda(x)v^{1-\lambda}(x)h(x)\mu(\mathrm{d}x)\le\left(\int u(x)h(x)\mu(\mathrm{d}x)\right)^\lambda\left(\int v(x)h(x)\mu(\mathrm{d}x)\right)^{1-\lambda}.$$

En appliquant cette inégalité avec $u(x)=\exp(\eta^T\mathbf{T}(x))$ et $v(x)=\exp(\tilde\eta^T\mathbf{T}(x))$, nous obtenons

$$\int\exp\left(\{\lambda\eta+(1-\lambda)\tilde\eta\}^T\mathbf{T}(x)\right)h(x)\mu(\mathrm{d}x)\\ \le\left(\int\exp\left(\eta^T\mathbf{T}(x)\right)h(x)\mu(\mathrm{d}x)\right)^\lambda\left(\int\exp\left(\tilde\eta^T\mathbf{T}(x)\right)h(x)\mu(\mathrm{d}x)\right)^{1-\lambda},$$

ce qui montre que l'espace des paramètres naturels est convexe. En prenant le logarithme des membres de gauche et de droite de l'inégalité précédente, nous obtenons, pour tout $\lambda\in]0,1[$:

$$A(\lambda\eta+(1-\lambda)\tilde\eta)\le\lambda A(\eta)+(1-\lambda)A(\tilde\eta)\,.$$ □

Théorème IV-4.9. *Soit ϕ une fonction mesurable telle que, pour tout $(\eta_1,\dots,\eta_d)\in\Xi$,*

$$\int|\phi(x)|\exp\left(\sum_{j=1}^k\eta_jT_j(x)\right)\mu(\mathrm{d}x)<\infty\,. \tag{IV-4.5}$$

Alors la fonction des variables complexes $(s_1,\dots,s_d)\in\mathbb{C}^d$

$$(s_1,\dots,s_d)\mapsto\Phi(s_1,\dots,s_d)=\int\phi(x)\exp\left(\sum_{j=1}^k s_jT_j(x)\right)\mu(\mathrm{d}x)$$

est définie en tous points $(s_1,\dots,s_d)$ tels que $(\mathrm{Re}(s_1),\dots,\mathrm{Re}(s_d))\in\Xi$. De plus

(i) La fonction Φ est analytique sur le domaine

$$R=\left\{(s_1,\dots,s_d)\in\mathbb{C}^d\,:\,(\mathrm{Re}(s_1),\dots,\mathrm{Re}(s_d))\in\Xi^o\right\}$$

(ii) Pour tout $p\in\mathbb{N}$, et tout d-uplet $(i_1,\dots,i_d)$ d'entiers naturels vérifiant $i_1+\cdots+i_d=p$, nous avons

$$\int|\phi(x)||T_1^{i_1}(x)\dots T_d^{i_d}(x)|\exp\left(\sum_{j=1}^k\mathrm{Re}(s_j)T_j(x)\right)\mu(\mathrm{d}x)<\infty$$

$$\frac{\partial^p\Phi(s_1,\dots,s_d)}{\partial s_1^{i_1}\dots\partial s_d^{i_d}}=\int\phi(x)T_1^{i_1}(x)\dots T_d^{i_d}(x)\exp\left(\sum_{j=1}^k s_jT_j(x)\right)\mu(\mathrm{d}x)\,.$$

Démonstration. Nous démontrons le résultat pour $p=1$, $i_1=1$ et $i_2=\cdots=i_d=0$. Soit $(s_1^0,\ldots,s_d^0)\in R$. En décomposant le facteur $\phi(x)\exp\left(s_2^0T_2(x)+\cdots+s_k^0T_k(x)\right)$ en partie réelle et imaginaire puis chacune de celles-ci en parties positives et négatives, nous pouvons réécrire

$$\Phi(s_1,s_2^0,\ldots,s_d^0)=\int\exp\left(s_1T_1(x)\right)\mu_1(\mathrm{d}x)-\int\exp\left(s_1T_1(x)\right)\mu_2(\mathrm{d}x)$$
$$+\mathrm{i}\int\exp\left(s_1T_1(x)\right)\mu_3(\mathrm{d}x)-\mathrm{i}\int\exp\left(s_1T_1(x)\right)\mu_4(\mathrm{d}x)\ .$$

Il suffit donc d'établir le résultat pour une intégrale de la forme

$$\psi(s_1)=\int\exp\left(s_1T_1(x)\right)\mu(\mathrm{d}x)\ .$$

Comme $(\mathrm{Re}(s_1^0),\ \ldots\ ,\mathrm{Re}(s_d^0))\in\Xi^o$, on peut choisir $\delta>0$ tel que $\psi(s_1)$ existe et est fini pour tout s_1 tel que $|s_1-s_1^0|\leq\delta$. Considérons la différence

$$\frac{\psi(s_1)-\psi(s_1^0)}{s_1-s_1^0}=\int\frac{\exp(s_1T_1(x))-\exp(s_1^0T_1(x))}{s_1-s_1^0}\mu(\mathrm{d}x)\ .$$

On peut écrire l'intégrande sous la forme

$$\exp(s_1^0T_1(x))\left[\frac{\exp[(s_1-s_1^0)T_1(x)]-1}{s_1-s_1^0}\right]$$

En appliquant l'inégalité

$$\left|\frac{\exp(az)-1}{z}\right|\leq\frac{\exp(\delta|a|)}{\delta}\quad\text{pour }|z|\leq\delta,$$

nous pouvons borner l'intégrande par

$$\frac{1}{\delta}|\exp(s_1^0T_1(x)+\delta|T_1(x)|)|\leq\frac{1}{\delta}|\exp[(s_1^0+\delta)T_1(x)]+\exp[(s_1^0-\delta)T_1(x)]|$$

pour $|s_1-s_1^0|\leq\delta$. Comme le terme de droite est intégrable, le théorème de convergence dominé montre que pour toute suite de points $\{s_1^{(n)},n\in\mathbb{N}\}$ convergent vers s_1^0, $\{\psi(s_1^{(n)})-\psi(s_1^{(0)})\}/(s_1^{(n)}-s_1^{(0)})$ tend vers

$$\int T_1(x)\exp\left(s_1^0T_1(x)\right)\mu(\mathrm{d}x)\ .$$

Le raisonnement pour les dérivées d'ordre plus élevé est exactement similaire. □

Dans la suite, nous notons, pour toute fonction mesurable $S:\mathsf{X}\to\mathbb{R}$ et $\eta\in\Xi$ vérifiant $\int|S(x)|q(\eta,x)\mu(\mathrm{d}x)<\infty$

$$Q_\eta(S)=\int S(x)q(\eta,x)\mu(\mathrm{d}x)\tag{IV-4.6}$$

Comme conséquence immédiate du résultat précédent, nous avons :

Corollaire IV-4.10. *(i) Pour tout $\eta=(\eta_1,\ldots,\eta_d)\in\Xi^o$ et $i,j\in\{1,\ldots,d\}$, nous avons*

$$Q_\eta(T_i)=\frac{\partial A}{\partial\eta_i}(\eta)$$
$$Q_\eta\{(T_i-Q_\eta(T_i))(T_j-Q_\eta(T_j))\}=\frac{\partial^2A}{\partial\eta_i\partial\eta_j}(\eta).$$

(ii) Pour tout $\eta=(\eta_1,\ldots,\eta_d)\in\Xi$ et tout $t=(t_1,\ldots,t_d)\in\mathbb{R}^d$ tel que $\eta+t=(\eta_1+t_1,\ldots,\eta_d+t_d)\in\Xi$, la fonction génératrice des moments des statistiques $\mathbf{T}$ est donnée par

$$M_\eta(t)=Q_\eta\left\{\exp\left[\mathbf{t}^T\mathbf{T}\right]\right\}=\exp\left(A(\eta+t)-A(\eta)\right)\ .$$

Définition IV-4.11 (Famille exponentielle de rang complet). *La famille exponentielle* $\{q(\eta,\cdot),\eta\in\Xi\}$ *engendrée par les statistiques* $\mathbf{T}$ *et* h *est de rang complet si et seulement si, pour tout* $\eta\in\Xi$*, la matrice* $H(\eta)=[H_{i,j}(\eta)]_{1\le i,j\le d}$ *est définie positive où*

$$H_{i,j}(\eta)=\left(\frac{\partial^2 A}{\partial\eta_i\partial\eta_j}(\eta)\right)_{1\le i,j\le d}. \tag{IV-4.7}$$

Théorème IV-4.12. *Soit* $\{q(\eta,\cdot),\eta\in\Xi\}$ *la famille exponentielle canonique engendrée par* $T_1,\dots,T_d$ *et* h*. Alors les assertions suivantes sont équivalentes.*

(i) La famille est de rang complet.

(ii) Pour tout $a_1,\dots,a_{d+1}$ *non identiquement nuls,*

$$Q_\eta\left(\left\{x\in\mathsf{X}:\sum_{i=1}^d a_iT_i(X)\neq a_{d+1}\right\}\right)=1,\quad \textit{pour tout } \eta\in\Xi^o. \tag{IV-4.8}$$

(iii) La fonction $\eta\to A(\eta)$ *est strictement convexe sur* Ξ^o*, i.e. pour tout* $\lambda\in]0,1[$*,* $\eta\in\Xi^o$*,* $\tilde\eta\in\Xi^o$*,*

$$A\left(\lambda\eta+(1-\lambda)\tilde\eta\right)<\lambda A(\eta)+(1-\lambda)A(\tilde\eta)$$

Remarque IV-4.13. Remarquons que pour $\eta,\tilde\eta\in\Xi$ et $A\in\mathscr{B}(\mathsf{X})$,

$$Q_\eta(A)=\int_A q(x;\eta)\mu(\mathrm{d}x)=\int_A q(x;\tilde\eta)\frac{\exp(\sum_{i=1}^d\eta_iT_i(x)-A(\eta))}{\exp(\sum_{i=1}^d\tilde\eta_iT_i(x)-A(\tilde\eta))}\mu(\mathrm{d}x)\,.$$

Par conséquent, si $\mathbb{P}_{\eta_0}(X\in A)=0$ (resp. 1) pour un $\eta_0\in\Xi$, alors $\mathbb{P}_{q(\eta,\cdot)}(X\in A)=0$ (resp. 1) pour tout $\eta\in\Xi$. De même, si pour $A\in\mathscr{B}(\mathsf{X})$ et $\eta_0\in\Xi$, $\mathbb{P}_{\eta_0}(X\in A)>0$, alors pour tout $\eta\in\Xi$, nous avons $\mathbb{P}_{q(\eta,\cdot)}(X\in A)>0$. En effet, la propriété découle de l'égalité précédente et du fait que

$$\frac{\exp(\sum_{i=1}^d\eta_iT_i(x)-A(\eta))}{\exp(\sum_{i=1}^d\tilde\eta_iT_i(x)-A(\tilde\eta))}>0,\quad \text{pour tout } x\in\mathsf{X},\ \eta,\tilde\eta\in\Xi.$$

Par conséquent, si la propriété (IV-4.8) est vérifiée pour $\eta_0\in\Xi$, alors elle est vérifiée pour tout $\eta\in\Xi$. ♢

Démonstration. (i) $\Longleftrightarrow$ (ii) Supposons qu'il existe $\mathbf{a}=(a_1,\dots,a_d)\in\mathbb{R}^d\setminus\{0\}$, $c\in\mathbb{R}$ et $\eta\in\Xi$ tel que :

$$\mathbb{P}_{q(\eta,\cdot)}\left(\sum_{i=1}^d a_iT_i(x)=c\right)=1.$$

Alors,

$$\mathrm{Var}_{q(\eta,\cdot)}\left(\mathbf{a}^T\mathbf{T}(X)\right)=\mathbf{a}^T\,\mathrm{Cov}_{q(\eta,\cdot)}\left(\mathbf{T}(X)\right)\mathbf{a}=0$$

et la famille exponentielle engendrée par $\mathbf{T}$ et h n'est pas de rang complet.
Réciproquement, si, pour $\mathbf{a}\neq 0$, $\mathrm{Var}_{q(\eta,\cdot)}(\mathbf{a}^T\mathbf{T}(X))=0$, alors, $\mathbb{P}_{q(\eta,\cdot)}(\mathbf{a}^T\mathbf{T}(X)=c)=1$, avec $c=Q_\eta[\mathbf{a}^T\mathbf{T}]$, et (IV-4.8) n'est pas vérifiée.

(i) $\Longleftrightarrow$ (iii) Si pour tout $\eta\in\Xi$ la matrice $\left(\partial^2A(\eta)/\partial\eta_i\partial\eta_j\right)_{1\le i,j\le d}$, est définie positive, alors la fonction A est strictement convexe. Supposons maintenant que (i) n'est pas vérifiée : il existe $\mathbf{a}\in\mathbb{R}^d\setminus\{0\}$

et $\eta_0 \in \Xi$ tel que $\mathrm{Var}_{q(\eta_0,\cdot)}(\mathbf{a}^T\mathbf{T}(X)) = 0$, ou, de façon équivalente, il existe $c \in \mathbb{R}$ tel que,

$$\begin{aligned}\mathbb{P}_{q(\eta_0,\cdot)}(\mathbf{a}^T\mathbf{T}(X) = c) = 1 &\Leftrightarrow \mathbb{P}_{q(\eta_0,\cdot)}(\lambda\mathbf{a}^T\mathbf{T}(X) = \lambda c) = 1, \quad \text{pour tout } \lambda \in [0,1],\\ &\Leftrightarrow \int h(x)\exp(\lambda\mathbf{a}^T\mathbf{T}(x))\exp({\eta_0}^T\mathbf{T}(X) - A(\eta_0))\mu(\mathrm{d}x) = \exp(\lambda c),\\ &\Leftrightarrow \int h(x)\exp(\eta_0 + \lambda\mathbf{a}^T\mathbf{T}(x))\mu(\mathrm{d}x) = \exp(\lambda c + A(\eta_0)).\end{aligned}$$

Donc, pour tout $\lambda \in [0,1]$, $\eta_0 + \lambda\mathbf{a} \in \Xi$ et

$$A(\eta_0 + \lambda\mathbf{a}) = A(\eta_0) + \lambda c\ . \tag{IV-4.9}$$

La fonction A n'est donc pas strictement convexe et (iii) n'est pas vérifiée. □

Théorème IV-4.14. *Soit $\{q(\eta,\cdot), \eta \in \Xi\}$ la famille exponentielle canonique engendrée par $T_1,\ldots,T_d$ et h. Le modèle exponentiel engendré par $\mathbf{T}$ et h est de rang complet si et seulement si l'application*

$$(\eta_1,\ldots,\eta_d) \mapsto \mathbf{e}(\eta_1,\ldots\eta_d) = \left(\frac{\partial A}{\partial\eta_1}(\eta_1,\ldots,\eta_d),\ldots,\frac{\partial A}{\partial\eta_d}(\eta_1,\ldots,\eta_d)\right) \tag{IV-4.10}$$

définit un difféomorphisme de Ξ^o sur $\mathbf{e}(\Xi^o)$. Pour tout $\eta \in \Xi^o$, la matrice jacobienne de ce difféomorphisme est donnée par

$$[\mathrm{J}_\mathbf{e}(\eta)]_{i,j} = \mathrm{Cov}_{q(\eta,\cdot)}(T_i(X),T_j(X)) = \frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\eta)\ , \quad 1 \le i,j \le d\ . \tag{IV-4.11}$$

Démonstration. Supposons que le modèle exponentiel engendré par $(\mathbf{T},h)$ est de rang complet. Le théorème IV-4.12-(iii) implique que la fonction $\eta \mapsto A(\eta)$ est strictement convexe. Elle est donc injective (en effet pour tout $\eta \neq \tilde\eta \in \Xi$ et $\lambda \in]0,1[$, on a $A(\lambda\eta + (1-\lambda)\tilde\eta) < \lambda A(\eta) + (1-\lambda)A(\tilde\eta)$ et donc $A(\eta) \neq A(\tilde\eta)$). L'équation (IV-4.11) découle de la définition de $\mathbf{e}$ et du corollaire IV-4.10. L'expression (IV-4.11) montre que la matrice jacobienne $\mathrm{J}_\mathbf{e}(\eta)$ est inversible et, comme $\mathbf{e}$ est injective, le théorème B.3 montre que $\mathbf{e}$ définit un difféomorphisme de Ξ sur $\mathbf{e}(\Xi)$.

Supposons maintenant que le modèle engendré par $(\mathbf{T},h)$ ne soit pas de rang complet. D'après le Théorème IV-4.12, il existe $\mathbf{a} \neq 0 \in \mathbb{R}^d$, $c \in \mathbb{R}$ et $\eta_0 \in \Xi$ tel que $\mathbb{P}_{q(\eta_0,\cdot)}(\mathbf{a}^T\mathbf{T}(X) = c) = 1$. Donc $\mathbb{P}_{q(\eta_0,\cdot)}(\mathbf{a}^T\mathbf{T}(X) = c) = 1$ pour tout $\eta \in \Xi$. En utilisant la remarque IV-4.13, ceci implique que, pour tout $\eta \in \Xi$ et $\lambda \in [0,1]$, $\mathbb{P}_{q(\eta,\cdot)}(\lambda\mathbf{a}^T\mathbf{T}(X) = \lambda c) = 1$. Par conséquent, en utilisant (IV-4.9), on obtient

$$A(\eta + \lambda\mathbf{a}) = \lambda c + A(\eta) \quad \Rightarrow \quad \mathbf{e}(\eta + \lambda\mathbf{a}) = \mathbf{e}(\eta)\ ,$$

ce qui montre que la fonction $\eta \to \mathbf{e}(\eta)$ n'est pas bijective. □

Exemple IV-4.15 (Vecteur gaussien p-dimensionnel). Soit $\mathbf{Y} = (Y_1,\ldots,Y_p)$ un vecteur Gaussien de moyenne $\boldsymbol{\mu}$ et de matrice de covariance Σ. Lorsque Σ est définie positive, alors, la variable $\mathbf{Y}$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$ donnée par

$$f(\mathbf{Y};\mu,\Sigma) = \det(\Sigma)^{-1/2}(2\pi)^{-p/2}\exp\left(-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right).$$

Par conséquent

$$\log f(\mathbf{Y};\boldsymbol{\mu},\Sigma) = -\frac{1}{2}\mathbf{Y}^T\Sigma^{-1}\mathbf{Y} + (\Sigma^{-1}\boldsymbol{\mu})^T\mathbf{Y} - \frac{1}{2}\left(\log(\det(\Sigma)) + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right) - \frac{p}{2}\log(2\pi),$$

ce qui montre que la famille de loi $\mathrm{N}(\mu,\Sigma)$ est une famille exponentielle $p(p+3)/2$-dimensionnelle, engendrée par la statistique

$$\mathbf{T}(\mathbf{Y}) = \left(\{Y_i\}_{1\le i\le p}, \{Y_iY_j\}_{1\le i\le j\le p}\right),$$

et la fonction $h(\mathbf{Y}) = 1$. La famille canonique associée à $(\mathbf{T}, h)$ est donnée par

$$q(\mathbf{X};\boldsymbol{\beta}) = \exp\left(\boldsymbol{\alpha}^T\mathbf{X} + \mathbf{X}^T\Gamma\mathbf{X} - A(\boldsymbol{\alpha},\boldsymbol{\gamma})\right),$$

où $\boldsymbol{\beta} = (\boldsymbol{\alpha},\boldsymbol{\gamma})$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_p)$, $\boldsymbol{\gamma} = \{\gamma_{i,j}\}_{1\le i\le j\le p}$ et Γ est la matrice $p\times p$ symétrique donnée par $\Gamma_{i,j} = \Gamma_{j,i} = \gamma_{i,j}$. Rappelons que toute matrice Γ symétrique est diagonalisable dans une base orthogonale, i.e. il existe une matrice unitaire U et une matrice diagonale Λ, telle que $\Gamma = U\Lambda U^T$. Notons que

$$\boldsymbol{\alpha}^T\mathbf{y} + \mathbf{y}^T\Gamma\mathbf{y} = (U\boldsymbol{\alpha})^T U\mathbf{y} + (U\mathbf{y})^T\Lambda(U\mathbf{y}).$$

Le changement de variable $y \mapsto U\mathbf{y}$ montre que

$$\int_{\mathbb{R}^p} \exp(\boldsymbol{\alpha}^T\mathbf{y} + \mathbf{y}^T\Gamma\mathbf{y})\mathrm{d}\mathbf{y} = \prod_{i=1}^p \int_{\mathbb{R}} \exp(\tilde{\alpha}_i z_i + \lambda_i z_i^2)\mathrm{d}z_i,$$

où $\tilde{\alpha}_i = (U\alpha)_i$, $i\in\{1,\ldots,p\}$ et λ_i est le i-ème élément diagonal de Λ. Ce produit d'intégrale est fini si et seulement si $\max_{1\le i\le p}\lambda_i < 0$, et par conséquent l'ensemble des paramètres canoniques est donné par :

$$\Xi = \left\{(\boldsymbol{\alpha},\boldsymbol{\gamma}) \in \mathbb{R}^p \times \mathbb{R}^{p(p+1)/2}, \quad \Gamma \quad \text{est définie négative}\right\}. \tag{IV-4.12}$$

L'ensemble Ξ est ouvert et nous pouvons donc appliquer le théorème IV-4.12. Nous allons établir que $\mathbf{T}$ est une statistique de rang $p(p+3)/2$. A cette fin, nous allons tout d'abord montrer que, pour $\mathbf{Z} = (Z_1,\ldots,Z_p)$ p v.a. gaussiennes i.i.d. centrées réduites, nous avons

$$\mathbb{P}_{\mathrm{N(0,I)}}\left(\mathbf{a}^T\mathbf{Z} + \mathbf{Z}^T B\mathbf{Z} = c\right) = 1, \quad \mathbf{a}\in\mathbb{R}^p, c\in\mathbb{R}, B\in\mathbb{R}^{p\times p},$$

si et seulement si $\|\mathbf{a}\| + \|B\| = 0$, où $\|B\|^2 = \mathrm{Tr}\left(BB^T\right)$. Remarquons en effet que, pour $1\le i\le j\le k\le l\le p$,

$$\mathrm{Cov}_{\mathrm{N(0,I)}}(Z_i,Z_j) = \delta_{i,j}, \quad \mathrm{Cov}_{\mathrm{N(0,I)}}(Z_i,Z_jZ_l) = 0, \quad \mathrm{Cov}_{\mathrm{N(0,I)}}(Z_iZ_j,Z_kZ_l) = \delta_{i,k}\delta_{j,l} + \delta_{i,l}\delta_{j,k},$$

et donc que :

$$\mathrm{Var}_{\mathrm{N(0,I)}}\left(\mathbf{a}^T\mathbf{Z} + \mathbf{Z}^T B\mathbf{Z}\right) = \|\mathbf{a}\|^2 + \|B\|^2.$$

ce qui établit la relation désirée. Nous avons, pour tout vecteur $\mathbf{a}\in\mathbb{R}^p$ et toute matrice $B\in\mathbb{R}^{p\times p}$,

$$\mathrm{Var}_{q(\boldsymbol{\beta},\cdot)}\left(\mathbf{a}^T\mathbf{Y} + \mathbf{Y}^T B\mathbf{Y}\right) = \mathrm{Var}\left(\tilde{\mathbf{a}}^T\mathbf{Z} + \mathbf{Z}^T\tilde{B}\mathbf{Z}\right),$$

avec $\tilde{\mathbf{a}} = (-\Lambda)^{-1/2}U\mathbf{a}$ et $\tilde{B} = U(-\Lambda)^{-1/2}B(-\Lambda)^{-1/2}U^T$, et le résultat précédent montre que

$$\begin{aligned}\mathrm{Var}_{q(\boldsymbol{\beta},\cdot)}\left(\mathbf{a}^T\mathbf{Y} + \mathbf{Y}^T B\mathbf{Y}\right) = 0 &\Leftrightarrow \tilde{\mathbf{a}} = 0 \quad \text{et} \quad \tilde{B} = 0,\\ &\Leftrightarrow \mathbf{a} = 0 \quad \text{et} \quad B = 0.\end{aligned}$$

Le théorème IV-4.12 montre que la famille $\mathbf{T}$ est de rang $p(p+3)/2$. ♢

sectionEstimateur des moments et maximum de vraisemblance pour la famille exponentielle

IV-4.0.2 Famille exponentielle canonique

Considérons la famille exponentielle canonique $\{q(\eta,\cdot),\eta\in\Xi\}$ engendrée par $(\mathbf{T},h)$. Supposons que cette famille soit de rang complet (voir la Définition IV-4.11). Nous observons un n-échantillon $X_1,\ldots,X_n$ i.i.d. de cette loi. L'estimateur des moments basé sur les statistiques $\mathbf{T}$ est défini comme la solution (si elle existe) du système de d équations à d inconnus

$$\hat{\mathbb{P}}_n\mathbf{T} = \frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i) = Q_\eta(\mathbf{T}) := \int \mathbf{T}(x)q(\eta,x)\mu(\mathrm{d}x)\,.$$

En utilisant le corollaire IV-4.10, ce système d'équations s'écrit de façon équivalente

$$\hat{\mathbb{P}}_n\mathbf{T} = \nabla A(\boldsymbol{\eta})\,. \tag{IV-4.13}$$

Comme nous l'avons établi dans le théorème IV-4.14, la fonction $\mathbf{e}$ définit un difféomorphisme de Ξ^o sur $\mathbf{e}(\Xi^o)$. La loi forte des grands nombres montre que, pour $\eta_0 \in \Xi^o$,

$$\lim_{n\to\infty} \mathbb{P}_{q(\eta_0,\cdot)} \left(\limsup_{n\to\infty} \hat{\mathbb{P}}_n \mathbf{T} \in \Xi^o \right) = 1 \,. \tag{IV-4.14}$$

D'autre part, sur l'événement $\{\hat{\mathbb{P}}_n \mathbf{T} \in \mathbf{e}(\Xi^o)\}$, nous avons

$$\hat{\eta}_n = \mathbf{e}^{-1}(\hat{\mathbb{P}}_n \mathbf{T}) \,. \tag{IV-4.15}$$

Nous prolongeons $\mathbf{e}^{-1}$ sur Ξ^c en posant $\mathbf{e}^{-1}(t) = \mathbf{0}_{d\times 1}$. En combinant le théorème II-1.9 et le théorème IV-4.14, nous obtenons

(i) La suite d'estimateurs $\{\eta_n,\, n \in \mathbb{N}\}$ est fortement consistante, pour tout $\eta_0 \in \Xi^o$, $\hat{\eta}_n \overset{\mathbb{P}_{\eta_0}-\text{p.s.}}{\longrightarrow} \eta_0$

(ii) La suite d'estimateurs $\{\eta_n,\, n \in \mathbb{N}\}$ est asymptotiquement normale, pour tout $\eta_0 \in \Xi^o$,

$$\sqrt{n}\,(\hat{\eta}_n - \eta_0) \overset{\mathbb{P}_{\eta_0}}{\Longrightarrow} \mathrm{N}(0, \{\mathrm{Var}_{q(\eta_0,\cdot)}(\mathbf{T}(X))\}^{-1}) \,. \tag{IV-4.16}$$

On remarque que dans ce cas particulier que, pour tout $\eta \in \Xi^o$,

$$\mathrm{J}_{\mathbf{e}}(\eta) = \mathrm{Var}_{q(\eta,\cdot)}(\mathbf{T}(X))$$

ce qui explique la forme particulièrement simple de la covariance limite dans ce cas particulier. Nous verrons dans la suite comment ce résultat se généralise. Nous allons maintenant montrer que l'estimateur $\hat{\eta}_n$ est aussi un estimateur de maximum de *vraisemblance*. La vraisemblance des observations est donnée par

$$\eta \mapsto q(\eta, X_1, \ldots, X_n) = \prod_{i=1}^n q(\eta, X_i) = \prod_{i=1}^n h(X_i) \cdot \exp\left(\sum_{i=1}^n \boldsymbol{\eta}^T \mathbf{T}(X_i) - nA(\eta) \right)$$

Posons $\ell_n(\eta, x) = \log q(\eta, x)$. Le théorème IV-4.9 montre que la *log-vraisemblance* est indéfiniment différentiable sur Ξ^o. Un calcul élémentaire montre que

$$\nabla \ell(\eta, x) = \mathbf{T}(x) - \nabla A(\eta) \,. \tag{IV-4.17}$$

Nous appelons dans la suite *score de Fisher* le gradient de la log-vraisemblance. En appliquant le corollaire IV-4.10, nous avons, pour tout $\eta \in \Xi^o$,

$$\mathbb{E}_{q(\eta,\cdot)}\left[\nabla \ell(\eta, X_1)\right] = Q_\eta\left(\mathbf{T}\right) - \nabla A(\eta) = \mathbf{0}_{d\times 1} \,.$$

Le système d'équations (IV-4.13) définissant l'estimateur des moments peut aussi s'interpréter comme la solution du système d'*équations de vraisemblance* donné par

$$n^{-1} \sum_{i=1}^n \nabla \ell(\eta, X_i) = 0 \,. \tag{IV-4.18}$$

Comme la fonction $\eta \mapsto A(\eta)$ est strictement convexe sur Ξ^o, la solution de (IV-4.18) (lorsqu'elle existe), correspond à l'unique maximum de la fonction de vraisemblance. L'estimateur des moments basés sur $\mathbf{T}$ coïncide pour la famille exponentielle canonique avec l'*estimateur de maximum de vraisemblance*.

L'équation (IV-4.16) montre que l'estimateur du maximum de vraisemblance est asymptotiquement normal. Nous appellerons *matrice d'information de Fisher* la covariance du score de Fisher

$$\mathbb{I}(\eta_0) = Q_{\eta_0}\left(\dot{\ell}_\eta \dot{\ell}_\eta^T\right) \,. \tag{IV-4.19}$$

En utilisant (IV-4.17), l'information de Fisher est donnée par

$$\mathbb{I}(\eta_0) = \mathrm{Var}_{q(\eta_0,\cdot)}(\mathbf{T}(X_1)) \,. \tag{IV-4.20}$$

L'équation (IV-4.16) montre que la covariance asymptotique de l'estimateur du maximum de vraisemblance est, dans ce cas, donnée par l'inverse de la matrice d'information de Fisher. Ces propriétés se généralisent à des modèles paramétriques réguliers (Section II-3.2).

IV-4.0.3 Famille exponentielle générale

Considérons le cas d'une famille exponentielle générale, dont la densité par rapport à la mesure de domination μ est donnée par (Définition IV-4.3).

$$p(\theta,x) = h(x)\exp\left[\varphi(\theta)^T\mathbf{T}(x) - B(\theta)\right], \quad x \in \mathsf{X}, \theta \in \Theta, \tag{IV-4.21}$$

où $\varphi = (\varphi^{(1)},\ldots,\varphi^{(d)})$ et $\mathbf{T} = (T_1,\ldots,T_d)$. Notons Ξ l'espace des paramètres naturels (voir Définition IV-4.1) Supposons que l'espace des paramètres Θ est un ouvert de $\mathbb{R}^d$ et que $\varphi = (\varphi_1,\ldots,\varphi_d) : \Theta \mapsto \Xi$, est un difféomorphisme de Θ sur Ξ^o. Dans ce cas, on peut passer des paramètres $\theta = (\theta^{(1)},\ldots,\theta^{(d)})$ aux paramètres canoniques $\eta = (\eta_1,\ldots,\eta_d)$ par un changement de variable "régulier", i.e. $\eta = \varphi(\theta)$ et $\theta = \varphi^{-1}(\eta)$. En particulier, pour tout $\theta \in \Theta$, $B(\theta) = A(\varphi(\theta))$ où la fonction A est définie par (IV-4.2).

L'estimateur des moments associés aux statistiques $\mathbf{T}$ est donc donné par

$$\hat{\theta}_n = \varphi^{-1}(\hat{\eta}_n) \tag{IV-4.22}$$

où $\hat{\eta}_n$ est l'estimateur des moments dans le modèle canonique (IV-4.15). Comme $\hat{\eta}$ est aussi l'estimateur du maximum de vraisemblance pour le modèle canonique et φ est un changement de variable régulier, la proposition I-2.18 montre que $\hat{\theta}_n$ est aussi l'estimateur du maximum de vraisemblance associé au modèle exponentiel général (IV-4.21). La log-vraisemblance de ce modèle est définie ici par

$$\theta \mapsto \ell_n(\theta) = n^{-1}\sum_{i=1}^n \log h(X_i) + n^{-1}\sum_{i=1}^n \varphi(\theta)^T\mathbf{T}(X_i) - A(\varphi(\theta)),$$

et l'estimateur du maximum de vraisemblance (confondu ici avec l'estimateur des moments) est la solution du système d'équations

$$\sum_{i=1}^n \nabla\ell(\theta,X_i) = \mathbf{0}_{d\times 1},$$

où $\theta \to \ell(\theta,x)$ est la log-vraisemblance d'une observation, donnée par

$$\ell(\theta,x) = \log p(\theta,x) = \log(h(x)) - \varphi(\theta)^T\mathbf{T}(x) - A(\varphi(\theta)) \tag{IV-4.23}$$

Si le modèle canonique est de rang complet, nous pouvons déduire de (IV-4.16), en appliquant la méthode-δ, la distribution limite de $\sqrt{n}(\hat{\theta}_n - \theta_0)$, où $\hat{\theta}_n$ est défini par (IV-4.22) :

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \overset{\mathbb{P}_{\theta_0}}{\Longrightarrow} \mathrm{N}\left(0, \{\mathrm{J}_\varphi(\theta_0)\}^{-1}\mathbb{I}^{-1}(\varphi(\theta_0))\{\mathrm{J}_\varphi(\theta_0)\}^{-T}\right), \tag{IV-4.24}$$

où

(i) $\mathbb{I}(\eta)$ est la matrice d'information de Fisher du modèle canonique (IV-4.19) évaluée au point η,

(ii) $\mathrm{J}_\varphi(\theta_0)$ est la matrice jacobienne de φ évaluée au point θ_0.

Le score de Fisher (le gradient de la log-vraisemblance (IV-4.23)) est donné, pour tout $\theta \in \Theta$, par

$$\nabla\ell(\theta,x) = \mathrm{J}_\varphi(\theta)\left\{\mathbf{T}(x) - \nabla A(\varphi(\theta))\right\}. \tag{IV-4.25}$$

Notons que, comme pour le modèle exponentiel canonique, l'espérance du score est nulle :

$$\int \nabla\ell(\theta,x)p(\theta,x)\mu(\mathrm{d}x) = 0,$$

car nous avons, en utilisant le corollaire IV-4.10

$$\int \mathbf{T}(x)p(\theta,x)\mu(\mathrm{d}x) = \int \mathbf{T}(x)q(\varphi(\theta),x)\mu(\mathrm{d}x) = A(\varphi(\theta)).$$

La matrice de Fisher du modèle exponentiel, défini comme la matrice de covariance du score est donnée par

$$\mathbb{J}(\theta) = \int \nabla\ell(\theta,x)\nabla\ell(\theta,x)^T p(\theta,x)\mu(\mathrm{d}x). \tag{IV-4.26}$$

En utilisant (IV-4.25), nous avons donc

$$\mathbb{J}(\theta) = \{\mathrm{J}_{\varphi}(\theta)\}\mathbb{I}(\phi(\theta))\{\mathrm{J}_{\varphi}(\theta)\}^T \ .$$

Par conséquent, (IV-4.27) montre que nous avons

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \stackrel{\mathbb{P}_{\theta_0}}{\Longrightarrow} \mathrm{N}\left(0, \mathbb{J}^{-1}(\theta)\right) \ , \tag{IV-4.27}$$

la variance limite de l'estimateur du maximum de vraisemblance est, dans ce cas aussi, l'inverse de la matrice d'information de Fisher.

Théorème IV-4.16. *Soit $\{p(\theta,\cdot),\in\Theta\}$ le modèle exponentiel engendré par $(\mathbf{T},h)$. Supposons que Θ soit un ouvert de $\mathbb{R}^d$ et que la fonction φ définisse un difféomorphisme de Θ^o sur l'ensemble des paramètres naturels Ξ. Supposons de plus que le modèle canonique associé soit de rang complet. Alors l'estimateur $\hat{\theta}_n$ des moments, basés sur les statistiques $\mathbf{T}$, coïncide avec l'estimateur du maximum de vraisemblance. De plus,*

(i) la suite d'estimateurs $\{\hat{\theta}_n,\, n\in\mathbb{N}\}$ est fortement consistante, i.e. pour tout $\theta_0\in\Theta$, $\hat{\theta}_n \stackrel{\mathbb{P}_{\theta_0}-\mathrm{p.s.}}{\longrightarrow} \theta_0$

(ii) la suite d'estimateurs $\{\hat{\theta}_n,\, n\in\mathbb{N}\}$ est asymptotiquement normale, i.e. pour tout $\theta_0\in\Theta$, $\hat{\theta}_n \stackrel{\mathbb{P}_{\theta_0}-\mathrm{p.s.}}{\longrightarrow} \theta_0$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \stackrel{\mathbb{P}_{\theta_0}}{\Longrightarrow} \mathrm{N}(0, \mathbb{J}^{-1}(\theta_0))$$

où $\mathbb{J}(\theta_0)$ est la matrice d'information de Fisher définie par (IV-4.26).

Chapitre IV-5

Modes de convergence et théorèmes limites

Nous rappelons dans ce chapitre les principaux modes de convergence des variables aléatoires, dont nous donnons des illustrations statistiques.

IV-5.1 Convergence en probabilité

Définition IV-5.1 (Convergence en Probabilité). *Soient* $\{X_n,\, n \in \mathbb{N}\}$ *une suite de vecteurs aléatoires à valeurs dans* $\mathbb{R}^d$ *et* X *un vecteur aléatoire à valeurs dans* $\mathbb{R}^d$*, définis sur un même espace de probabilité* $(\Omega, \mathscr{F}, \mathbb{P})$*. Nous dirons que la suite* $\{X_n,\, n \in \mathbb{N}\}$ converge en probabilité *vers* X *et nous noterons*

$$X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X\ ,$$

si pour tout $\varepsilon > 0$*, nous avons :*

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

Exemple IV-5.2. Soient $\{X_k,\, k \in \mathbb{N}^*\}$ une suite de variables aléatoires indépendantes de Bernoulli, de probabilité de succès $p = 1/2$: pour tout $k \geq 1$, $\mathbb{P}(X_k = 1) = 1/2$ (succès) et $\mathbb{P}(X_k = 0) = 1/2$ (échec). Pour $n \in \mathbb{N}$, on note T_n le nombre de fois que, dans un tirage $(X_1, \ldots, X_n)$, un succès est suivi d'un échec. En notant $I_k = \mathbb{1}_{\{X_k=1, X_{k+1}=0\}}$, nous avons donc $T_n = \sum_{k=1}^{n-1} I_k$, et par conséquent $\mathbb{E}[T_n] = (n-1)/4$ et

$$\mathrm{Var}(T_n) = \sum_{i=1}^{n-1} \mathrm{Var}(I_i) + 2\sum_{i=1}^{n-2} \mathrm{Cov}(I_i,\, I_{i+1}) = \frac{3(n-1)}{16} - \frac{2(n-2)}{16} = \frac{n+1}{16}\ .$$

En appliquant l'inégalité de Bienayme-Tchebychev (Lemme IV-1.2), nous avons, pour tout $\delta > 0$,

$$\mathbb{P}\left(|T_n - (n-1)/4| \geq \delta\right) \leq \frac{n+1}{16\delta^2}\ .$$

soit aussi

$$\mathbb{P}\left(|T_n/n - (1-1/n)/4| \geq \varepsilon\right) \leq \frac{n+1}{16\varepsilon^2 n^2} \to 0\ .$$

L'intuition est donc que $T_n/n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 1/4$. Montrons-le rigoureusement. En écrivant que

$$\begin{aligned}\{|T_n/n - 1/4| \geq \varepsilon\} &\subset \{|T_n/n - (n-1)/(4n)| + 1/(4n) \geq \varepsilon\} \\ &\subset \{|T_n/n - (n-1)/(4n)| \geq \varepsilon/2\} \bigcup \{1/(4n) \geq \varepsilon/2\}\end{aligned}$$

et puisque le dernier événement est toujours faux pour tout n assez grand, on en déduit que pour tout n assez grand,

$$\mathbb{P}(|T_n/n - 1/4| \geq \varepsilon) \leq \mathbb{P}(|T_n/n - (n-1)/(4n)| \geq \varepsilon/2) \leq \frac{(n-1)}{4\varepsilon^2 n^2}$$

ce qui montre que $n^{-1}T_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 1/4$. ♢

Proposition IV-5.3. *Soient $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, définis sur un espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. Supposons qu'il existe $p > 0$ tel que $\lim_{n\to\infty} \mathbb{E}[\|X_n - X\|^p] = 0$. Alors $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X$.*

Démonstration. Par l'inégalité de Markov (Lemme IV-1.1), pour tout $\varepsilon > 0$,

$$\mathbb{P}(\|X_n - X\| > \varepsilon) \leq \frac{\mathbb{E}[\|X_n - X\|^p]}{\varepsilon^p} \to 0\ .$$ □

Comme le montre la proposition suivante, la convergence en probabilité d'un vecteur aléatoire est équivalente à la convergence de ses coordonnées.

Proposition IV-5.4. *Soient $\{X_n,\ n \in \mathbb{N}\}$ et $\{Y_n,\ n \in \mathbb{N}\}$ deux suites de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et $\mathbb{R}^p$, X et Y deux vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et $\mathbb{R}^p$, définis sur le même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. $(X_n, Y_n) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} (X, Y)$ si et seulement si $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X$ et $Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} Y$.*

Démonstration. C'est une conséquence directe des inégalités

$$\|x_1 - y_1\| \leq \|(x_1, y_1) - (x_2, y_2)\| \leq \|x_1 - y_1\| + \|x_2 - y_2\|\ .$$ □

Proposition IV-5.5. *Soient $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, définis sur le même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. Supposons que $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X$. Alors, il existe une suite décroissante $\{\delta_n,\ n \in \mathbb{N}\}$ telle que $\lim_{n\to\infty} \delta_n = 0$ et vérifiant*

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

Démonstration. Pour tout $k \in \mathbb{N}^*$, il existe n_k tel que, pour tout $n \geq n_k$, $\mathbb{P}(\|X_n - X\| \geq 1/k) \leq 1/k$. Sans perte de généralité, on peut supposer que la suite $\{n_k,\ k \in \mathbb{N}^*\}$ est croissante. On pose $\delta_n = 1/k$ pour $n_k \leq n < n_{k+1}$. Il est clair que $\lim_{n\to\infty} \delta_n = 0$ et $\lim_{n\to\infty} \mathbb{P}(\|X_n - X\| \geq \delta_n) = 0$. □

Théorème IV-5.6. *Soient $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, définis sur le même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. Soit $g : \mathbb{R}^d \mapsto \mathbb{R}^m$ une fonction continue en tout point d'un ensemble C tel que $\mathbb{P}(X \in C) = 1$. Si $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X$, alors $g(X_n) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} g(X)$.*

Démonstration. Soit $\varepsilon > 0$. Pour tout $\delta > 0$, soit B_δ l'ensemble des points x tels qu'il existe y vérifiant $\|x-y\| \leq \delta$ et $\|g(x)-g(y)\| \geq \varepsilon$. Si $x \notin B_\delta$ et $\|g(x_n)-g(x)\| \geq \varepsilon$, alors $\|x_n - x\| \geq \delta$. Nous avons donc :

$$\mathbb{P}(\|g(X_n)-g(X)\| \geq \varepsilon) \leq \mathbb{P}(X \in B_\delta) + \mathbb{P}(\|X_n - X\| \geq \delta).$$

Le second terme du membre de droite tend vers 0 car $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X$. Nous avons $\mathbb{P}(X \in B_\delta \cap C^c) = 0$ et $\lim_{\delta \to 0} \mathbb{P}(X \in B_\delta \cap C) = 0$ par continuité de g. □

Corollaire IV-5.7. *Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ définis sur $(\Omega, \mathscr{F}, \mathbb{P})$. Soit $g : \mathbb{R}^d \mapsto \mathbb{R}^m$ une fonction continue au point $c \in \mathbb{R}^d$. Si $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} c$, alors $g(X_n) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} g(c)$.*

Démonstration. Il suffit d'appliquer le théorème IV-5.6 avec $X = c$. □

IV-5.2 Convergence presque-sûre

Définition IV-5.8 (Convergence presque-sûre). *Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, définis sur le même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. Nous dirons que la suite $\{X_n,\ n \in \mathbb{N}\}$* converge presque-sûrement *vers X et nous noterons*

$$X_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} X\ ,$$

si, pour tout $\varepsilon > 0$,

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

Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et soit $\{A_n,\ n \in \mathbb{N}\}$ une suite d'éléments de $\mathscr{F}$. Formons pour chaque $n \geq 1$, le sous-ensemble

$$B_n = \bigcup_{p=n}^{\infty} A_p.$$

Lorsque n croît, B_n décroît, et on peut donc parler de la limite *décroissante* de la suite $\{B_n,\ n \in \mathbb{N}\}$, égale, par définition, à $\bigcap_{n=1}^{\infty} B_n$. C'est cette limite qu'on appelle la *limite supérieure* de la suite $\{A_n,\ n \in \mathbb{N}\}$ (voir section A.1)

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{p=n}^{\infty} A_p.$$

On voit aisément que si $\omega \in \limsup_n A_n$ alors il existe une infinité d'indices $\{p_k(\omega), k \in \mathbb{N}\}$ tels que $\omega \in A_{p_k(\omega)}$. En particulier, lorsque $A_n = \{\omega \in \Omega : \|X_n(\omega) - X(\omega)\| > \varepsilon\}$, on a

$$A_n = \limsup_n \{\|X_n - X\| > \varepsilon\} = \{\limsup_n \|X_n - X\| > \varepsilon\}. \tag{IV-5.1}$$

Proposition IV-5.9. *Soient $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, définis sur le même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. La suite de vecteurs aléatoires $\{X_n,\ n \in \mathbb{N}\}$ converge presque-sûrement vers le vecteur aléatoire X si et seulement si il existe un événement $\Omega_0 \in \mathscr{F}$ de probabilité 1 tel que, pour tout $\omega \in \Omega_0$, $\lim_{n \to \infty} X_n(\omega) = X(\omega)$.*

Démonstration. La condition est suffisante. Pour tout $\omega \in \Omega_0$, le nombre d'indices n tels que $|X_n(\omega) - X(\omega)| \geq \varepsilon$ est fini. On déduit, en utilisant (IV-5.1),

$$\limsup_n \{\|X_n - X\| \geq \varepsilon\} = \{\limsup_n \|X_n - X\| \geq \varepsilon\} \subseteq \Omega_0^c \,.$$

La condition est nécessaire. Pour tout $k \in \mathbb{N}^*$, introduisons les événements

$$N_k = \limsup_n \{\|X_n - X\| \geq 1/k\} = \{\limsup_n \|X_n - X\| \geq 1/k\}.$$

Par hypothèse, $\mathbb{P}(N_k) = 0$ pour tout $k \geq 1$. Si on définit $N = \bigcup_{k=1}^{\infty} N_k$, on a bien $\mathbb{P}(N) = 0$. Reste à montrer que si $\omega \notin N$ alors $\lim X_n(\omega) = X(\omega)$ ou encore, de façon équivalente : si $\omega \notin N$, pour tout $\varepsilon > 0$, il n'existe qu'un nombre fini de n tels que $\|X_n(\omega) - X(\omega)\| \geq \varepsilon$. En effet, choisissons $k \geq 1$ tel que $\varepsilon > 1/k$. Comme $\omega \notin N$ et que $N_k \subset N$, alors $\omega \notin N_k$. D'après la définition de N_k, $\|X_n(\omega) - X(\omega)\| \geq 1/k$ n'a lieu que pour un nombre fini de n. A fortiori $\|X_n(\omega) - X(\omega)\| \geq \varepsilon$ n'a lieu que pour un nombre fini de n. □

Comme le montre le résultat ci-dessous, la convergence presque-sûre implique la convergence en probabilité.

Théorème IV-5.10. *Soient $\{X_n,\ n \in \mathbb{N}\}$ et X des variables aléatoires à valeurs dans $\mathbb{R}^d$, définies sur le même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. Si $X_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} X$ alors $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X$.*

Démonstration. Soit $\varepsilon > 0$. Posons $A_n := \bigcup_{k \geq n} \{\|X_k - X\| \geq \varepsilon\}$. On a

$$\{\|X_n - X\| \geq \varepsilon\} \subseteq A_n \,, \qquad \lim_n A_n = \bigcap_n A_n.$$

Par suite, et en utilisant le théorème de convergence monotone (voir Théorème A.37), il vient

$$\lim_n \mathbb{P}(\|X_n - X\| \geq \varepsilon) \leq \lim_n \mathbb{P}(A_n) = \mathbb{P}(\lim_n A_n) = \mathbb{P}\left(\bigcap_n A_n\right) .$$

Puisque $X_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} X$, $\mathbb{P}\left(\bigcap_n A_n\right) = 0$, ce qui conclut la preuve. □

Nous aurons souvent recours au Lemme de Borel-Cantelli qui donne une condition suffisante maniable pour établir que $\mathbb{P}(\limsup_n A_n) = 0$.

Lemme IV-5.11 (Lemme de Borel-Cantelli). *Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et $\{A_n,\ n \in \mathbb{N}\}$ une suite d'éléménts de $\mathscr{F}$. Alors :*

$$\sum_{n \geq 0} \mathbb{P}(A_n) < \infty \Rightarrow \mathbb{P}\left(\limsup_n A_n\right) = 0. \tag{IV-5.2}$$

Démonstration. Remarquons que $\left\{\bigcup_{p \geq n} A_p, n \in \mathbb{N}\right\}$ est une suite décroissante d'événements dont la limite est $\limsup_n A_n$, et l'on a

$$\mathbb{P}\left(\limsup_n A_n\right) = \lim \downarrow \mathbb{P}\left(\bigcup_{p \geq n} A_p\right) .$$

Or $\mathbb{P}\left(\bigcup_{p \geq n} A_p\right) \leq \sum_{p \geq n} \mathbb{P}(A_p)$ qui tend vers zero quand $n \to \infty$ par hypothèse, ce qui montre (IV-5.2). □

Le lemme de Borel-Cantelli permet d'établir la condition suffisante de convergence presque-sûre suivante.

Lemme IV-5.12. *Soient $\{X_n,\, n\in\mathbb{N}\}$ une suite de vecteurs aléatoires à valeur dans $\mathbb{R}^d$ et X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, définis sur le même espace de probabilité $(\Omega,\mathscr{F},\mathbb{P})$. Soit $\{\varepsilon_n,\, n\in\mathbb{N}\}$ une suite numérique convergeant vers 0. Si*

$$\sum_{n=0}^{\infty}\mathbb{P}(\|X_n-X\|\geq\varepsilon_n)<\infty,$$

alors $X_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} X$.

Démonstration. D'après le lemme de Borel-Cantelli (voir Lemme IV-5.11), il existe un ensemble $N\in\mathscr{F}$ de probabilité nulle tel que, pour tout $\omega\notin N$, $\|X_n(\omega)-X(\omega)\|\geq\varepsilon_n$ seulement pour un nombre fini d'indices. Donc, à partir d'un certain rang $N(\omega)$, $\|X_n(\omega)-X(\omega)\|<\varepsilon_n$. □

Théorème IV-5.13. *Soient $\{X_n,\, n\in\mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, définis sur le même espace de probabilité $(\Omega,\mathscr{F},\mathbb{P})$. Soit $g:\mathbb{R}^d\mapsto\mathbb{R}^m$ une fonction continue en tout point d'un ensemble C tel que $\mathbb{P}(X\in C)=1$. Si $X_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} X$, alors $g(X_n) \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} g(X)$.*

Démonstration. On note $\Omega_0=\{\omega\in\Omega:\ X(\omega)\in C\}$ et

$$\Omega_1=\left\{\omega\in\Omega:\ \limsup_n X_n(\omega)=\liminf_n X_n(\omega)=X(\omega)\right\}.$$

Sous les hypothèses du théorème, $\mathbb{P}(\Omega_0)=\mathbb{P}(\Omega_1)=1$, ce qui entraine $\mathbb{P}(\Omega_0\cap\Omega_1)=1$. Or, pour tout $\omega\in\Omega_0\cap\Omega_1$, nous avons $\lim_{n\to\infty}g(X_n(\omega))=g(X(\omega))$. Ce qui conclut la preuve d'après la Proposition IV-5.9. □

IV-5.3 Loi des grands nombres

Dans de nombreux cas, la consistance d'un estimateur découle simplement de la loi des grands nombres. Cette loi se démontre de façon élémentaire sous l'hypothèse que les variables concernées sont décorrelées : c'est l'objectif de la section IV-5.3.1. Dans la IV-5.3.2, nous montrons une version plus forte de la convergence en loi, qui permet d'établir la consistance forte de certains estimateurs.

IV-5.3.1 Loi faible des grands nombres

Définition IV-5.14 (Variables décorrélées). *Soient $X_1,\ldots,X_n$, des variables aléatoires définies sur $(\Omega,\mathscr{F},\mathbb{P})$ et de carré intégrable, $\mathbb{E}[X_k^2]<\infty$ pour tout $k\in\{1,\ldots,n\}$. Les variables aléatoires $X_1,\ldots,X_n$ sont décorrélées si, pour tout $1\leq k\neq\ell\leq n$,*

$$\mathrm{Cov}(X_k,X_\ell):=\mathbb{E}[(X_k-\mathbb{E}[X_k])(X_\ell-\mathbb{E}[X_\ell])]=0\,.$$

Le principal intérêt des variables décorrélées est donné par le lemme suivant.

Lemme IV-5.15 (Lemme de Bienaymé). *Soient $X_1,\dots,X_n$, des variables aléatoires définies sur $(\Omega,\mathscr{F},\mathbb{P})$ et de carré intégrable, $\mathbb{E}[X_k^2]<\infty$ pour tout $k\in\{1,\dots,n\}$. Supposons de plus que les variables aléatoires $X_1,\dots,X_n$ sont décorrélées. Alors,*

$$\mathrm{Var}\left(\sum_{k=1}^n X_k\right)=\sum_{k=1}^n \mathrm{Var}(X_k)\,.$$

Démonstration. Un calcul immédiat montre que

$$\mathrm{Var}\left(\sum_{k=1}^n X_k\right)=\sum_{k=1}^n \mathbb{E}\left[(X_k-\mathbb{E}[X_k])^2\right]+\sum_{1\le k\neq \ell\le n}\mathbb{E}\left[(X_k-\mathbb{E}[X_k])(X_\ell-\mathbb{E}[X_\ell])\right]=\sum_{k=1}^n \mathrm{Var}(X_k)\,.$$

□

Ce lemme permet d'établir, par des arguments élémentaires, une loi des grands nombres, attribuée à Markov.

Théorème IV-5.16 (Loi faible des grands nombres pour des variables décorrélées). *Soit $\{X_n,\ n\in\mathbb{N}\}$ une suite de variables aléatoires de carré intégrables : pour tout $k\in\mathbb{N}$, $\mathbb{E}[X_k^2]<\infty$. Supposons que*

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{k=1}^n \mathrm{Var}(X_k)=0\,,$$

et que les variables aléatoires $\{X_k,\ k\in\mathbb{N}\}$ sont décorrélées. Alors

$$\lim_{n\to\infty}\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^n\{X_k-\mathbb{E}[X_k]\}\right)^2\right]=0\,, \tag{IV-5.3}$$

et

$$\frac{1}{n}\sum_{k=1}^n\{X_k-\mathbb{E}[X_k]\}\overset{\mathbb{P}-\text{prob}}{\longrightarrow}0\,. \tag{IV-5.4}$$

Démonstration. En utilisant le lemme de Bienaymé (Lemme IV-5.15), nous avons

$$\mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^n\{X_k-\mathbb{E}[X_k]\}\right)=\frac{1}{n^2}\sum_{k=1}^n\mathrm{Var}(X_k)\,,$$

ce qui montre (IV-5.3). La relation (IV-5.4) est une conséquence de la proposition IV-5.3. □

Corollaire IV-5.17 (Loi faible pour des variables i.i.d. de carré intégrable). *Soit $\{Y_n,\ n\in\mathbb{N}\}$ une suite de variables définies sur $(\Omega,\mathscr{F},\mathbb{P})$, i.i.d. et de carré intégrable. Alors*

$$\bar{Y}_n:=\frac{1}{n}\sum_{i=1}^n Y_i\overset{\mathbb{P}-\text{prob}}{\longrightarrow}\mathbb{E}[Y_1]\,.$$

On peut affaiblir la condition de moment du corollaire IV-5.17 en tronquant les variables aléatoires. Le résultat est important, mais sa preuve est plus technique et peut être omise en première lecture.

Théorème IV-5.18 (Loi faible des grands nombres pour des variables i.i.d. intégrables). *Soit* $\{X_k,\, k\in\mathbb{N}\}$ *des variables aléatoires définies sur* $(\Omega,\mathscr{F},\mathbb{P})$ *i.i.d. et intégrables. Alors*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \mathbb{E}[X_1]\,.$$

Démonstration. Sans perte de généralité (quitte à remplacer X_k par $X_k-\mathbb{E}[X_k]$), supposons que $\mathbb{E}[X_1]=0$. Soit $\varepsilon>0$. Posons $R_n:=\sum_{k=1}^n X_k$ et $T_n:=\sum_{k=1}^n X_k\mathbb{1}_{\{|X_k|\le n\varepsilon^3\}}$. Notons tout d'abord que, sur l'événement $\bigcap_{k=1}^n\{|X_k|\le n\varepsilon^3\}$, $R_n=T_n$. Par suite,

$$\begin{aligned}
&\mathbb{P}\left(|R_n-\mathbb{E}[T_n]|\ge n\varepsilon\right)\\
&=\mathbb{P}\left(|R_n-\mathbb{E}[T_n]|\ge n\varepsilon,\bigcap_{k=1}^{n}\{|X_k|\le n\varepsilon^3\}\right)+\mathbb{P}\left(|R_n-\mathbb{E}[T_n]|\ge n\varepsilon,\bigcup_{k=1}^{n}\{|X_k|> n\varepsilon^3\}\right)\\
&\le \mathbb{P}(|T_n-\mathbb{E}[T_n]|\ge n\varepsilon)+\mathbb{P}\left(\bigcup_{k=1}^{n}\{|X_k|>n\varepsilon^3\}\right)\,. \qquad \text{(IV-5.5)}
\end{aligned}$$

En utilisant l'inégalité de Bienayme-Tchebychev (Lemme IV-1.2), nous avons

$$\begin{aligned}
\mathbb{P}(|T_n-\mathbb{E}[T_n]|\ge n\varepsilon)&\le\frac{1}{n\varepsilon^2}\operatorname{Var}(X_1\mathbb{1}_{|X_1|\le n\varepsilon^3})\le\frac{1}{n\varepsilon^2}\mathbb{E}\left[X_1^2\mathbb{1}_{|X_1|\le n\varepsilon^3}\right]\\
&\le\frac{n\varepsilon^3}{n\varepsilon^2}\mathbb{E}[|X_1|]\le\varepsilon\mathbb{E}[|X_1|]\,. \qquad \text{(IV-5.6)}
\end{aligned}$$

D'autre part, nous avons

$$\mathbb{P}\left(\bigcup_{k=1}^{n}\{|X_k|>n\varepsilon^3\}\right)\le\sum_{k=1}^{n}\mathbb{P}(|X_k|>n\varepsilon^3)=n\mathbb{P}(|X_1|>n\varepsilon^3)\,.$$

En utilisant l'inégalité $\mathbb{1}_{\{|X_1|>n\varepsilon^3\}}\le(|X_1|/n\varepsilon^3)\mathbb{1}_{\{|X_1|>n\varepsilon^3\}}$, nous avons

$$n\mathbb{P}(|X_1|>n\varepsilon^3)\le\frac{1}{\varepsilon^3}\mathbb{E}[|X_1|\mathbb{1}_{|X_1|>n\varepsilon^3}]$$

et le théorème de convergence dominée implique

$$\limsup_{n\to\infty} n\mathbb{P}(|X_1|>n\varepsilon^3)=0\,.$$

Par conséquent, en combinant ce résultat avec (IV-5.6), nous obtenons

$$\limsup_{n\to\infty}\mathbb{P}\left(|R_n-\mathbb{E}[T_n]|\ge n\varepsilon\right)\le\varepsilon\mathbb{E}[|X_1|]\,. \qquad \text{(IV-5.7)}$$

Notons que

$$\frac{\mathbb{E}[T_n]}{n}=\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k\mathbb{1}_{\{|X_k|\le n\varepsilon^3\}}]=\mathbb{E}[X_1\mathbb{1}_{\{|X_1|\le n\varepsilon^3\}}]$$

et le théorème de convergence dominée implique que

$$\lim_{n\to\infty} n^{-1}\mathbb{E}[T_n]=\mathbb{E}[X_1]=0\,. \qquad \text{(IV-5.8)}$$

En utilisant (IV-5.7) et (IV-5.8), nous avons

$$\begin{aligned}
\limsup_n\mathbb{P}\left(\left|\frac{R_n}{n}\right|\ge\varepsilon\right)&\le\limsup_n\mathbb{P}\left(|R_n-\mathbb{E}[T_n]|\ge n\varepsilon/2\right)+\limsup_n\mathbb{P}\left(|\mathbb{E}[T_n]|\ge n\varepsilon/2\right)\\
&\le\varepsilon\mathbb{E}[|X_1|]\,,
\end{aligned}$$

ce qui conclut la preuve car ε est arbitraire. □

Nous établirons plus loin (voir Théorème IV-5.38) ce résultat en utilisant d'autres outils basés sur la fonction caractéristique.

Il est possible de généraliser le Théorème IV-5.18 à des variables qui ne sont pas intégrables. La seule chose que nous avons à supposer est que $\mathbb{E}[X_1^+] < \infty$ ou $\mathbb{E}[X_1^-] < \infty$.

Théorème IV-5.19 (Loi faible des grands nombres pour des variables i.i.d.). *Soit* $\{X_k,\ k \in \mathbb{N}\}$ *une suite de variables aléatoires définies sur le même espace* $(\Omega, \mathscr{F}, \mathbb{P})$*, i.i.d. et telles que* $\mathbb{E}[X_1^+] < \infty$ *ou* $\mathbb{E}[X_1^-] < \infty$*. Alors*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \mathbb{E}[X_1]\,.$$

Démonstration. Considérons le cas où $\mathbb{E}[X_1^-] < \infty$, l'autre cas se traitant de manière entièrement symétrique. Si $\mathbb{E}[X_1^+] < \infty$, alors la variable aléatoire X_1 est intégrable et le résultat découle de Théorème IV-5.18. Supposons donc que $\mathbb{E}[X_1^+] = \infty$. Notons que pour tout $c > 0$, nous avons

$$\mathbb{E}[|X_1 \wedge c|] \leq c + \mathbb{E}[X_1^-] < \infty\,.$$

En utilisant Théorème IV-5.18, nous obtenons donc que

$$\frac{1}{n}\sum_{k=1}^{n} X_k \wedge c \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \mathbb{E}[X_1 \wedge c]\,.$$

D'autre part, nous avons $n^{-1}\sum_{k=1}^{n}(X_k \wedge c) \leq n^{-1}\sum_{k=1}^{n} X_k$. Pour tout $M > 0$, comme $\mathbb{E}[X_1] = \infty$, nous pouvons choisir c tel que $\mathbb{E}[X_1 \wedge c] > M$. Nous avons donc

$$\lim_{n\to\infty} \mathbb{P}\left(n^{-1}\sum_{k=1}^{n} X_k \geq M\right) \geq \lim_{n\to\infty} \mathbb{P}\left(n^{-1}\sum_{k=1}^{n}(X_k \wedge c) \geq M\right) = 1.$$

Comme M est arbitraire, ceci montre que $n^{-1}\sum_{k=1}^{n} X_k \overset{\mathbb{P}-\text{prob}}{\longrightarrow} +\infty$. □

Exemple IV-5.20 (Loi de Cauchy). Soit $\{X_k,\ k \in \mathbb{N}\}$ une suite de variables aléatoires i.i.d. Pour que la loi des grands nombres soit satisfaite, il est indispensable de supposer que $\mathbb{E}[X_1^+] < \infty$ ou $\mathbb{E}[X_1^-] < \infty$. Nous allons voir dans cet exemple que si $\mathbb{E}[X_1^+] = \infty$ et $\mathbb{E}[X_1^-] = \infty$ alors la loi des grands nombres n'est pas vérifiée. Une loi de Cauchy de paramètre $\theta = (\alpha, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$ est une loi dont la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ est donnée par

$$p(\theta, x) = \frac{\sigma}{\pi(\sigma^2 + (x-\alpha)^2)}\,.$$

Les paramètres α et σ sont ici respectivement les paramètres de localisation et d'échelle. On montre aisément que la fonction caractéristique (Définition IV-5.34) d'une loi de Cauchy de paramètre $\theta = (\alpha, \sigma)$ est donnée par

$$\phi_\theta(t) = \exp(\mathrm{i}\alpha t - \sigma|t|)\,.$$

Cela entraine en particulier que, si X_1 suit une loi de Cauchy de paramètre (α, σ), alors λX_1 suit une loi de Cauchy de paramètres $(\lambda\alpha, |\lambda|\sigma)$. Soient Y_1 et Y_2 deux variables indépendantes distribuées suivant des lois de Cauchy de paramètres (α_1, σ_1) et (α_2, σ_2). La fonction caractéristique de la somme $Y = Y_1 + Y_2$ est donnée par

$$\exp(\mathrm{i}\alpha_1 t - \sigma_1|t|)\exp(\mathrm{i}\alpha_2 t - \sigma_2|t|) = \exp(\mathrm{i}(\alpha_1+\alpha_2)t - (\sigma_1+\sigma_2)|t|)$$

et donc Y est aussi distribuée suivant une loi de Cauchy de paramètre $(\alpha_1 + \alpha_2, \sigma_1 + \sigma_2)$. Supposons maintenant que $\{X_n,\ n \in \mathbb{N}\}$ soit une suite i.i.d. de variables aléatoires i.i.d. distribuées suivant une loi de Cauchy de paramètres $(0,1)$. La moyenne empirique $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ est aussi distribuée suivant une loi de Cauchy de paramètre $(0,1)$, et donc la moyenne empirique ne converge pas vers une constante. ♢

IV-5.3.2 Loi forte des grands nombres

La loi forte des grands nombres est un résultat essentiel en probabilité et en statistique. Nous donnons ici la version due à Kolmogorov (1933).

Théorème IV-5.21 (Loi forte des grands nombres). *Soit $\{X_n,\, n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur $(\Omega, \mathscr{F}, \mathbb{P})$, i.i.d. et intégrables. Alors*

$$\bar{X}_n := n^{-1} \sum_{i=1}^{n} X_i \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \mathbb{E}[X_1] \ .$$

La preuve de ce théorème est délicate et requiert l'utilisation d'outils probabilistes plus sophistiqués que ceux dont nous disposons. Nous donnons ci-dessous une version de la loi forte des grands nombres pour des variables identiquement distribuées (mais qui ne sont pas nécessairement indépendantes), de *variance finie* (l'intégrabilité suffit normalement) et *décorrélées*.

Théorème IV-5.22. *Soit $\{X_n,\, n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur $(\Omega, \mathscr{F}, \mathbb{P})$, intégrables, identiquement distribuées et de variance finie et non-corrélées. On a alors :*

$$n^{-1} \sum_{i=1}^{n} X_i \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \mathbb{E}[X_1] \ .$$

Démonstration. Nous allons utiliser l'inégalité de Markov (Lemme IV-1.1) et le lemme de Borel-Cantelli (Lemme IV-5.11). Sans perte de généralité, on suppose $\mathbb{E}[X_1] = 0$. Pour un entier positif m, on considère les variables aléatoires $\{Y_{m^2}, m \in \mathbb{N}\}$ et $\{Z_m, m \in \mathbb{N}\}$ où

$$Y_n := \sum_{i=1}^{n} X_i, \qquad Z_m := \sup_{1 \le k \le 2m+1} \left(|X_{m^2+1} + \ldots + X_{m^2+k}| \right) .$$

Admettons avoir les propriétés suivantes : lorsque m tend vers l'infini,

$$\frac{Y_{m^2}}{m^2} \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} 0 \tag{IV-5.9}$$

et

$$\frac{Z_m}{m^2} \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} 0 \ . \tag{IV-5.10}$$

On utilise la majoration suivante :

$$\left| \frac{Y_n}{n} \right| \le \left| \frac{Y_{m(n)^2}}{m(n)^2} \right| + \left| \frac{Z_{m(n)}}{m(n)^2} \right| ,$$

où $m(n)$ est l'entier m tel que

$$m^2 < n \le (m+1)^2 \ .$$

Comme $\lim_{n\to\infty} m(n) = +\infty$, on conclut en utilisant (IV-5.9) et (IV-5.10).

Prouvons les deux relations admises ; considérons (IV-5.9). Pour un $\varepsilon > 0$ quelconque, l'inégalité de Bienayme-Tchebychev donne

$$\mathbb{P}\left(|Y_{m^2}/m^2| \ge \varepsilon \right) \le \frac{\mathrm{Var}(Y_{m^2})}{m^4 \varepsilon^2} = \frac{m^2 \sigma^2}{m^4 \varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \frac{1}{m^2} \ .$$

On conclut en utilisant l'inégalité de Borel-Cantelli. Considérons maintenant (IV-5.10). Posons $|\xi_{m,k}| = |X_{m^2+1} + \cdots + X_{m^2+k}|$. Notons tout d'abord que

$$\mathbb{P}\left(\left|\frac{Z_m}{m^2}\right| \geq \varepsilon\right) = \mathbb{P}\left(|Z_m| \geq m^2\varepsilon\right) \leq \mathbb{P}\left(\bigcup_{k=1}^{2m+1} \{|\xi_{m,k}| \geq m^2\varepsilon\}\right) \leq \sum_{k=1}^{2m+1} \mathbb{P}(|\xi_{m,k}| \geq m^2\varepsilon)\,.$$

L'inégalité de Bienayme-Tchebychev nous donne d'autre part

$$\mathbb{P}(|\xi_{m,k}| \geq m^2\varepsilon) \leq \frac{\mathrm{Var}(\xi_{m,k})}{m^4\varepsilon^2} \leq \frac{(2m+1)\sigma^2}{m^4\varepsilon^2}$$

puisque $\xi_{m,k}$ est la somme de moins de $2m+1$ variables aléatoires. En combinant les trois dernières inégalités, on voit que

$$\mathbb{P}\left(\left|\frac{Z_m}{m^2}\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}\frac{(2m+1)^2}{m^4}\,.$$

On conclut en utilisant encore le Lemme de Borel-Cantelli. □

Exemple IV-5.23 (Corrélation empirique). Soit $\{(X_n, Y_n),\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires de $\mathbb{R}^2$, i.i.d. et de même loi que (X,Y). On suppose que $\mathbb{E}[X^2] < \infty$ et $\mathbb{E}[Y^2] < \infty$. Soit

$$\rho := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}\,,$$

le coefficient de corrélation de X et Y. En appliquant l'inégalité de Cauchy-Schwarz, nous établissons aisément que $|\rho| \leq 1$. Si $|\rho| = 1$ (cas d'égalité dans l'inégalité de Cauchy-Schwarz), il existe des constantes a, b, c telles que $a + bX + cY = 1 \quad \mathbb{P}-\text{p.s.}$ Une valeur du coefficient de corrélation proche de 1 implique une forte dépendance linéaire entre les variables X et Y. Considérons le *coefficient de corrélation empirique* défini par

$$R_n := \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}, \qquad \text{où } \bar{X}_n := \frac{1}{n}\sum_{k=1}^n X_k, \quad \bar{Y}_n := \frac{1}{n}\sum_{k=1}^n Y_k\,.$$

Nous allons démontrer que $R_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \rho$. Notons que

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

Par la loi forte des grands nombres, $C_n := n^{-1}\sum_{i=1}^n X_i Y_i \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \mathbb{E}[XY]$, $\bar{X}_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \mathbb{E}[X]$ et $\bar{Y}_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \mathbb{E}[Y]$. De façon similaire, on montre que $V_n := n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \mathrm{Var}(X)$ et $W_n := n^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \mathrm{Var}(Y)$. Considérons la fonction

$$f(s,t,\,u,\,v,\,w) = \frac{s - tu}{\sqrt{v}\sqrt{w}}\,, \quad -\infty < s,t,u < \infty, v, w > 0\,.$$

Cette fonction est continue sur l'ensemble

$$S = \{(s,\,t,\,u,\,v,\,w) : -\infty < s,t,u < \infty, v, w > 0, (s-tu)^2 \leq vw\}\,.$$

Nous avons clairement $\mathbb{P}((C_n, \bar{X}_n, \bar{Y}_n, V_n, W_n) \in S) = 1$. Le théorème de continuité (Théorème IV-5.13) montre que $R_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \rho$. ♢

IV-5.4 Convergence en loi

Dans cette section, nous introduisons les notations, définitions et résultats essentiels de la convergence en loi. Les résultats principaux seront présentés sans preuve, mais nous donnerons des illustrations de leur utilisation.

Définition IV-5.24 (Convergence étroite). *Soient $\{\mu_n,\ n\in\mathbb{N}\}$ une suite de probabilités sur $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$ et μ une probabilité sur $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$. Nous dirons que la suite $\{\mu_n,\ n\in\mathbb{N}\}$* converge étroitement *vers μ, ce que nous notons $\mu_n \overset{w}{\Rightarrow} \mu$, si, pour toute fonction continue bornée $f:\mathbb{R}^d\to\mathbb{R}$,*

$$\lim_{n\to\infty}\int f\mathrm{d}\mu_n = \int f\mathrm{d}\mu\ .$$

Définition IV-5.25 (Convergence en loi). *Pour tout $n\in\mathbb{N}$, soit $(\Omega_n,\mathscr{F}_n,\mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n,\mathscr{F}_n,\mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$. Soit $(\Omega,\mathscr{F},\mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega,\mathscr{F},\mathbb{P})$ et à valeurs dans $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$. Nous dirons que la suite $\{X_n,\ n\in\mathbb{N}\}$* converge en loi *vers X, ce que nous noterons $X_n \Longrightarrow X$ si la suite des mesures images de $\mathbb{P}_n$ par X_n, $\{\mathbb{P}_n^{X_n}, n\in\mathbb{N}\}$ converge étroitement vers la mesure image de $\mathbb{P}$ par X, $\mathbb{P}^X$ ou, de façon équivalente, si pour toute fonction continue bornée $f:\mathbb{R}^d\to\mathbb{R}$:*

$$\lim_{n\to\infty}\mathbb{E}_n[f(X_n)] = \mathbb{E}[f(X)]\ .$$

Dans la définition de la convergence en loi, contrairement aux définitions de convergence en probabilité ou presque-sûre, rien n'oblige à définir les vecteurs aléatoires $\{X_n,\ n\in\mathbb{N}\}$ et X sur le même espace de probabilité ; on s'intéresse uniquement à la convergence étroite des lois des vecteurs aléatoires X_n vers la loi de X.

Théorème IV-5.26 (Théorème de Portmanteau). *Pour tout $n\in\mathbb{N}$, soit $(\Omega_n,\mathscr{F}_n,\mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n,\mathscr{F}_n,\mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$. Soit $(\Omega,\mathscr{F},\mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega,\mathscr{F},\mathbb{P})$ et à valeurs dans $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$. Les assertions suivantes sont équivalentes.*

(i) $X_n \Longrightarrow X$,

(ii) Pour toute fonction f bornée et lipschitzienne, $\lim_{n\to\infty}\mathbb{E}_n[f(X_n)] = \mathbb{E}[f(X)]$,

(iii) $\liminf_{n\to\infty}\mathbb{P}_n(X_n\in G)\geq\mathbb{P}(X\in G)$ *pour tout ensemble ouvert G,*

(iv) $\limsup_{n\to\infty}\mathbb{P}_n(X_n\in F)\leq\mathbb{P}(X\in F)$ *pour tout ensemble F fermé,*

(v) Pour tout pavé $A=\prod_{i=1}^d]a_i,b_i]$ dont la frontière ∂A vérifie $\mathbb{P}(X\in\partial A)=0$ on a

$$\lim_{n\to\infty}\mathbb{P}_n(X_n\in A) = \mathbb{P}(X\in A)$$

(vi) Pour toute fonction continue et positive

$$\liminf_{n\to\infty}\mathbb{E}_n[f(X_n)]\geq\mathbb{E}[f(X)]\ .$$

Démonstration. Voir par exemple [**?**, Theorem 2.1]. □

En particulier, la suite $\{X_n,\ n\in\mathbb{N}\}$ converge en loi vers X si et seulement si

$$\lim_{n\to\infty}F_{X_n}(x_1,\ \ldots,\ x_d) = F_X(x_1,\ \ldots,\ x_d)$$

en tout point $(x_1,\ \ldots\ ,\ x_d)$ où F_X est continue.

Exemple IV-5.27 (un exemple élémentaire). Soit $\{X_n,\ n\in\mathbb{N}\}$ une suite de variables aléatoires définies sur le même espace de probabilité $(\Omega,\mathscr{F},\mathbb{P})$ telles que pour tout n, $\mathbb{P}(X_n=1/n)=1$: la loi de X_n est égale à $\delta_{1/n}$. Donc pour toute fonction continue bornée, $\lim_{n\to\infty}\mathbb{E}[f(X_n)]=\lim_{n\to\infty}f(1/n)=f(0)$. La loi limite est donc δ_0.

Les fonctions de répartition des variables aléatoires $X=0$ et X_n sont respectivement

$$F_X(x)=\begin{cases}0 & \text{si } x<0\\ 1 & \text{si } x\geq 0\end{cases}\quad,\qquad F_{X_n}(x)=\begin{cases}0 & \text{si } x<1/n\\ 1 & \text{si } x\geq 1/n\end{cases}\quad.$$

On voit que la fonction de répartition F_X est continue partout sauf en $x=0$. On peut vérifier que $\lim_n F_{X_n}(x)=F_0(x)$ pour tout $x>0$ et tout $x<0$; mais que $\lim_n F_{X_n}(0)\neq F_0(0)$. ♢

Exemple IV-5.28 (Sanity check !). Pour $n\geq 1$, soit X_n une variable aléatoire de loi uniforme et à valeurs dans $\{1/n,2/n,\cdots,(n-1)/n,1\}$. On a donc $\mathbb{P}(X_n=k/n)=1/n$ pour tout $1\leq k\leq n$. La fonction de répartition de la variable X_n est donnée, pour $x\in[0,1]$ par $F_{X_n}(x)=\lfloor nx\rfloor/n$. Or, pour tout $x\in[0,1]$, $\lim_{n\to\infty}F_n(x)=x$; et la fonction $x\mapsto x$ sur $[0,1]$ est la fonction de répartition d'une loi uniforme sur $[0,1]$. Par conséquent, le théorème IV-5.26 item (v) montre que $X_n\Longrightarrow \mathrm{Unif}([0,1])$.

On peut aussi prouver cette convergence en utilisant une autre caractérisation : pour toute fonction f continue sur $[0,1]$, nous avons

$$\mathbb{E}[f(X_n)]=\frac{1}{n}\sum_{k=1}^{n}f(k/n)\to_{n\to\infty}\int_0^1 f(x)\mathrm{d}x\,,$$

et le terme de droite se relit comme $\mathbb{E}[f(X)]$ où $X\sim\mathrm{Unif}([0,1])$. ♢

Le théorème suivant donne les liens entre la convergence en loi et la convergence en probabilité.

Théorème IV-5.29. *Soit $(\Omega,\mathscr{F},\mathbb{P})$ un espace de probabilité. Soient $\{X_n,\ n\in\mathbb{N}\}$ une suite de vecteurs aléatoires et X un vecteur aléatoire définis sur $(\Omega,\mathscr{F},\mathbb{P})$ et à valeurs dans $\mathbb{R}^d$. Alors*

(i) $X_n\overset{\mathbb{P}-\text{prob}}{\longrightarrow}X$ *implique* $X_n\Longrightarrow X$,

(ii) $X_n\overset{\mathbb{P}-\text{prob}}{\longrightarrow}c$, *où c est une constante, si et seulement si* $X_n\Longrightarrow c$.

Démonstration. (i) Soit $f:\mathbb{R}^d\to\mathbb{R}$ Lipschitzienne bornée. Notons

$$\|f\|_\infty:=\sup_x|f(x)|\quad\text{and}\quad|f|_{\mathrm{Lip}}:=\sup_{x\neq y}\frac{|f(x)-f(y)|}{\|x-y\|}\,.$$

Pour tout $\varepsilon>0$,

$$|\mathbb{E}[f(X_n)]-\mathbb{E}[f(X)]|\leq\varepsilon|f|_{\mathrm{Lip}}+2|f|_\infty\mathbb{P}(\|X_n-X\|\geq\varepsilon).$$

Le second terme du membre de droite tend vers 0 et le premier peut être rendu arbitrairement petit. Donc, $\lim_{n\to\infty}\mathbb{E}[f(X_n)]=\mathbb{E}[f(X)]$ pour toute fonction f Lipschitzienne bornée. Nous concluons en appliquant le théorème IV-5.26.

(ii) Soit $\varepsilon>0$ et soit $\mathrm{B}(c,\varepsilon):=\left\{x\in\mathbb{R}^d\ :\ \|x-c\|<\varepsilon\right\}$ la boule ouverte de centre c et de rayon ε. Nous avons $\mathbb{P}(\|X_n-c\|\geq\varepsilon)=\mathbb{P}(X_n\in\mathrm{B}(c,\varepsilon)^c)$. Si $X_n\Longrightarrow c$, le théorème IV-5.26 montre que $\limsup_n\mathbb{P}(X_n\in\mathrm{B}(c,\varepsilon)^c)\leq\mathbb{P}(X\in\mathrm{B}(c,\varepsilon)^c)=0$. □

IV-5.4.1 Opérations sur les limites en loi

Théorème IV-5.30 (Transformation continue). *Pour tout $n\in\mathbb{N}$, soit $(\Omega_n,\mathscr{F}_n,\mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n,\mathscr{F}_n,\mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$. Soit $(\Omega,\mathscr{F},\mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega,\mathscr{F},\mathbb{P})$ et à valeurs dans $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$. Soit $g:\mathbb{R}^d\mapsto\mathbb{R}^m$ une fonction continue en tout point d'un ensemble C tel que $\mathbb{P}(X\in C)=1$. Si $X_n\Longrightarrow X$ alors $g(X_n)\Longrightarrow g(X)$.*

Démonstration. Par définition, $\{g(X_n) \in F\} = \{X_n \in g^{-1}(F)\}$. Pour tout fermé F, on a :

$$g^{-1}(F) \subset \overline{g^{-1}(F)} \subset (g^{-1}(F) \cup C^c); \tag{IV-5.11}$$

C^c désigne le complémentaire de C. Seule la seconde inclusion est non-triviale ; soit $x \in \overline{g^{-1}(F)}$ et montrons que seulement deux cas sont possibles : soit $x \in C^c$, soit $x \in g^{-1}(F)$. Par définition, comme $x \in \overline{g^{-1}(F)}$, il existe une suite $\{x_n,\, n \in \mathbb{N}\}$ d'éléments de $g^{-1}(F)$ telle que $\lim_{m\to\infty} x_m = x$. Si $x \in C$, $g(x_m) \to g(x)$, car g est continue au point x, et comme $g(x_m) \in F$ et F est fermé, $g(x) \in F$, ce qui implique que $x \in g^{-1}(F)$.

On déduit de (IV-5.11) et du théorème de Portmanteau (Théorème IV-5.26), comme $X_n \Longrightarrow X$ et $\mathbb{P}(X \in C^c) = 0$:

$$\begin{aligned}\limsup \mathbb{P}(g(X_n) \in F) &\le \limsup \mathbb{P}(X_n \in \overline{g^{-1}(F)}) \le \mathbb{P}(X \in \overline{g^{-1}(F)}) \\ &\le \mathbb{P}(X \in g^{-1}(F) \cup C^c) = \mathbb{P}(X \in g^{-1}(F)) = \mathbb{P}(g(X) \in F),\end{aligned}$$

et, donc, en appliquant de nouveau le théorème de Portmanteau, $g(X_n) \Longrightarrow g(X)$. □

La proposition IV-5.4 montre que la convergence en probabilité d'une suite de vecteurs aléatoires $X_n = (X_{n,1}, \ldots, X_{n,k})$ est équivalente à la convergence de chacune de ses composantes. Le résultat analogue pour la convergence en loi est faux : la convergence en loi d'une suite de vecteurs aléatoires est une propriété plus forte que la convergence en loi de chacune de ses composantes $X_{n,i}$, comme le montre l'exemple suivant :

Exemple IV-5.31. Considérons le vecteur aléatoire (X_n, Y_n) défini de la façon suivante : pour tout n, X_n est distribuée suivant une loi normale centrée réduite et $Y_n = (-1)^n X_n$. Nous avons donc $X_n \Longrightarrow \mathrm{N}(0,1)$; et puisque pour tout n, Y_n suit aussi une loi normale centrée réduite, nous avons aussi $Y_n \Longrightarrow \mathrm{N}(0,1)$. Donc, chaque composante du vecteur (X_n, Y_n) converge en loi.

Si nous prenons $f(x,y) = \phi(x+y)$ où ϕ est une fonction continue bornée de $\mathbb{R}$ dans $\mathbb{R}$, nous avons $f(X_n, Y_n) = \phi(2X_n)$ lorsque n est pair et $f(X_n, Y_n) = \phi(0)$ sinon. Par suite, la suite $\{f(X_n, Y_n), n \in \mathbb{N}\}$ ne peut pas converger et en utilisant le Théorème IV-5.30, le vecteur (X_n, Y_n) ne peut pas converger en loi. ♢

Le théorème suivant clarifie les relations entre les différentes définitions de convergence et celles entre convergence d'un vecteur et convergence de ses coordonnées.

Théorème IV-5.32. *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n, Y_n des vecteurs aléatoires, définis sur $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ et à valeur respectivement dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ et $(\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p))$. Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathscr{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.*

(i) (cas $d = p$) Si $X_n \Longrightarrow X$ et $X_n - Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$, alors $Y_n \Longrightarrow X$.

(ii) Soit $c \in \mathbb{R}^p$. Si $X_n \Longrightarrow X$ et $Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} c$, alors $(X_n, Y_n) \Longrightarrow (X, c)$.

Démonstration. (i) Soit $f : \mathbb{R}^d \to \mathbb{R}$ Lipschitzienne bornée, de constante de Lipschitz notée $|f|_{\mathrm{Lip}}$. Pour tout $\varepsilon > 0$,

$$\begin{aligned}|\mathbb{E}_n[f(X_n)] - \mathbb{E}_n[f(Y_n)]| &= \left|\mathbb{E}_n\left[(f(X_n) - f(Y_n))\left(\mathbb{1}_{\|X_n - Y_n\| \le \varepsilon} + \mathbb{1}_{\|X_n - Y_n\| \ge \varepsilon}\right)\right]\right| \\ &\le |f|_{\mathrm{Lip}}\varepsilon + 2|f|_\infty \mathbb{P}_n(\|X_n - Y_n\| \ge \varepsilon)\,.\end{aligned}$$

Le second terme tend vers 0 puisque $X_n - Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$ (à ε fixé) et le premier peut être rendu arbitrairement petit en choisissant ε petit. Donc $\mathbb{E}_n[f(X_n)]$ et $\mathbb{E}_n[f(Y_n)]$ ont la même limite. Enfin, par deux applications du théorème IV-5.26, on a $\lim_n \mathbb{E}_n[f(X_n)] = \mathbb{E}[f(X)]$ et $Y_n \Longrightarrow X$.

(ii) Remarquons que

$$\left|\begin{bmatrix} X_n \\ Y_n \end{bmatrix} - \begin{bmatrix} X_n \\ c \end{bmatrix}\right| = \left|\begin{bmatrix} 0 \\ Y_n - c \end{bmatrix}\right| \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0\,.$$

En utilisant (i) et la décomposition

$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} - \begin{bmatrix} X \\ c \end{bmatrix} = \begin{bmatrix} X_n \\ Y_n \end{bmatrix} - \begin{bmatrix} X_n \\ c \end{bmatrix} + \begin{bmatrix} X_n \\ c \end{bmatrix} - \begin{bmatrix} X \\ c \end{bmatrix}$$

il suffit de prouver que $(X_n, c) \Longrightarrow (X, c)$ Pour toute fonction continue bornée $f : (x, y) \to f(x, y)$, la fonction $f(\cdot, c) : x \to f(x, c)$ est continue et bornée et $|\mathbb{E}[f(X_n, c)] - \mathbb{E}[f(X, c)]| \to 0$, car $X_n \Longrightarrow X$; par suite, $(X_n, c) \Longrightarrow (X, c)$. □

La propriété (ii) et le théorème de continuité (théorème IV-5.30) montrent que pour toute fonction $g : (x, y) \to g(x, y)$ continue sur un ensemble $C \times \{c\}$ où C est tel que $\mathbb{P}(X \in C) = 1$, on a $g(X_n, Y_n) \Longrightarrow g(X, c)$. Des applications particulières de ce principe sont souvent regroupées sous la forme du lemme suivant, connu sous le nom de lemme de Slutsky.

Lemme IV-5.33 (Lemme de Slutsky). *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n, Y_n des v.a. réelles, définies sur $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$. Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et X une v.a. réelle définie sur $(\Omega, \mathscr{F}, \mathbb{P})$.*

Si $X_n \Longrightarrow X$ et $Y_n \Longrightarrow c$ où c est une constante, alors

(i) $X_n + Y_n \Longrightarrow X + c$;

(ii) $Y_n X_n \Longrightarrow cX$;

(iii) Si $c \neq 0$, $Y_n^{-1} X_n \Longrightarrow c^{-1} X$.

Le procédé de Cramér-Wold (Théorème IV-5.37) permet d'étendre ce résultat au cas vectoriel/matriciel, pour peu que, dans *(i)*, c soit un vecteur de même dimension que X, et dans *(ii)* et *(iii)*, $\{Y_n,\ n \in \mathbb{N}\}$ et c soient des matrices (avec c inversible pour *(iii)*) de dimension adaptée à celle des vecteurs $\{X_n,\ n \in \mathbb{N}\}$.

IV-5.4.2 Caractérisation par la fonction caractéristique

Le théorème de Portmanteau (Théorème IV-5.26) montre que, pour établir la convergence en loi, il suffit de s'intéresser à un sous-ensemble des fonctions continues bornées, par exemple, les fonctions lipschitziennes bornées, mais cette classe peut encore être réduite. Nous allons en fait démontrer dans cette partie qu'il suffit de s'intéresser à une seule fonction, la *fonction caractéristique*.

Définition IV-5.34 (Fonction caractéristique). *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ défini sur $(\Omega, \mathscr{F}, \mathbb{P})$. La fonction caractéristique du vecteur aléatoire X est définie par*

$$t = (t_1, \ldots, t_d) \to \mathbb{E}[\mathrm{e}^{\mathrm{i}t^T X}] .$$

Exemple IV-5.35 (Fonction caractéristique d'une gaussienne multidimensionnelle). Supposons que le vecteur aléatoire $X \in \mathbb{R}^d$ est distribué suivant une loi normale d'espérance μ et de covariance Σ, ce que nous notons $\mathrm{N}(\mu, \Sigma)$. Alors X a même loi que $\mu + \sqrt{\Sigma} Y$ où $Y \sim \mathrm{N}(0, I)$, et $\sqrt{\Sigma}$ désigne une matrice $d \times d$ satisfaisant $\sqrt{\Sigma}\sqrt{\Sigma} = \Sigma$.

En dimension $d = 1$, la fonction caractéristique d'une loi $\mathrm{N}(0, 1)$ est donnée par (voir Lemme **??**) :

$$\mathbb{E}\left[\mathrm{e}^{\mathrm{i}tX}\right] = \exp\left(-\frac{1}{2}t^2\right) . \tag{IV-5.12}$$

Dans le cas $X \sim \mathrm{N}(\mu, \Sigma)$, remarquons que

$$\phi_X(t) = \mathbb{E}[\mathrm{e}^{\mathrm{i}t^T X}] = \mathbb{E}[\mathrm{e}^{\mathrm{i}t^T (\mu + \sqrt{\Sigma} Y)}] = \mathrm{e}^{\mathrm{i}t^T \mu}\ \mathbb{E}[\mathrm{e}^{\mathrm{i}\sqrt{\Sigma}^T t^T Y}] .$$

Or Y a ses composantes indépendantes de loi $\mathrm{N}(0,1)$ donc pour tout $t \in \mathbb{R}^d$

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}t^T Y}] = \prod_{k=1}^{d} \mathbb{E}[\mathrm{e}^{\mathrm{i}t_k^T Y_k}] = \exp\left(-\frac{1}{2}|t|^2\right) .$$

On en déduit que

$$\phi_X(t) = \exp\left(it^T\mu\right) \exp\left(-\frac{1}{2}t^T\Sigma t\right) . \qquad \diamondsuit$$

Pour tout $t \in \mathbb{R}^d$, la fonction $x \to \mathrm{e}^{\mathrm{i}t^T x}$ est continue et bornée. Par conséquent, si $X_n \Longrightarrow X$, $\mathbb{E}_n[\mathrm{e}^{\mathrm{i}t^T X_n}] \to \mathbb{E}[\mathrm{e}^{\mathrm{i}t^T X}]$. Le théorème de Levy montre que la réciproque est vraie.

Théorème IV-5.36 (Lévy). *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathscr{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Alors*

1. *$X_n \Longrightarrow X$ si et seulement si $\mathbb{E}_n[\exp(\mathrm{i}t^T X_n)] \to \mathbb{E}[\exp(\mathrm{i}t^T X)]$ pour tout $t \in \mathbb{R}^d$.*
2. *Si il existe une fonction ϕ, définie sur $\mathbb{R}^d$ et continue en 0, telle que, pour tout $t \in \mathbb{R}^d$, la suite $\{\mathbb{E}_n[\exp(\mathrm{i}t^T X_n)], n \in \mathbb{N}\}$ converge vers $\phi(t)$, alors ϕ est la fonction caractéristique d'un vecteur aléatoire X à valeurs dans $\mathbb{R}^d$ et $X_n \Longrightarrow X$.*

Démonstration. Voir par exemple [**?**, Théorème 26.3 et Corollaire 1]. □

La fonction caractéristique d'un vecteur aléatoire $X = (X_1, \ldots, X_d)$ en $t \in \mathbb{R}^d$ peut être vue comme la fonction caractéristique de la variable aléatoire $Y := t^T X$ évaluée au point 1 : $\psi(t) = \mathbb{E}[\mathrm{e}^{\mathrm{i}t^T X}] = \phi(1)$ où $\phi : u \in \mathbb{R} \to \mathbb{E}[\mathrm{e}^{\mathrm{i}uY}]$.

Cette observation, exploitée dans le théorème suivant, est très utile pour prouver la convergence en loi de vecteurs. Elle est connue sous le nom de *procédé de Cramér–Wold*. Elle permet de réduire les problèmes de convergence de vecteurs aléatoires à des problèmes de convergence de variables aléatoires.

Théorème IV-5.37 (Cramér-Wold). *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathscr{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. $X_n \Longrightarrow X$ si et seulement si, pour tout $t \in \mathbb{R}^d$, $t^T X_n \Longrightarrow t^T X$.*

Démonstration. Supposons que $X_n \Longrightarrow X$. Alors, pour tout $t \in \mathbb{R}^d$ et tout $u \in \mathbb{R}$, $\mathbb{E}_n[\mathrm{e}^{\mathrm{i}ut^T X_n}] \to \mathbb{E}[\mathrm{e}^{\mathrm{i}u(t^T X)}]$, et donc $t^T X_n \Longrightarrow t^T X$ par application du théorème de Levy (théorème IV-5.36).

Réciproquement, supposons que, pour tout $t \in \mathbb{R}^d$, $t^T X_n \Longrightarrow t^T X$. Alors $\mathbb{E}_n[\mathrm{e}^{\mathrm{i}(t^T X_n)}] \to \mathbb{E}[\mathrm{e}^{\mathrm{i}(t^T X)}]$ et donc $X_n \Longrightarrow X$, encore par application du théorème de Levy. □

Le Théorème IV-5.36 permet d'obtenir une autre preuve de la loi faible des grands nombres que celle donnée dans la démonstration du théorème IV-5.18.

Théorème IV-5.38 (Loi faible des grands nombres). *Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur $(\Omega, \mathscr{F}, \mathbb{P})$, i.i.d. et intégrables. Alors*

$$n^{-1} \sum_{i=1}^{n} X_i \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \mathbb{E}[X_1] .$$

Démonstration. On pose $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ et $\mu := \mathbb{E}[X_1]$. Notons $\psi_n(t)$ la fonction caractéristique de $\bar{X}_n$ et $\phi(t)$ la fonction caractéristique de X_1. Les variables aléatoires étant i.i.d., on a

$$\psi_n(t) = \mathbb{E}\left[\exp\left(\mathrm{i}tn^{-1}\sum_{k=1}^n X_k\right)\right] = \prod_{k=1}^n \mathbb{E}[\exp(\mathrm{i}tn^{-1}X_k)] = \left(\phi(n^{-1}t)\right)^n.$$

Comme $\mathbb{E}[|X_1|] < \infty$ existe, la fonction ϕ est dérivable en 0 et

$$\left[\phi(n^{-1}t)\right]^n = \left(1 + \frac{\mathrm{i}t}{n}\mu + o(n^{-1})\right)^n \quad \to_{n\to\infty} \quad \mathrm{e}^{\mathrm{i}t\mu}; \tag{IV-5.13}$$

(la déribabilité de la fonction ϕ s'établit par application du théorème de convergence dominée, voir Théorème A.40 et Proposition A.42 ; pour ceux qui ne sont pas familiers du passage à la limite (IV-5.13), on peut l'établir en utilisant les mêmes arguments que ceux de la démonstration du Théorème IV-5.39).

Dans (IV-5.13), le membre de droite est la fonction caractéristique de la variable aléatoire qui vaut μ avec probabilité 1. Le théorème de Levy IV-5.36 montre que $\bar{X}_n \Longrightarrow \mu$ et donc $\bar{X}_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \mu$ par le Théorème IV-5.29. □

IV-5.5 Théorème de la limite centrale

IV-5.5.1 T.L.C. pour des v.a. indépendantes et de même loi

Le théorème de la limite centrale (T.L.C.) donne des conditions sous lesquelles des sommes normalisées de v.a. indépendantes de moyenne nulle convergent en loi vers une gaussienne. Ce résultat joue un rôle majeur en statistique (voir Le Cam, 1986, pour une histoire de ce théorème).

Théorème IV-5.39 (TLC pour des v.a. i.i.d.). *Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires de $\mathbb{R}^d$, définis sur le même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$, i.i.d. et possédant un moment d'ordre 2. On note μ l'espérance et Σ la matrice de covariance. Alors :*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \Longrightarrow \mathrm{N}(0,\Sigma),$$

Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de variables aléatoires réelles i.i.d. d'espérance μ et de variance $\sigma^2 > 0$. On note F_n la fonction de répartition de $\sqrt{n}(\overline{X}_n - \mu)$ où $\overline{X}_n := n^{-1}\sum_{k=1}^n X_k$. On remarque que

$$\begin{aligned}\mathbb{P}\left(\mu - a/\sqrt{n} < \bar{X}_n \le \mu + a/\sqrt{n}\right) &= \mathbb{P}(-a < \sqrt{n}(\bar{X}_n - \mu) \le a) \\ &= F_n(a) - F_n(-a) \\ &\to \Phi(a/\sigma) - \Phi(-a/\sigma)\ ,\end{aligned}$$

où Φ est la fonction de répartition d'une variable gaussienne centrée réduite. L'information apportée par le théorème de la limite centrale précise le résultat donné par la loi faible des grands nombres.

Démonstration. (du Théorème IV-5.39) D'après le procédé de Cramér-Wold (Théorème IV-5.37), il suffit de montrer que pour tout $t \in \mathbb{R}^d$,

$$\sqrt{n}t^T Z_n \Longrightarrow t^T W, \qquad \text{où} \qquad W \sim \mathrm{N}(0,\Sigma),\ \ Z_n := n^{-1}\sum_{k=1}^n (X_k - \mu). \tag{IV-5.14}$$

Notons que Z_n est un vecteur aléatoire centré et de matrice de covariance Σ, de sorte que $\sqrt{n}t^T Z_n$ est une v.a. réelle centrée et de variance $t^T\Sigma t$.

Soit $t \in \mathbb{R}^d$. Si t est tel que $t^T \Sigma t = 0$, alors pour tout $n \geq 1$ $\sqrt{n} t^T Z_n$ est une v.a. réelle constante, égale à son espérance donc nulle. Par suite, $\sqrt{n} t^T Z_n \Longrightarrow 0$ ce qui est bien équivalent à (IV-5.14) puisque $t^T W$ est une v.a. gaussienne centrée de variance $t^T \Sigma t = 0$ (donc une v.a. constante égale à son espérance).

Dans la suite, on prend t tel que $\sigma^2 := t^T \Sigma t > 0$. Posons, pour tout $i = 1, \ldots, n$,

$$Y_i := \sigma^{-1}(X_i - \mu), \qquad \bar{Y}_n := n^{-1} \sum_{i=1}^{n} Y_i.$$

Les variables aléatoires $\{Y_n,\ n \in \mathbb{N}\}$ sont de moyenne nulle et de variance unité : $\mathbb{E}[Y_i] = 0$ et $\mathrm{Var}(Y_i) = 1$. Notons ϕ_n la fonction caractéristique associée à la variable aléatoire $\sqrt{n}\bar{Y}_n$, et ψ celle associée à la v.a. Y_1. Nous allons démontrer que, pour tout t :

$$\lim_{n \to \infty} \phi_n(t) = \exp(-t^2/2);$$

nous conclurons en utilisant le théorème de Lévy et le résultat (IV-5.12).

Les variables aléatoires $\{Y_n,\ n \in \mathbb{N}\}$ étant i.i.d. de même loi que Y (disons), nous avons, pour tout $t \in \mathbb{R}$:

$$\phi_n(t) = [\psi(n^{-1/2}t)]^n \, .$$

Nous allons montrer que pour tout $t \in \mathbb{R}$,

$$[\psi(n^{-1/2}t)]^n \to_{n \to \infty} \mathrm{e}^{-t^2/2} \, .$$

Dans la suite, $t \in \mathbb{R}$ est fixé. Remarquons tout d'abord que

$$\begin{aligned} \left| [\psi(n^{-1/2}t)]^n - \mathrm{e}^{-t^2/2} \right| &= \left| [\psi(n^{-1/2}t)]^n - [\mathrm{e}^{-t^2/(2n)}]^n \right| \\ &\leq n \left| \psi(n^{-1/2}t) - \mathrm{e}^{-t^2/(2n)} \right| , \end{aligned}$$

car $|\psi(n^{-1/2}t)| \leq 1$ et $\mathrm{e}^{-t^2/n} \leq 1$. Nous avons, en utilisant l'inégalité triangulaire

$$n|\psi(t/\sqrt{n}) - \mathrm{e}^{-t^2/2n}| \leq n|\psi(t/\sqrt{n}) - (1 - t^2/2n)| + n|(1 - t^2/2n) - \mathrm{e}^{-t^2/2n}|.$$

Nous allons utiliser deux inégalités élémentaires

$$|\mathrm{e}^{\mathrm{i}u} - 1 - \mathrm{i}u| \leq |u|^2/2 \tag{IV-5.15}$$

$$|\mathrm{e}^{\mathrm{i}u} - 1 - \mathrm{i}u - (\mathrm{i}u)^2/2| \leq |u|^3/6 \, . \tag{IV-5.16}$$

En posant $u = \mathrm{i}t^2/2n \geq 0$, nous obtenons en utilisant (IV-5.15)

$$n|(1 - t^2/2n) - \mathrm{e}^{-t^2/2n}| \leq n(t^2/2n)^2/2 = t^4/8n \to_{n \to \infty} 0 \, .$$

Comme $\mathbb{E}[Y] = 0$ et $\mathrm{Var}(Y) = \mathbb{E}[Y^2] = 1$, nous avons

$$\begin{aligned} n|\psi(t/\sqrt{n}) - (1 - t^2/2n)| &= n|\mathbb{E}[\mathrm{e}^{\mathrm{i}tY/\sqrt{n}} - (1 + \mathrm{i}tY/\sqrt{n} + \mathrm{i}^2 t^2 Y^2/2n)]| \\ &\leq \ n\mathbb{E}[|\mathrm{e}^{\mathrm{i}tY/\sqrt{n}} - (1 + \mathrm{i}tY/\sqrt{n} + \mathrm{i}^2 t^2 Y^2/2n)|] \, . \end{aligned}$$

En utilisant (IV-5.15) avec $u = tY/\sqrt{n}$, nous avons

$$\begin{aligned} |\mathrm{e}^{\mathrm{i}tY/\sqrt{n}} - (1 + \mathrm{i}tY/\sqrt{n} + \mathrm{i}^2 t^2 Y^2/2n)| &\leq |\mathrm{e}^{\mathrm{i}tY/\sqrt{n}} - (1 + \mathrm{i}tY/\sqrt{n})| + t^2 Y^2/2n \\ &\leq \ t^2 Y^2/2n + t^2 Y^2/2n \\ &= \ t^2 Y^2/n \, . \end{aligned}$$

D'autre part, en utilisant cette fois l'inégalité (IV-5.16), nous obtenons

$$|\mathrm{e}^{\mathrm{i}tY/\sqrt{n}} - (1 + \mathrm{i}tY/\sqrt{n} + \mathrm{i}^2 t^2 Y^2/2n)| \leq |t|^3 |Y|^3/6n^{3/2} \, .$$

Par suite, pour tout ensemble A, on a

$$|\mathrm{e}^{\mathrm{i}tY/\sqrt{n}} - (1 + \mathrm{i}tY/\sqrt{n} + \mathrm{i}^2 t^2 Y^2/2n)| \leq (t^2Y^2/n)\mathbb{1}_A + (|t|^3|Y|^3/6n^{3/2})\mathbb{1}_{A^c} .$$

Pour tout $\delta > 0$ et $n \in \mathbb{N}$, posons $A = A(\delta,\, n) := \{|Y| > \delta n^{1/6}\}$. Il vient pour tout $n \geq 1$,

$$\begin{aligned} n\mathbb{E}[|\mathrm{e}^{\mathrm{i}tY/\sqrt{n}} - (1 + \mathrm{i}tY/\sqrt{n} + \mathrm{i}^2 t^2 Y^2/2n)|] & \\ \leq n\mathbb{E}[(t^2Y^2/n)\mathbb{1}_A] + n\mathbb{E}[(|t|^3|Y|^3/6n^{3/2})\mathbb{1}_{A^c}] & \\ \leq t^2\mathbb{E}[Y^2\mathbb{1}\left\{|Y| > \delta n^{1/6}\right\}] + |t|^3\delta^3/6 . & \end{aligned}$$

Par conséquent, pour tout $\varepsilon > 0$, nous pouvons choisir δ tel que $|t|^3\delta^3/6 \leq \varepsilon/2$ puis (en appliquant le théorème de convergence monotone en utilisant $\mathbb{E}[Y^2] = 1$) N tel que pour tout $n \geq N$ nous ayons

$$t^2\mathbb{E}[Y^2\mathbb{1}_{\{|Y|>\delta n^{1/6}\}}] \leq \varepsilon/2 .$$ □

Exemple IV-5.40 (Statistique de Student). Soit $\{Y_n,\, n \in \mathbb{N}\}$ une suite de v.a. i.i.d. centrées et de variance σ^2. Considérons la *t-statistique*

$$T_n := \sqrt{n}\frac{\bar{Y}_n}{S_n^2} ,$$

où

$$\bar{Y}_n := n^{-1}\sum_{k=1}^{n} Y_k, \qquad S_n^2 := n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

désignent resp. la moyenne et la variance empirique. Nous allons montrer que T_n est asymptotiquement normale i.e. $\{T_n,\, n \in \mathbb{N}\}$ converge en loi vers une loi gaussienne.

Remarquons tout d'abord que, par application de la loi faible des grands nombres (Théorème IV-5.18) et du théorème de continuité (Théorème IV-5.6), nous avons :

$$S_n^2 = n^{-1}\sum_{i=}^{n} Y_i^2 - (\bar{Y}_n)^2 \quad \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \quad (\mathbb{E}Y_1^2 - (\mathbb{E}Y_1)^2) = \mathrm{Var}(Y_1).$$

Le théorème de continuité implique aussi que $S_n = \sqrt{S_n^2}$ converge en probabilité vers $\sqrt{\mathrm{Var}(Y_1)} = \sigma$. Le théorème de la limite centrale (théorème IV-5.39) montre que $\sqrt{n}\bar{Y}_n \Longrightarrow \mathrm{N}(0, \sigma^2)$ et le lemme de Slutsky (Lemme IV-5.33) implique que $T_n \Longrightarrow \mathrm{N}(0,1)$. ♢

IV-5.5.2 T.L.C. pour des v.a. indépendantes

Il existe une autre méthode de preuve du théorème de la limite centrale due à Lindeberg (1922), qui permet de généraliser le T.L.C. à des variables aléatoires indépendantes mais qui ne sont pas nécessairement identiquement distribuées. Ce résultat s'applique donc en toute généralité à des tableaux triangulaires de v.a. indépendantes.

Théorème IV-5.41 (Lindeberg–Feller). *Soit* $\{k_n,\ n \in \mathbb{N}\}$ *une suite d'entiers croissante. Soit* $\{(Y_{n,i})_{i=1}^{k_n}, n \in \mathbb{N}\}$ *un tableau triangulaire de vecteurs aléatoires définis sur le même espace de probabilité* $(\Omega, \mathscr{F}, \mathbb{P})$*, indépendants, centrés et tels que* $\mathbb{E}[\|Y_{n,i}\|^2] < \infty$ *pour* $i \in \{1, \ldots, k_n\}$*. Supposons les* conditions de Lindeberg–Feller *vérifiées :*

$$\lim_{n\to\infty}\sum_{i=1}^{k_n}\mathbb{E}[\|Y_{n,i}\|^2\,\mathbb{1}_{\{\|Y_{n,i}\|>\varepsilon\}}] = 0, \quad \textit{pour tout } \varepsilon > 0, \tag{IV-5.17}$$

$$\lim_{n\to\infty}\sum_{i=1}^{k_n}\mathrm{Var}(Y_{n,i}) = \Sigma. \tag{IV-5.18}$$

Alors, la suite $\{\sum_{i=1}^{k_n} Y_{n,i}, n \in \mathbb{N}\}$ *converge en loi vers une gaussienne centrée et de matrice de covariance* Σ.

En utilisant l'inégalité de Markov (Lemme IV-1.1), il est aisé de démontrer qu'une condition suffisante pour (IV-5.17) est

$$\exists \delta > 0, \qquad \lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}[\|Y_{n,i}\|^{2+\delta}] = 0\,. \tag{IV-5.19}$$

Démonstration. Par le procédé de Cramér-Wold (Théorème IV-5.37), il suffit de montrer ce résultat en dimension un.

La méthode de Lindeberg repose sur la comparaison entre les sommes partielles $\sum_{i=1}^{k_n} Y_{n,i}$ et $\sum_{i=1}^{k_n} X_{n,i}$ où $\{(X_{n,i})_{i=1}^{k_n}, n \in \mathbb{N}\}$ est un tableau triangulaire de v.a. *gaussiennes* indépendantes, centrées, telles que,

— pour tout n et pour tout $i \in \{1,\dots,k_n\}$, $\mathrm{Var}(Y_{n,i}) = \mathrm{Var}(X_{n,i})$,

— pour tout n et tout $i, j \in \{1,\dots,k_n\}$, les v.a. $X_{n,i}$ et $Y_{n,j}$ sont indépendantes.

Nous allons tout d'abord montrer que, sous ces deux conditions, il est possible de contrôler la différence entre les sommes partielles construites à partir des tableaux triangulaires $(X_{n,i}, i \in \{1,\dots,k_n\})$ et $(Y_{n,i}, i \in \{1,\dots,k_n\})$ de telle sorte que la convergence en loi de la somme partielle $R_n := \sum_{i=1}^{k_n} X_{n,i}$ implique la convergence en loi de la somme partielle $T_n := \sum_{i=1}^{k_n} Y_{n,i}$.

Soit f une fonction deux fois différentiable avec une dérivée seconde bornée et Lipschitzienne, *i.e.*,

$$|f''|_{\mathrm{Lip}} := \sup_{(x,y)\in\mathbb{R}\times\mathbb{R}, x\neq y} \frac{|f''(x) - f''(y)|}{|x-y|} < \infty, \tag{IV-5.20}$$

où f'' est la dérivée seconde de f. On a la décomposition

$$\mathbb{E}\left[f(R_n)\right] - \mathbb{E}\left[f(T_n)\right] = \sum_{k=1}^{k_n} \left(\mathbb{E}\left[f(R_{n,k}+X_{n,k})\right] - \mathbb{E}\left[f(R_{n,k}+Y_{n,k})\right]\right),$$

où $R_{n,k} := \left(\sum_{j<k} X_{n,j}\right) + \left(\sum_{j>k} Y_{n,j}\right)$. Développons $f(R_{n,k}+X_{n,k})$ au voisinage de $R_{n,k}$:

$$f(R_{n,k}+X_{n,k}) = f(R_{n,k}) + X_{n,k} f'(R_{n,k}) + \frac{X_{n,k}^2}{2} f''(R_{n,k}) + \frac{X_{n,k}^2}{2}\left[f''(R_{n,k}+\theta_{n,k}X_{n,k}) - f''(R_{n,k})\right],$$

où $\theta_{n,k} \in [0,1]$. Développons de même $f(R_{n,k}+Y_{n,k})$. Notons que la v.a. $R_{n,k}$ est indépendante, par construction, des v.a. $X_{n,k}$ et $Y_{n,k}$. Par conséquent, nous avons, pour tout $k \in \{1,\dots,n\}$,

$$\mathbb{E}[f'(R_{n,k})(X_{n,k}-Y_{n,k})] = \mathbb{E}[f'(R_{n,k})]\left(\mathbb{E}[X_{n,k}] - \mathbb{E}[Y_{n,k}]\right) = 0,$$

en utilisant que $\mathbb{E}[X_{n,k}] = \mathbb{E}[Y_{n,k}] = 0$. De la même façon, comme par construction des v.a $X_{n,k}$ et $Y_{n,k}$ nous avons $\mathbb{E}[X_{n,k}^2] = \mathbb{E}[Y_{n,k}^2]$, on en déduit

$$\mathbb{E}[f''(R_{n,k})(X_{n,k}^2 - Y_{n,k}^2)] = \mathbb{E}[f''(R_{n,k})]\left(\mathbb{E}[X_{n,k}^2] - \mathbb{E}[Y_{n,k}^2]\right) = 0\,.$$

Comme f'' est une fonction Lipshitzienne, pour tout $\varepsilon > 0$, nous avons, pour tout $\theta \in [0,1]$,

$$\begin{aligned}
&|f''(R_{n,k}+\theta X_{n,k}) - f''(R_{n,k})| \le |f''|_{\mathrm{Lip}} |X_{n,k}|,\\
&|f''(R_{n,k}+\theta Y_{n,k}) - f''(R_{n,k})| \le |f''|_{\mathrm{Lip}} |Y_{n,k}| \mathbb{1}_{\{|Y_{n,k}|\le\varepsilon\}} + 2|f''|_\infty \mathbb{1}_{\{|Y_{n,k}|>\varepsilon\}}\,,
\end{aligned}$$

où $|f''|_\infty := \sup_{x\in\mathbb{R}} |f''(x)|$ (qui est fini par hypothèse). Nous utilisons ici deux majorations différentes pour des raisons qui deviendront transparentes dans la suite de la preuve. Notons tout d'abord que, pour tout $\varepsilon > 0$,

$$\begin{aligned}
&\left|\mathbb{E}[f(R_{n,k}+X_{n,k})] - \mathbb{E}[f(R_{n,k}+Y_{n,k})]\right| \\
&\qquad\qquad \le \frac{1}{2}|f''|_{\mathrm{Lip}} \left(\mathbb{E}\left[|X_{n,k}|^3\right] + \mathbb{E}\left[|Y_{n,k}\,\mathbb{1}_{\{|Y_{n,k}|^3\le\varepsilon\}}\right]\right) + |f''|_\infty \mathbb{E}\left[|Y_{n,k}\,\mathbb{1}_{\{Y_{n,k}^2>\varepsilon\}}\right]\,.
\end{aligned}$$

Remarquons que $\mathbb{E}\left[|X_{n,k}|^3\right] = \sigma_{n,k}^3 m_3$ où $m_3 := \mathbb{E}[|Z|^3]$ avec $Z \sim \mathrm{N}(0,1)$ et $\sigma_{n,k}^2 := \mathbb{E}[X_{n,k}^2] = \mathbb{E}[Y_{n,k}^2]$. Remarquons aussi que

$$\mathbb{E}\left[|Y_{n,k}|^3 \mathbb{1}_{\{|Y_{n,k}|\leq\varepsilon\}}\right] \leq \varepsilon\mathbb{E}\left[|Y_{n,k}|^2\right] = \varepsilon\sigma_{n,k}^2.$$

Ces inégalités conduisent à la majoration :

$$|\mathbb{E}f(R_n) - \mathbb{E}f(T_n)| \leq \frac{1}{2}|f''|_{\mathrm{Lip}}\left(m_3\sum_{k=1}^{k_n}\sigma_{n,k}^3 + \varepsilon\sum_{k=1}^{k_n}\sigma_{n,k}^2\right) + |f''|_\infty\sum_{k=1}^{k_n}\mathbb{E}\left[Y_{n,k}^2\mathbb{1}_{\{|Y_{n,k}|>\varepsilon\}}\right].$$

On a d'autre part

$$\sum_{k=1}^{k_n}\sigma_{n,k}^3 \leq \sum_{k=1}^{k_n}\sigma_{n,k}^2 \max_{k\in\{1,\ldots,k_n\}}(\sigma_{n,k}).$$

Or (IV-5.18) implique que pour tout $1 \leq k \leq k_n$, et tout $\varepsilon > 0$, nous avons

$$\sigma_{n,k}^2 \leq \varepsilon^2 + \mathbb{E}\left[Y_{n,k}^2\mathbb{1}_{\{|Y_{n,k}|>\varepsilon\}}\right] \leq \varepsilon^2 + \sum_{j=1}^{k_n}\mathbb{E}\left[Y_{n,j}^2\mathbb{1}_{\{|Y_{n,j}|>\varepsilon\}}\right].$$

et donc

$$\limsup_{n\to\infty}\max_{k\in\{1,\ldots,k_n\}}\sigma_{n,k}^2 = 0.$$

D'où, pour tout $\varepsilon > 0$,

$$|\mathbb{E}f(R_n) - \mathbb{E}f(T_n)| \leq \frac{1}{2}|f''|_{\mathrm{Lip}}\left(m_3\sum_{k=1}^{k_n}\sigma_{n,k}^2\left(\varepsilon^2 + \sum_{j=1}^{k_n}\mathbb{E}\left[Y_{n,j}^2\mathbb{1}_{\{|Y_{n,j}|>\varepsilon\}}\right]\right)^{1/2} + \varepsilon\sum_{k=1}^{k_n}\sigma_{n,k}^2\right)$$
$$+ |f''|_\infty\sum_{k=1}^{k_n}\mathbb{E}\left[Y_{n,k}^2\mathbb{1}_{\{|Y_{n,k}|>\varepsilon\}}\right].$$

Par hypothèse, quand n tend vers l'infini, la série $\sum_k \mathbb{E}\left[Y_{n,k}^2\mathbb{1}_{\{|Y_{n,k}|>\varepsilon\}}\right]$ tend vers zéro et $\sum_k \sigma_{n,k}^2$ tend vers une constante σ^2. D'où, pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty}|\mathbb{E}f(R_n) - \mathbb{E}f(T_n)| \leq \frac{1}{2}|f''|_{\mathrm{Lip}}(m_3+1)\sigma^2\varepsilon.$$

En faisant tendre, dans un deuxième temps, ε vers zéro, on obtient donc une limite nulle. Comme les $(X_{n,j}, j \in \{1,\ldots,k_n\})$ sont des v.a. gaussiennes, la proposition **??** permet de conclure.

Ces majorations étant en particulier valables pour $f(x) = \exp(-itx)$, ce qui permet de conclure que $\{T_n,\, n\in\mathbb{N}\}$ et $\{R_n,\, n\in\mathbb{N}\}$ ont même limite en loi. Enfin, R_n étant la somme de gaussiennes indépendantes, c'est une gaussienne centrée de variance $\sum_{i=1}^{k_n}\mathrm{Var}(X_{n,i})$. Puisque la fonction caractéristique de R_n vaut

$$t \mapsto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{k_n}\mathrm{Var}(X_{n,i})\right)t^2\right) = \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{k_n}\mathrm{Var}(Y_{n,i})\right)t^2\right)$$

elle converge vers $\exp(-\sigma^2t^2/2)$ et on reconnaît la fonction caractéristique d'une loi gaussienne centrée de matrice de covariance Σ. On en déduit la convergence en loi de $\{R_n,\, n\in\mathbb{N}\}$ vers cette gaussienne, par application du Théorème de Levy (théorème IV-5.36). □

Exemple IV-5.42. Soit $\{Y_n,\, n\in\mathbb{N}\}$ une suite de variables indépendantes. Supposons que, pour tout $n\in\mathbb{N}$, Y_n est distribuée suivant une loi de Bernoulli de paramètre $p_n \in [0,1]$ et que $\sum_{i=1}^{\infty} p_i(1-p_i) = \infty$. Posons

$$s_n^2 := \sum_{i=1}^{n}\mathrm{Var}(Y_i) = \sum_{i=1}^{n} p_i(1-p_i).$$

Par construction, $\lim_{n\to\infty} s_n = \infty$. Posons $Y_{n,i} = s_n^{-1}(Y_i - p_i)$. Clairement, $\sum_{i=1}^n \mathrm{Var}(Y_{n,i}) = 1$. D'autre part, nous avons

$$\mathbb{E}[|Y_{n,i}|^3] = s_n^{-3}\left\{p_i^3(1-p_i) + (1-p_i)^3 p_i\right\} \le 2s_n^{-3} p_i(1-p_i)\,,$$

ce qui implique

$$\sum_{i=1}^n \mathbb{E}[|Y_{n,i}|^3] \le \frac{2}{s_n^3}\sum_{i=1}^n p_i(1-p_i) = \frac{2}{s_n} \to 0,$$

car $\lim_{n\to\infty} s_n = \infty$. Par conséquent la condition (IV-5.17) est satisfaite avec $\varepsilon = 1$. Le Théorème IV-5.41 montre que

$$s_n^{-1}\sum_{i=1}^n (Y_i - p_i) \Longrightarrow \mathrm{N}(0,1)\,.$$

Exemple IV-5.43 (Théorème de Hájek-Sidak). Soit $\{X_n,\ n\in\mathbb{N}^*\}$ une suite de variables aléatoires réelles i.i.d. d'espérance μ et de variance $\sigma^2 \in]0,\infty[$. Soit $\{(c_{n,i})_{i=1}^n, n\in\mathbb{N}\}$ un tableau triangulaire de constantes tel que, pour tout $n \ge 0$, $\sum_{i=1}^n c_{n,i}^2 > 0$ et

$$\delta_n := \max_{1\le i\le n} \frac{c_{n,i}^2}{\sum_{j=1}^n c_{n,j}^2} \to 0\,. \tag{IV-5.21}$$

Alors, nous avons

$$\frac{\sum_{i=1}^n c_{n,i}(X_i-\mu)}{\sigma\sqrt{\sum_{i=1}^n c_{n,i}^2}} \Longrightarrow \mathrm{N}(0,\ 1)\,. \tag{IV-5.22}$$

Pour établir ce résultat, posons $s_n^2 := \sigma^2\sum_{i=1}^n c_{n,i}^2$ et $Y_{n,i} := s_n^{-1}c_{n,i}(X_i-\mu)$ pour $i\in\{1,\dots,n\}$. Clairement, $\sum_{i=1}^n \mathrm{Var}(Y_{n,i}) = 1$. Nous avons de plus

$$\begin{aligned}\mathbb{E}\left[Y_{n,i}^2 \mathbb{1}_{\{|Y_{n,i}|>\varepsilon\}}\right] &= s_n^{-2}c_{n,i}^2\mathbb{E}\left[(X_1-\mu)^2\mathbb{1}_{\{c_{n,i}|X_1-\mu|>\varepsilon s_n\}}\right]\\ &\le c_{n,i}^2\mathbb{E}\left[(X_1-\mu)^2\mathbb{1}_{\{\sqrt{\delta_n}|X_1-\mu|>\varepsilon\sigma\}}\right]\,.\end{aligned}$$

Par conséquent, par le théorème de convergence dominée (Théorème A.40), nous avons

$$\frac{1}{s_n^2}\sum_{i=1}^n c_{n,i}^2\mathbb{E}\left[(X_1-\mu)^2\mathbb{1}_{\{\sqrt{\delta_n}|X_1-\mu|>\varepsilon\sigma\}}\right] = \frac{\mathbb{E}[(X_1-\mu)^2\mathbb{1}_{\{\sqrt{\delta_n}|X_1-\mu|>\varepsilon\sigma\}}]}{\sigma^2} \to 0\,.$$

De façon intuitive, la condition (IV-5.21) montre que, lorsque $n\to\infty$, la contribution de chaque terme dans la somme

$$\sum_{i=1}^n \frac{c_{n,i}(X_i-\mu)}{\sigma\sqrt{\sum_{j=1}^n c_{n,j}^2}}, \tag{IV-5.23}$$

est négligeable. Cette condition est indispensable pour obtenir une loi limite. Si par exemple $c_{n,1} = 1$ et $c_{n,j} = 0$ pour tout $j\in\{2,\dots,n\}$, la somme (IV-5.23) est toujours égale à $(X_1-\mu)/\sigma$ et il n'y a bien évidemment pas de passage à la limite. ♢

IV-5.5.3 Vitesse dans le T.L.C.

Soit $\{X_n,\ n\in\mathbb{N}\}$ une suite de variables aléatoires réelles i.i.d. d'espérance μ et de variance $\sigma^2 > 0$; on pose

$$\overline{X}_n := n^{-1}\sum_{k=1}^n X_k\,, \qquad Y_n := \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}\,.$$

Le théorème de la limite centrale montre que $Y_n \Longrightarrow \mathrm{N}(0,1)$ ce qui implique en particulier que pour tout $x\in\mathbb{R}$, $\lim_{n\to\infty} F_n(x) = \Phi(x)$ (voir Théorème IV-5.26) où F_n (resp. Φ) désigne la fonction de répartition de Y_n (resp. d'une loi $\mathrm{N}(0,1)$). Comme la fonction $x\mapsto\Phi(x)$ est continue, le théorème de Polya (Théorème IV-5.44) montre que la convergence est uniforme en x.

Le théorème de Polya, qui est une conséquence du théorème de Dini, montre que si la fonction de répartition de la limite est continue, alors la convergence est uniforme.

Théorème IV-5.44 (Théorème de Polya). *Soit $\{X_n,\ n\in\mathbb{N}\}$ une suite de variables aléatoires réelles, de fonction de répartition notée F_n. Soit X une variable aléatoire réelle de fonction de répartition F. Si F est continue et si $X_n \Longrightarrow X$, alors*

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|=0.$$

Démonstration. La fonction F étant continue, il existe des points $-\infty = x_0 < x_1 < \ldots < x_k = \infty$ tels que $F(x_i) = i/k$. F_n et F étant croissantes, nous avons, pour tout $x\in\mathbb{R}^d$, en choisissant i tel que $x_{i-1}\le x\le x_i$:

$$\begin{aligned}F_n(x)-F(x)&\le F_n(x_i)-F(x_{i-1})=F_n(x_i)-F(x_i)+1/k\\ F_n(x)-F(x)&\ge F_n(x_{i-1})-F(x_i)=F_n(x_{i-1})-F(x_{i-1})-1/k.\end{aligned}$$

Donc $|F_n(x)-F(x)|$ est borné par $\sup_i|F_n(x_i)-F(x_i)|+1/k$ pour tout x. Par conséquent,

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\le\lim_{n\to\infty}\sup_{i\in\{0,\ldots,k\}}|F_n(x_i)-F(x_i)|+1/k=1/k,$$

ce qui permet de conclure, en choisissant k arbitrairement grand. □

Le raisonnement s'étend sans difficulté au cas multidimensionnel mais nous omettons cet énoncé, la fonction de répartition étant mieux adaptée à la dimension un.

Ce résultat ne fournit pas par contre de contrôle de l'erreur que nous commettons lorsque nous approchons F_n par Φ. Le théorème de Berry-Esseen donne un contrôle de l'erreur d'approximation dans le théorème de la limite centrale.

Théorème IV-5.45 (Berry-Esseen). *Soient $\{X_n,\ n\in\mathbb{N}\}$ une suite de variables aléatoires réelles définies sur $(\Omega,\mathscr{F},\mathbb{P})$, i.i.d. et telles que $\mathbb{E}[|X_1|^3]<\infty$. Notons μ l'espérance de X_1 et σ^2 sa variance. Il existe une constante universelle C (ne dépendant pas de la loi de X_1) telle que, pour tout x et n,*

$$\left|\mathbb{P}\left(\sqrt{n}\left(\frac{\overline{X}_n-\mu}{\sigma}\right)\le x\right)-\Phi(x)\right|\le\frac{C}{\sqrt{n}}\frac{\mathbb{E}[|X_1-\mu|^3]}{\sigma^3}\,, \tag{IV-5.24}$$

où Φ est la fonction de répartition de la loi $\mathrm{N}(0,1)$.

Démonstration. Voir [**?**, Chapitre XVI, Section V, Théorème 1] dans le cas où $C=3$. □

Le théorème de Berry-Esseen est vérifié avec la constante $C=0.4784$. On ne connaît pas aujourd'hui la plus petite valeur de la constante C, mais on sait que le résultat n'est pas vérifié pour $C<0.4097$ [**?**].

Si $\mathbb{E}[|X_1-\mu|^3]/\sigma^3<\infty$, le terme de droite de (IV-5.24) tend vers 0 et donc le terme de gauche tend uniformément vers 0 en x. Si la loi de la variable aléatoire X_1 appartient à une famille de distributions telle que

$$\frac{\mathbb{E}[|X_1-\mu|^3]}{\sigma^3}\le B\,,$$

pour $B<\infty$, alors la convergence est aussi uniforme par rapport à la loi de X_1 !

Exemple IV-5.46 (Erreur d'approximation pour une loi de Bernoulli). Soit $(X_1,\ldots,X_n)$ un n-échantillon de loi de Bernoulli de paramètre $\theta\in\Theta=[0,1]$. L'espérance de X_1 est θ, sa variance est $\theta(1-\theta)$ et le moment centré d'ordre 3 vaut

$$\mathbb{E}[|X_1-\theta|^3]=\sigma_\theta(1-2\sigma_\theta)\quad\text{où }\sigma_\theta^2:=\theta(1-\theta)\,.$$

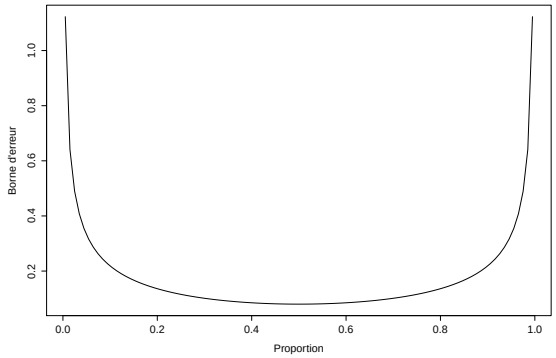

FIGURE IV-5.1 – Borne de Berry-Esseen en fonction de la proportion $\theta \in]0,1[$ pour $n = 100$

En prenant $C = 0.8$ dans la borne de Berry-Esseen, nous avons, pour tout $\theta \in \Theta$,

$$\left|\mathbb{P}_\theta\left(\sqrt{n}(\bar{X}_n - \theta)/\sigma_\theta \leq x\right) - \Phi(x)\right| \leq \frac{0.8}{\sqrt{n}} \frac{\sigma_\theta(1-2\sigma_\theta)}{\sigma_\theta^{3/2}} .$$

Nous avons représenté cette borne dans la Figure IV-5.1 pour $n = 100$. Nous voyons que l'erreur d'approximation est minimale quand $\theta = 1/2$ mais que l'erreur est grande quand $\theta \to 0$ ou $\theta \to 1$, ce qui est assez logique. D'autres méthodes d'approximation doivent être appliquées dans ce cas (approximation Poissonnienne par exemple). Même lorsque $\theta = 0.5$, l'erreur est égale à 0.08 pour $n = 1e2$, 0.025 pour $n = 1e3$ et 0.008 pour $n = 1e4$. Il faut donc être vigilant au fait que, même dans des cas très favorables, l'erreur dans la borne de Berry-Esseen est assez grande. C'est le prix à payer pour disposer d'un contrôle uniforme en x, de meilleures bornes peuvent être obtenues si on relâche cette contrainte. ♢

IV-5.5.4 La δ–méthode

Nous savons déjà que, si la suite d'estimateurs $\{T_n,\ n \in \mathbb{N}\}$ converge en probabilité vers θ et que g est continue au point θ, alors $\{g(T_n), n \in \mathbb{N}\}$ converge en probabilité vers $g(\theta)$ (voir Théorème IV-5.6). Si nous savons de plus que $\sqrt{n}(T_n - \theta)$ converge en loi vers une distribution limite, pouvons nous affirmer qu'il en est de même pour $\sqrt{n}\{g(T_n) - g(\theta)\}$? La réponse est affirmative si la fonction g est différentiable au point θ : de façon heuristique, nous avons :

$$\sqrt{n}\{g(T_n) - g(\theta)\} \simeq g'(\theta)\sqrt{n}(T_n - \theta) ,$$

et donc, si $\sqrt{n}\{T_n - \theta\} \Longrightarrow T$, alors $\sqrt{n}\{g(T_n) - g(\theta)\} \Longrightarrow g'(\theta)T$. En particulier, si $\sqrt{n}(T_n - \theta) \Longrightarrow \mathrm{N}(0, \sigma^2(\theta))$, alors $\sqrt{n}\{g(T_n) - g(\theta)\} \Longrightarrow \mathrm{N}(0, [g'(\theta)]^2\sigma^2(\theta))$. La même question se pose lorsque $\mathbf{T}_n = (T_{n,1}, \ldots, T_{n,p})$ est un vecteur aléatoire et $g = (g_1, \cdots, g_m)$ est une fonction de $\mathbb{R}^p \mapsto \mathbb{R}^m$. Le résultat ci-dessus s'étend directement en remplaçant la dérivée par la matrice Jacobienne de g au point θ, notée $\mathrm{J}_g(\theta)$ et définie par

$$\mathrm{J}_g(\theta) := \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_1}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial \theta_1} & \cdots & \frac{\partial g_m}{\partial \theta_p} \end{pmatrix} . \tag{IV-5.25}$$

Théorème IV-5.47 (δ-méthode). *Soit D_g un sous-ensemble ouvert de $\mathbb{R}^p$ et $\theta \in D_g$. Soit $g : D_g \mapsto \mathbb{R}^m$ une fonction différentiable au point θ. Soit $\{\mathbf{T}_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires tels que, pour tout $n \in \mathbb{N}$,*

$$\mathbb{P}(\mathbf{T}_n \in D_g) = 1 .$$

Supposons que

$$r_n(\mathbf{T}_n - \theta) \Longrightarrow \mathbf{G}\,,$$

pour une suite $\{r_n,\, n \in \mathbb{N}\}$ *croissante telle* $\lim_{n\to\infty} r_n = \infty$. *Alors*

$$r_n\{g(\mathbf{T}_n) - g(\theta)\} - r_n \mathrm{J}_g(\theta)(\mathbf{T}_n - \theta) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0\,.$$

et donc

$$r_n\{g(\mathbf{T}_n) - g(\theta)\} \Longrightarrow \mathrm{J}_g(\theta)\mathbf{G}\,.$$

Démonstration. Comme $\{r_n(\mathbf{T}_n - \theta), n \in \mathbb{N}\}$ converge en loi, par le lemme IV-5.33 nous avons

$$\mathbf{T}_n - \theta = r_n^{-1}\, r_n\{\mathbf{T}_n - \theta\} \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0\,.$$

La différentiabilité de la fonction g au point $\theta \in D_g$ implique que

$$g(\theta + h) = g(\theta) + \mathrm{J}_g(\theta)h + R(h)$$

où $\lim_{\|h\|\to 0} \|R(h)\| / \|h\| = 0$ et $R(0) = 0$. Il existe donc une fonction $\tilde{R}(h)$, vérifiant $R(h) = \|h\|\tilde{R}(h)$, continue en zero et telle que $\tilde{R}(0) = 0$. On a donc

$$r_n\left(g(\mathbf{T}_n) - g(\theta)\right) - \mathrm{J}_g(\theta)\left\{r_n(\mathbf{T}_n - \theta)\right\} = r_n \|\mathbf{T}_n - \theta\|\, \tilde{R}(\mathbf{T}_n - \theta)\,.$$

Puisque $\mathbf{T}_n - \theta \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$, $\tilde{R}(\mathbf{T}_n - \theta) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$; de plus, par le Théorème IV-5.30, $r_n \|\mathbf{T}_n - \theta\| \Longrightarrow \|G\|$. Nous concluons en appliquant le lemme de Slutsky (Lemme IV-5.33). □

Exemple IV-5.48. Soit $\{X_n,\, n \in \mathbb{N}\}$ une suite de variables aléatoires telles que

$$r_n(X_n - \mu) \Longrightarrow \mathrm{N}(0, \sigma^2)$$

où $\{r_n,\, n \in \mathbb{N}\}$ est une suite à termes positifs et croissante telle que $\lim_{n\to\infty} r_n = \infty$. Que pouvons-nous dire de la distribution limite des variables aléatoires $X_n^2, \exp(X_n), 1/X_n, \log|X_n|$? Nous avons $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \mu$ et le théorème de continuité (théorème IV-5.6) implique que :

$$X_n^2 \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \mu^2, \qquad \exp(X_n) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \exp(\mu),$$

et si $\mu \neq 0$

$$\frac{1}{X_n} \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \frac{1}{\mu}, \qquad \log|X_n| \overset{\mathbb{P}-\text{prob}}{\longrightarrow} \log|\mu|.$$

De plus,

— posons $g(x) = x^2$. Cette fonction est continûment différentiable et $g'(x) = 2x$. Si $\mu \neq 0$, alors $g'(\mu) = 2\mu \neq 0$. Une application du théorème IV-5.47 montre que, pour $\mu \neq 0$,

$$r_n(X_n^2 - \mu^2) \Longrightarrow \mathrm{N}(0, 4\mu^2\sigma^2)\,.$$

Pour $\mu = 0$, $g'(\mu) = 0$ et l'application de théorème IV-5.47 montre que

$$r_n X_n^2 \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0\,.$$

On peut néanmoins obtenir un résultat plus précis, en exhibant une vitesse de convergence, en appliquant le théorème de continuité (théorème IV-5.30). En effet,

$$r_n\left(r_n X_n^2\right) = (r_n X_n)^2 \Longrightarrow \sigma^2 X^2$$

où X est une variable aléatoire gaussienne $\mathrm{N}(0,1)$. Par conséquent, X^2 suit une loi du χ^2 centrée à un degré de liberté.

— Pour tout μ, le théorème IV-5.47 appliqué à la fonction $g(x) = \mathrm{e}^x$ montre que $r_n(\exp(X_n) - \exp(\mu)) \Longrightarrow \mathrm{N}(0, \exp(2\mu)\sigma^2)$.

— Nous supposons ici que pour tout $n \in \mathbb{N}$, $\mathbb{P}(X_n \neq 0) = 1$ (autrement, la variable aléatoire $1/X_n$ n'est pas définie). Pour $\mu \neq 0$, une application de théorème IV-5.47 à la fonction $g(x) = 1/x$ qui est différentiable au point $\mu \neq 0$ montre que

$$r_n(1/X_n - 1/\mu) \Longrightarrow \mathrm{N}(0, \sigma^2/\mu^4) \; .$$

Le cas $\mu = 0$ n'est pas couvert par le théorème IV-5.47, mais on peut dans ce contexte encore utiliser le théorème de continuité (Théorème IV-5.30) qui montre que

$$1/(r_n X_n) \Longrightarrow 1/(\sigma X) \; ,$$

où X est une variable aléatoire gaussienne $\mathrm{N}(0,1)$.

— Nous supposons ici encore que pour tout $n \in \mathbb{N}$, $\mathbb{P}(X_n = 0) = 1$. Pour $\mu > 0$, le théorème IV-5.47 appliqué avec $g(x) = \log(|x|)$ ($g'(\mu) = 1/\mu$) montre que

$$r_n\{\log(|X_n|) - \log(|\mu|)\} \Longrightarrow \mathrm{N}(0, \sigma^2/\mu^2) \; . \tag{IV-5.26}$$

Pour $\mu < 0$, nous pouvons appliqué encore le théorème IV-5.47. Dans ce cas, $g(\mu) = -1/\mu$ et donc nous avons encore (IV-5.26).

Le cas $\mu = 0$ n'est pas couvert par le théorème IV-5.47, mais le théorème de continuité montre que $\log|r_n X_n| \Longrightarrow \log|\mathrm{N}(0,\sigma^2)|$. ♢

Exemple IV-5.49 (Variance empirique). Soit $\{X_n,\, n \in \mathbb{N}\}$ une suite de variables aléatoires i.i.d. Supposons que $\mathbb{E}[X_1^4] < \infty$ et notons $m_k := \mathbb{E}[X_1^k]$, pour $k = 1, \ldots, 4$. Considérons l'estimateur de la variance empirique

$$S_n^2 := n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 \; , \qquad \text{où } \bar{X}_n := n^{-1} \sum_{i=1}^n X_i .$$

La version multidimensionnelle du théorème de la limite centrale montre que :

$$\sqrt{n} \left(\begin{bmatrix} n^{-1}\sum_{i=1}^n X_i \\ n^{-1}\sum_{i=1}^n X_i^2 \end{bmatrix} - \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} \right) \quad \Longrightarrow \quad \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Gamma \right), \qquad \text{où} \qquad \Gamma := \begin{bmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{bmatrix} .$$

Posons $\phi(x,y) := y - x^2$, de sorte que

$$S_n^2 = \phi\left(\bar{X}_n, \frac{1}{n} \sum_{i=1}^n X_i^2 \right) .$$

La fonction ϕ est différentiable au point (m_1, m_2) et la matrice jacobienne de ϕ au point (m_1, m_2) est le vecteur $[-2m_1, 1]$. Par conséquent,

$$\sqrt{n}(S_n^2 - (m_2 - m_1^2)) \Longrightarrow -2m_1 T_1 + T_2,$$

où (T_1, T_2) est un vecteur gaussien centré de covariance Γ. $-2m_1 T_1 + T_2$ est une variable aléatoire gaussienne de moyenne nulle et sa variance est donnée par

$$[-2m_1, 1] \begin{bmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{bmatrix} \begin{bmatrix} -2m_1 \\ 1 \end{bmatrix}$$

Si $m_1 = 0$ (les variables aléatoires X_i sont centrées), cette variance est simplement égale à $m_4 - 2m_2^2$. Remarquons que la statistique S_n^2 est invariante par translation : la loi de S_n^2 est inchangée si nous calculons la statistique à partir des variables aléatoires $Y_i := X_i - m_1$. Par suite,

$$\sqrt{n}(S_n^2 - \mu_2) \Longrightarrow \mathrm{N}(0, \mu_4 - \mu_2^2),$$

où μ_k sont les moments des variables aléatoires recentrées, $\mu_k := \mathbb{E}\left[(X_1 - m_1)^k\right]$, $k = 1, \ldots, 4$. ♢

Exemple IV-5.50 (Coefficient de corrélation (voir Exemple IV-5.23)). Soit $\{(X_i, Y_i),\, i \in \mathbb{N}\}$ une suite i.i.d. de vecteurs aléatoires de dimension 2. On suppose que $\mathbb{E}[X_1^4] + \mathbb{E}[Y_1^4] < \infty$. Le coefficient de corrélation du couple (X_1, Y_1) est donné par

$$\rho := \frac{\mathrm{Cov}(X_1, Y_1)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(Y_1)}} \; .$$

Le coefficient de corrélation empirique est donné par

$$\hat{\rho}_n := \frac{n^{-1}\sum_{i=1}^n X_i Y_i - \bar{X}_n \bar{Y}_n}{\sqrt{n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}_n^2}\sqrt{n^{-1}\sum_{i=1}^n Y_i^2 - \bar{Y}_n^2}}.$$

où

$$\bar{X}_n := n^{-1}\sum_{k=1}^n X_k, \qquad \bar{Y}_n := n^{-1}\sum_{k=1}^n Y_k.$$

Posons

$$\mathbf{T}_n := \left(\bar{X}_n, \bar{Y}_n,\, n^{-1}\sum_{i=1}^n X_i^2,\, n^{-1}\sum_{i=1}^n Y_i^2,\, n^{-1}\sum_{i=1}^n X_i Y_i\right),$$

et

$$\nu := (\mathbb{E}[X_1],\, \mathbb{E}[Y_1],\, \mathbb{E}[X_1^2],\, \mathbb{E}[Y_1^2],\, \mathbb{E}[X_1 Y_1]),$$

et notons Σ la matrice de covariance du vecteur aléatoire $(X_1,\, Y_1,\, X_1^2,\, Y_1^2,\, X_1 Y_1)$. Le T.L.C. (Théorème IV-5.39) montre que

$$\sqrt{n}\,(\mathbf{T}_n - \nu) \Longrightarrow \mathrm{N}(0, \Sigma)\,. \qquad \diamondsuit$$

Considérons la transformation

$$g(u_1,\, u_2,\, u_3,\, u_4,\, u_5) := \frac{u_5 - u_1 u_2}{\sqrt{(u_3 - u_1^2)(u_4 - u_2^2)}}\,.$$

En utilisant la δ-méthode avec $p = 5$ et $m = 1$ (Théorème IV-5.47), nous avons

$$\sqrt{n}(\hat{\rho}_n - \rho) \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0,\, \gamma^2)\,,$$

où $\gamma^2 := \mathrm{J}_g(\nu)\Sigma\mathrm{J}_g(\nu)^T$. L'expression de γ^2 est assez compliquée dans le cas général (elle dépend de tous les moments joints jusqu'à l'ordre 4). Elle a une expression simple dans certains cas particuliers : par exemple, si la loi du couple $(X_1,\, Y_1)$ est gaussienne, alors

$$\sqrt{n}(\hat{\rho}_n - \rho) \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0,\, (1-\rho^2)^2)\,.$$

IV-5.6 Convergence des moments

Par définition, $X_n \Longrightarrow X$ implique que pour toute fonction continue *bornée* f, $\mathbb{E}_n[f(X_n)] \to \mathbb{E}[f(X)]$. Le fait que la fonction f soit *bornée* n'est pas superflue, et il est facile de construire des exemples de suite de variables aléatoires vérifiant $X_n \Longrightarrow X$ et pour lesquelles nous n'avons pas $\mathbb{E}_n[f(X_n)] \to \mathbb{E}[f(X)]$ pour f une fonction continue non bornée (il est tout à fait possible que $\mathbb{E}_n[f(X_n)]$ ne soit d'ailleurs tout simplement pas défini, car $\mathbb{E}_n[|f(X_n)|] = \infty$ alors que $\mathbb{E}[|f(X)|] < \infty$!). Nous avons toutefois le résultat suivant :

Proposition IV-5.51. *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathscr{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Si $X_n \Longrightarrow X$, alors*

$$\mathbb{E}[\|X\|] \leq \liminf_{n\to\infty} \mathbb{E}_n[\|X_n\|]\,.$$

Démonstration. En utilisant le théorème IV-5.30, nous avons $\|X_n\| \Longrightarrow \|X\|$, ce qui implique que $\mathbb{P}_n(\|X_n\| > t) \to \mathbb{P}(\|X\| > t)$ sauf en un nombre au plus dénombrable de points (les points de discontinuité de la fonction de répartition). En utilisant le Lemme de Fatou (voir Lemme A.39), nous avons donc

$$\mathbb{E}[\|X\|] = \int_0^\infty \mathbb{P}(\|X\| > t)\mathrm{d}t \leq \liminf_{n\to\infty} \int_0^\infty \mathbb{P}_n(\|X_n\| > t)\mathrm{d}t = \liminf_{n\to\infty} \mathbb{E}_n[\|X_n\|]\,. \qquad \square$$

Corollaire IV-5.52. *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathscr{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathscr{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.*

Soit $f : \mathbb{R}^d \mapsto \mathbb{R}^m$ une fonction continue en tout point d'un ensemble C tel que $\mathbb{P}(X \in C) = 1$. Si $X_n \Longrightarrow X$ alors

$$\mathbb{E}[\|f(X)\|] \leq \lim \inf_{n\to\infty} \mathbb{E}_n[\|f(X_n)\|] .$$

Définition IV-5.53 (Uniforme intégrabilité). *Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ définie sur un espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. La suite $\{X_n,\ n \in \mathbb{N}\}$ est dite* uniformément intégrable *si*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}[\|X_n\|\, \mathbb{1}_{\{\|X_n\|\geq M\}}] = 0 .$$

Remarquons tout d'abord que l'uniforme intégrabilité implique que, pour tout $n \in \mathbb{N}$, $\mathbb{E}[\|X_n\|]$ est majoré indépendamment de n. En effet, il existe M et $C > 0$ tels que, pour tout n

$$\mathbb{E}[\|X_n\|\, \mathbb{1}_{\{\|X_n\|\geq M\}}] \leq C,$$

ce qui implique que pour tout n,

$$\mathbb{E}[\|X_n\|] = \mathbb{E}[\|X_n\|\, \mathbb{1}_{\{\|X_n\|\leq M\}}] + \mathbb{E}[\|X_n\|\, \mathbb{1}_{\{\|X_n\|\geq M\}}] \leq M + C .$$

Nous donnons ci-dessous quelques conditions suffisantes d'uniforme intégrabilité.

Théorème IV-5.54. *(i) Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires définis sur $(\Omega, \mathscr{F}, \mathbb{P})$, à valeurs dans $\mathbb{R}^d$. S'il existe $\delta > 0$ tel que $\sup_{n\geq 1} \mathbb{E}[\|X_n\|^{1+\delta}] < \infty$, alors la suite $\{X_n,\ n \in \mathbb{N}\}$ est uniformément intégrable.*

(ii) Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et X une v.a. intégrable. Si pour tout $n \in \mathbb{N}$, $\|X_n\| \leq X$, alors la suite $\{X_n,\ n \in \mathbb{N}\}$ est uniformément intégrable.

(iii) Soient $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeur dans $\mathbb{R}^d$ et $\{Y_n,\ n \in \mathbb{N}\}$ une suite de v.a. réelles. Si pour tout $n \in \mathbb{N}$, $\|X_n\| \leq Y_n$ et que la suite $\{Y_n,\ n \in \mathbb{N}\}$ est uniformément intégrable, alors la suite $\{X_n,\ n \in \mathbb{N}\}$ est uniformément intégrable.

(iv) Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$. Si $\mathbb{E}[\|X_1\|] < \infty$, alors la suite $\{X_n,\ n \in \mathbb{N}\}$ est uniformément intégrable.

(v) Soient $\{X_n,\ n \in \mathbb{N}\}$ et $\{Y_n,\ n \in \mathbb{N}\}$ deux suites de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$. Si les suites $\{X_n,\ n \in \mathbb{N}\}$ et $\{Y_n,\ n \in \mathbb{N}\}$ sont uniformément intégrables, alors la suite $\{X_n + Y_n, n \in \mathbb{N}\}$ est uniformément intégrable.

(vi) Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et $\{Y_n,\ n \in \mathbb{N}\}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}$. Si la suite $\{X_n,\ n \in \mathbb{N}\}$ est uniformément intégrable et la suite $\{Y_n,\ n \in \mathbb{N}\}$ est bornée (presque-sûrement), alors la suite $\{X_n Y_n, n \in \mathbb{N}\}$ est uniformément intégrable.

Démonstration. Ces propriétés sont toutes élémentaires.

(i) Par l'inégalité de Markov (voir Lemme IV-1.1), pour tout $M > 0$,

$$\limsup_{n\to\infty} \mathbb{E}[\|X_n\|\, \mathbb{1}_{\{\|X_n\|\geq M\}}] \leq M^{-\delta} \limsup_{n\to\infty} \mathbb{E}[\|X_n\|^{1+\delta}] .$$

(ii) La condition $\|X_n\| \le X$ implique $\{\|X_n\| \ge M\} \subset \{X \ge M\}$. Par conséquent, pour tout $M \ge 0$,

$$\|X_n\| \mathbb{1}_{\{\|X_n\| \ge M\}} \le X \mathbb{1}_{\{X \ge M\}} ,$$

ce qui implique que, en appliquant le théorème de convergence dominée (voir Théorème A.40),

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \ge M\}}] \le \lim_{M\to\infty} \mathbb{E}[X \mathbb{1}_{\{X \ge M\}}] = 0 .$$

(iii) Pour tout $n \in \mathbb{N}$, nous avons

$$\|X_n\| \mathbb{1}_{\{\|X_n\| \ge M\}} \le Y_n \mathbb{1}_{\{Y_n \ge M\}} ,$$

et donc

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \ge M\}}] \le \lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}[Y_n \mathbb{1}_{\{Y_n \ge M\}}] = 0 .$$

(iv) Nous avons $\mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \ge M\}}] = \mathbb{E}[\|X_1\| \mathbb{1}_{\{\|X_1\| \ge M\}}]$ et on conclut en appliquant le théorème de convergence dominée.

(v) Remarquons tout d'abord que

$$\begin{aligned} \|X_n + Y_n\| \mathbb{1}_{\{\|X_n+Y_n\| \ge M\}} &\le (\|X_n\| + \|Y_n\|) \mathbb{1}_{\{\|X_n\|+\|Y_n\| \ge M\}} \\ &\le 2\|X_n\| \mathbb{1}_{\{\|X_n\|+\|Y_n\| \ge M\}} \mathbb{1}_{\{\|X_n\| \ge \|Y_n\|\}} + 2\|Y_n\| \mathbb{1}_{\{\|X_n\|+\|Y_n\| \ge M\}} \mathbb{1}_{\{\|Y_n\| > \|X_n\|\}} \\ &\le 2\|X_n\| \mathbb{1}_{\{2\|X_n\| \ge M\}} + 2\|Y_n\| \mathbb{1}_{\{2\|Y_n\| \ge M\}} . \end{aligned}$$

La preuve en découle immédiatement.

(vi) Par hypothèse, il existe une constante $C > 0$ telle que $\mathbb{P}(\sup_n \|Y_n\| \le C) = 1$. Par suite

$$\mathbb{E}\left[\|X_n Y_n\| \mathbb{1}_{\{\|X_n Y_n\| \ge M\}}\right] \le C\,\mathbb{E}\left[\|X_n\| \mathbb{1}_{\{\|X_n\| \ge M/C\}}\right]$$

ce qui conclut la preuve. □

Comme le montre le théorème ci-dessous, l'uniforme intégrabilité permet de relier la convergence en loi et la convergence des moments.

Théorème IV-5.55. *Soit $f : \mathbb{R}^k \mapsto \mathbb{R}$ une fonction borélienne continue en tout point de $C \in \mathscr{B}(\mathbb{R}^k)$. Supposons que $X_n \Longrightarrow X$ et $\mathbb{P}(X \in C) = 1$. Alors, $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ si et seulement si la suite $\{f(X_n), n \in \mathbb{N}\}$ est uniformément intégrable.*

Démonstration. Nous ne montrons ici que la réciproque. Posons $Y_n := f(X_n)$ et supposons que Y_n est uniformément intégrable. Nous allons montrer que $\mathbb{E}[Y_n] \to \mathbb{E}[Y]$, où $Y := f(X)$. Nous supposons sans perte de généralité que Y_n est positive (il suffit autrement de raisonner sur les parties positives et négatives séparément). Le théorème de continuité montre que $Y_n \Longrightarrow Y$. Nous notons $a \wedge b := \inf(a,b)$. L'inégalité triangulaire donne d'une part

$$\mathbb{E}[Y \wedge M] \le |\mathbb{E}[Y_n \wedge M] - \mathbb{E}[Y \wedge M]| + \mathbb{E}[Y_n \wedge M] \tag{IV-5.27}$$

et d'autre part

$$|\mathbb{E}[Y_n] - \mathbb{E}[Y]| \le |\mathbb{E}[Y_n] - \mathbb{E}[Y_n \wedge M]| + |\mathbb{E}[Y_n \wedge M] - \mathbb{E}[Y \wedge M]| + |\mathbb{E}[Y \wedge M] - \mathbb{E}[Y]|. \tag{IV-5.28}$$

Comme la fonction $y \to y \wedge M$ est continue et bornée, $|\mathbb{E}[Y_n \wedge M] - \mathbb{E}[Y \wedge M]| \to 0$ quand $n \to \infty$. Le second terme de la partie droite de (IV-5.27) est majoré indépendamment de M (voir ci-dessus), donc $\mathbb{E}[Y] < \infty$. Le premier terme et le troisième terme de la partie droite de (IV-5.28) peuvent être rendus arbitrairement petits en utilisant respectivement l'uniforme intégrabilité et $\mathbb{E}[Y] < \infty$, ce qui achève la démonstration de l'implication réciproque. □

Exemple IV-5.56. Soit $\{X_n,\, n\in\mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ telle que $X_n \Longrightarrow X$ et $\limsup \mathbb{E}[\|X_n\|^p] < \infty$ pour $p>0$. Alors, pour tout $0\le q<p$, nous avons $\mathbb{E}[\|X_n\|^q]\to\mathbb{E}[\|X\|^q]$. Le théorème précédent montre en effet qu'il suffit de vérifier que la suite $Y_n=\|X_n\|^q$ est uniformément intégrable. Cette propriété découle de l'inégalité de Markov :

$$\begin{aligned}\mathbb{E}\left[\|X_n\|^q\,\mathbb{1}_{\{\|X_n\|^q\ge M\}}\right] &= \mathbb{E}\left[\|X_n\|^q\,\mathbb{1}_{\{\|X_n\|^q/M\ge 1\}}\right]\\ &\le \mathbb{E}[\|X_n\|^q\,(\|X_n\|^q/M)^{(p-q)/q}] = M^{1-p/q}\mathbb{E}[\|X_n\|^p]\,,\end{aligned}$$

ce qui implique bien l'uniforme intégrabilité car $\sup_{n\ge0}\mathbb{E}[\|X_n\|^p]<\infty$. ♢

IV-5.7 Symboles *o* et *O* stochastiques

Définition IV-5.57 (suite bornée en probabilité). *Soit $\{X_n,\, n\in\mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ définis sur un espace de probabilité $(\Omega,\mathscr{F},\mathbb{P})$. La suite $\{X_n,\, n\in\mathbb{N}\}$ est dite* bornée en probabilité *si*

$$\lim_{M\to\infty}\limsup_{n\to\infty}\mathbb{P}(\|X_n\|\ge M)=0\,.$$

Si, pour $p>0$, nous avons $\sup_{n\in\mathbb{N}}\mathbb{E}[\|X_n\|^p]<\infty$, l'inégalité de Markov (Lemme IV-1.1) montre que $\{X_n,\, n\in\mathbb{N}\}$ est bornée en probabilité, car

$$\sup_{n\ge0}\mathbb{P}(\|X_n\|\ge M)\le M^{-p}\sup_{n\ge0}\mathbb{E}[\|X_n\|^p]\,.$$

On montre aisément que

Lemme IV-5.58. *Soit $\{X_n,\, n\in\mathbb{N}\}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$, définies sur l'espace de probabilité $(\Omega,\mathscr{F},\mathbb{P})$. Soit X un vecteur aléatoire à valeur $\mathbb{R}^d$. Si $X_n\Longrightarrow X$, alors la suite $\{X_n,\, n\in\mathbb{N}\}$ est bornée en probabilité.*

Il est pratique de disposer de notations simples pour exprimer qu'une suite tend vers 0 en probabilité ou est bornée en probabilité. Soit $\{X_n,\, n\in\mathbb{N}\}$ une suite de variables aléatoires (scalaires ou vectorielles), nous posons

$$X_n = o_P(1) \quad \text{signifie} \quad X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0, \tag{IV-5.29}$$

$$X_n = O_P(1) \quad \text{si la suite } \{X_n,\, n\in\mathbb{N}\} \text{ est bornée en probabilité} \tag{IV-5.30}$$

Plus généralement, pour $\{R_n,\, n\in\mathbb{N}\}$ une suite de variables aléatoires réelles, nous posons

$$X_n = o_P(R_n) \quad \text{signifie} \quad X_n = Y_n R_n \quad \text{avec} \quad Y_n = o_P(1)\,, \tag{IV-5.31}$$

$$X_n = O_P(R_n) \quad \text{signifie} \quad X_n = Y_n R_n \quad \text{avec} \quad Y_n = O_P(1)\,. \tag{IV-5.32}$$

Si $\{X_n,\, n\in\mathbb{N}\}$ et $\{R_n,\, n\in\mathbb{N}\}$ sont des suites déterministes, les symboles o_P et O_P coïncident avec les notations de Landau o et O couramment utilisées en analyse. Les règles de calcul des symboles o_P et O_P coïncident avec les règles de Landau. A titre d'exemple,

$$\begin{gathered} o_P(1)+o_P(1)=o_P(1),\\ o_P(1)+O_P(1)=O_P(1),\\ O_P(1)o_P(1)=o_P(1),\\ (1+o_P(1))^{-1}=1+o_P(1),\\ o_P(R_n)=R_n o_P(1),\quad O_P(R_n)=R_n O_P(1)\,.\end{gathered}$$

Pour s'assurer de la validité de ces règles, il suffit de les re-écrire explicitement avec des suites et d'utiliser les résultats classiques énoncés ci-dessus. Par exemple, si $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$ et $Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$, le théorème IV-5.32 implique que $(X_n, Y_n) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} (0,0)$, ce qui équivaut à $(X_n, Y_n) \Longrightarrow (0,0)$. Le théorème de continuité (appliqué à $f : (x,y) \mapsto x+y$) implique $X_n + Y_n \Longrightarrow 0$, qui équivaut à $X_n + Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$. La troisième règle est une façon concise d'écrire : si $\{X_n,\ n \in \mathbb{N}\}$ est bornée en probabilité et $Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$, alors $X_n Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$. Si $X_n \Longrightarrow X$, alors ce résultat découle du lemme de Slutsky (car $X_n \Longrightarrow X$ et $Y_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} c$ implique que $Y_n X_n \Longrightarrow cX$, donc si $c = 0$, $Y_n X_n \Longrightarrow 0$ qui équivaut à $Y_n X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$). Dans le cas où $\{X_n,\ n \in \mathbb{N}\}$ ne converge pas en probabilité, on peut donner aisément une preuve directe.

La règle de calcul suivante est utile pour les développements asymptotiques.

Lemme IV-5.59. *Soit R une fonction définie sur un voisinage $D \subset \mathbb{R}^k$ de 0 telle que $R(0) = 0$. Soit $\{X_n,\ n \in \mathbb{N}\}$ une suite de v.a. à valeurs dans D telle que $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} 0$. Alors, pour tout $p > 0$,*
(i) *Si $R(h) = o(\|h\|^p)$ quand $h \to 0$, alors $R(X_n) = o_P(\|X_n\|^p)$,*
(ii) *Si $R(h) = O(\|h\|^p)$ quand $h \to 0$, alors $R(X_n) = O_P(\|X_n\|^p)$.*

Démonstration. Définissons $g(h) = R(h)/\|h\|^p$ pour $h \neq 0$ et $g(0) = 0$: $R(X_n) = g(X_n)\,\|X_n\|^p$.
(i) La fonction g est continue en 0 et le théorème de continuité (Théorème IV-5.6) montre que $g(X_n) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} g(0) = 0$. La deuxième assertion se démontre de façon similaire. □

Exemple IV-5.60 (Variance empirique). Soit $\{Y_k,\ k \in \mathbb{N}\}$ une suite de variables aléatoires i.i.d. telles que $m_4 := \mathbb{E}[Y_1^4] < \infty$; on note μ l'espérance de Y_1 et σ^2 la variance de Y_1.

Considérons l'estimateur de la variance empirique, $S_n^2 := n^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, où $\bar{Y}_n := n^{-1}\sum_{i=1}^n Y_i$ est la moyenne empirique. Nous allons montrer que $\sqrt{n}(S_n^2 - \sigma^2)$ converge en loi vers une gaussienne centrée. Nous avons :

$$S_n^2 = n^{-1}\sum_{i=1}^n (Y_i - \mu)^2 - (\bar{Y}_n - \mu)^2$$
$$\sqrt{n}(S_n^2 - \sigma^2) = n^{-1/2}\sum_{i=1}^n \Big((Y_i - \mu)^2 - \sigma^2\Big) - \sqrt{n}(\bar{Y}_n - \mu)^2.$$

Notons $Z_i := (Y_i - \mu)^2 - \sigma^2$. Les variables aléatoires $\{Z_i,\ i \in \mathbb{N}\}$ sont i.i.d. centrées et de variance finie. Nous pouvons donc utiliser le T.L.C. (Théorème IV-5.39).

$$n^{-1/2}\sum_{i=1}^n Z_i \quad \Longrightarrow \quad \mathrm{N}(0,\gamma), \quad \text{où } \gamma := \mathbb{E}(Y_1 - \mu)^4 - \sigma^4\,.$$

Le T.L.C. montre aussi que $\sqrt{n}(\bar{Y}_n - \mu) \quad \Longrightarrow \quad \mathrm{N}(0,\sigma^2)$; par conséquent la suite $\{\sqrt{n}(\bar{Y}_n - \mu), n \in \mathbb{N}\}$ est bornée en probabilité i.e. $\bar{Y}_n - \mu = O_P(n^{-1/2})$ (d'après le lemme IV-5.58). Le lemme IV-5.59 montre que $n^{1/2}(\bar{Y}_n - \mu)^2 = O_P(n^{-1/2}) = o_P(1)$. Le lemme de Slutsky (Lemme IV-5.33) montre donc :

$$\sqrt{n}(S_n^2 - \sigma^2) \quad \Longrightarrow \quad \mathrm{N}(0,\gamma).$$ ♢

Cinquième partie

Annexe mathématique

Annexe A

Eléments de théorie de la mesure

Dans ce chapitre, nous présentons de façon succincte les principaux résultats de la théorie de la mesure et de l'intégration utilisés dans le livre. Ces notes sont concises et le lecteur soucieux d'approfondir ces notions pourra consulter, en français [**?**, **?**].

A.1 Tribus et Mesurabilité

A.1.1 Tribus

Définition A.1 (Algèbre, Tribu). *Soit* E *un ensemble et* $\mathscr{E}$ *un ensemble de parties de* E*. On dit que* $\mathscr{E}$ *est une* algèbre *sur* E *si elle vérifie les propriétés suivantes :*

— $\mathsf{E} \in \mathscr{E}$,

— $\mathscr{E}$ *est stable par passage au complémentaire : pour tout* $A \in \mathscr{E}$*, on a* $A^c := \mathsf{E} \setminus A \in \mathscr{E}$,

— $\mathscr{E}$ *est stable par réunion finie : pour* A, B *éléments de* $\mathscr{E}$*,* $A \cup B \in \mathscr{E}$*.*

On dit que $\mathscr{E}$ *est une* tribu *sur* E *(ou* σ-algèbre*) si elle vérifie les propriétés suivantes :*

— $\mathsf{E} \in \mathscr{E}$,

— $\mathscr{E}$ *est stable par passage au complémentaire : pour tout* $A \in \mathscr{E}$*, on a* $A^c \in \mathscr{E}$,

— $\mathscr{E}$ *est stable par réunion dénombrable : pour toute suite* $\{A_n,\, n \in \mathbb{N}\}$ *d'éléments de* $\mathscr{E}$*,* $\bigcup_{n=0}^{\infty} A_n \in \mathscr{E}$*.*

Remarquons que si $\mathscr{E}$ est une tribu sur E, alors $\emptyset \in \mathscr{E}$ puisque c'est le complémentaire (dans E) de E. Remarquons aussi que si $\{A_n,\, n \in \mathbb{N}\}$ est une suite d'éléments de $\mathscr{E}$ alors $\bigcap_{n=0}^{\infty} A_n \in \mathscr{E}$ puisque cet ensemble est le complémentaire de $\bigcup_{n=0}^{\infty} A_n^c$. Par suite, $\mathscr{E}$ est stable par *intersection* dénombrable.

En considérant des suites $\{A_n,\, n \in \mathbb{N}\}$ constantes à partir d'un certain rang, on prouve que $\mathscr{E}$ est stable par réunion et intersection finies. La notion de tribu est plus générale que celle d'algèbre, qui n'est pas assez riche pour développer une théorie satisfaisante de la mesure ou de l'intégration : en effet, la notion de σ-additivité est nécessaire dès lors que l'on veut parler de limites d'ensembles ou de fonctions.

De la Définition A.1 et des propriétés de stabilité par intersection et réunion finies ou dénombrables, on déduit les propriétés suivantes : Si $\mathscr{E}$ est une tribu, alors

(i) Si $A, B \in \mathscr{E}$, alors $A \setminus B := A \cap B^c \in \mathscr{E}$,

(ii) Si $A, B \in \mathscr{E}$ alors $A \Delta B := (A \setminus B) \cup (B \setminus A) \in \mathscr{E}$.

Pour une suite $\{A_n,\, n \in \mathbb{N}\}$ de parties de E, définissons la limite inférieure et la limite supérieure de ces ensembles par

$$\liminf_{n\to\infty} A_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k, \qquad \limsup_{n\to\infty} A_n := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

$\liminf_{n\to\infty} A_n$ est l'ensemble des éléments de E qui appartiennent à tous les A_n à partir d'un certain rang, et $\limsup_{n\to\infty} A_n$ est l'ensemble des éléments de E qui appartiennent à une infinité de A_n. On a

(iii) Si $\{A_n,\, n\in\mathbb{N}\}$ est une suite d'éléments de $\mathscr{E}$, alors

$$\liminf_n A_n \in \mathscr{E}, \qquad \limsup_n A_n \in \mathscr{E}\,.$$

Il n'est souvent pas possible de décrire les éléments d'une tribu. On remarque que l'ensemble des parties de E, noté $\mathscr{P}(\mathsf{E})$, est une tribu sur E et que l'intersection d'une famille quelconque de tribus sur E est une tribu sur E. Donc, étant donné $\mathscr{C}\subset\mathscr{P}(\mathsf{E})$, on peut considérer la plus petite tribu contenant $\mathscr{C}$.

Définition A.2 (Tribu engendrée). *Soit $\mathscr{C}\in\mathscr{P}(\mathsf{E})$. La* tribu engendrée *par $\mathscr{C}$ (sur E) est l'intersection de toutes les tribus sur E contenant $\mathscr{C}$. Cette tribu se note $\sigma(\mathscr{C})$.*

Définition A.3 (Tribu borélienne). *Soit E un espace topologique et soit $\mathscr{O}$ la classe des ouverts de E. La tribu $\sigma(\mathscr{O})$ s'appelle la* tribu borélienne *de E et se note $\mathscr{B}(\mathsf{E})$.*

Il est facile de voir que la tribu borélienne est aussi engendrée par les fermés. Dans la plupart des cas, nous considérerons l'espace topologique $\mathsf{E}=\mathbb{R}^d$ muni de sa topologie usuelle (associée à la distance euclidienne). Dans ce cas, $\mathscr{B}(\mathbb{R}^d)$ est aussi la tribu engendrée par les boules, par les pavés et même par les pavés à coordonnées rationnelles (cette dernière famille ayant l'avantage d'être dénombrable). En dimension $d=1$, outre la tribu borélienne $\mathscr{B}(\mathbb{R})$, on considérera

$$\begin{aligned}
&\mathscr{B}(\mathbb{R}^+) := \left\{A\in\mathscr{B}(\mathbb{R})\,,\, A\subset\mathbb{R}^+\right\},\\
&\mathscr{B}(\overline{\mathbb{R}}) := \sigma(\mathscr{B}(\mathbb{R}),\{+\infty\},\{-\infty\}),\\
&\mathscr{B}(\overline{\mathbb{R}^+}) := \sigma(\mathscr{B}(\mathbb{R}^+),\{+\infty\}).
\end{aligned}$$

On prolonge à $\overline{\mathbb{R}^+}$ les relations d'ordre, l'addition et la multiplication de $\mathbb{R}^+$ en posant

pour tout $a\in\mathbb{R}^+ : a<+\infty$.
pour tout $a\in\overline{\mathbb{R}^+} : a+(+\infty)=(+\infty)+a=+\infty$.
pour tout $a\in\overline{\mathbb{R}^+}\setminus\{0\} : a(+\infty)=(+\infty)a=+\infty$.
$0(+\infty)=(+\infty)0=0$.

Le ratio $0/0$ n'est pas défini. On prolonge à $\overline{\mathbb{R}}$ les relations d'ordre, l'addition et la multiplication de $\mathbb{R}$ en posant

pour tout $x\in\mathbb{R}$, $-\infty<x<+\infty$.
pour tout $x\in\,]-\infty,+\infty] : x+(+\infty)=(+\infty)+x=+\infty$.
pour tout $x\in[-\infty,+\infty[: x+(-\infty)=(-\infty)+x=-\infty$.
pour tout $x\in[-\infty,+\infty] : x(+\infty)=+\infty$ si $x\geq 0$ et $-\infty$ sinon.
pour tout $x\in[-\infty,+\infty] : x(-\infty)=-\infty$ si $x\geq 0$ et $+\infty$ sinon.

Quand l'espace d'arrivée est $\mathbb{R}$, $\overline{\mathbb{R}^+}$, $\overline{\mathbb{R}}$ ou $\mathbb{R}^d$, cet espace est, sauf mention explicite, muni de sa tribu borélienne.

A.1.2 Classes monotones (ou λ-systèmes)

Définition A.4 (Classe monotone, λ-système). *Soit* E *un ensemble. Une famille* $\mathscr{M}$ *d'éléments de* $\mathscr{P}(\mathsf{E})$ *est appelée une* classe monotone *(ou un* λ-système*) sur* E *si*

(i) $\mathsf{E} \in \mathscr{M}$,

(ii) $A, B \in \mathscr{M}$, $A \subset B \Longrightarrow B \setminus A \in \mathscr{M}$,

(iii) pour toute suite $\{A_n,\ n \in \mathbb{N}\}$ *d'éléments de* $\mathscr{M}$ *tels que* $A_n \subset A_{n+1}$, $\bigcup_{n=1}^{\infty} A_n \in \mathscr{M}$.

Une tribu est une classe monotone. L'intersection d'une famille quelconque de classes monotones est une classe monotone. Pour une famille quelconque $\mathscr{C}$ de sous-ensembles de E, il existe une *plus petite classe monotone contenant* $\mathscr{C}$ qui est l'intersection de toutes les classes monotones contenant $\mathscr{C}$.

Lemme A.5. *Soit* E *un ensemble. Si* $\mathscr{N}$ *est une classe monotone sur* E*, stable par intersection finie, alors* $\mathscr{N}$ *est une tribu sur* E.

Démonstration. Par définition d'une classe monotone, $\mathsf{E} \in \mathscr{N}$. De plus, comme $\mathscr{N}$ est stable par différence, si $A \in \mathscr{N}$ alors $A^c \in \mathscr{N}$. Montrons la stabilité par union dénombrable.

Soit $\{B_k,\ k \in \mathbb{N}\}$ une suite d'éléments de $\mathscr{N}$; pour tout $n \in \mathbb{N}$, posons $A_n := \bigcup_{k=1}^{n} B_k \in \mathscr{N}$. Alors la stabilité par intersection finie et par passage au complémentaire impliquant la stabilité par union finie, on a $A_n \in \mathscr{N}$. Donc $\{A_n,\ n \in \mathbb{N}\}$ est une suite croissante d'éléments de $\mathscr{N}$. Par définition de la classe monotone, on a donc $\bigcup_{n=1}^{\infty} A_n \in \mathscr{N}$ ce qui entraîne $\bigcup_{k=1}^{\infty} B_k \in \mathscr{N}$ puisque $\bigcup_{n=1}^{\infty} A_n = \bigcup_{k=1}^{\infty} B_k \in \mathscr{N}$. □

Théorème A.6. *Soit* E *un ensemble et* $\mathscr{M}$ *une classe monotone sur* E*. Soient* $\mathscr{C} \subset \mathscr{M}$*, une famille d'ensembles stable par intersection finie. Alors* $\sigma(\mathscr{C}) \subset \mathscr{M}$.

Démonstration. Soit $\mathscr{S}$ la plus petite classe monotone contenant $\mathscr{C}$. Nous allons montrer que $\mathscr{S}$ est stable par intersection finie ; nous conclurons par Lemme A.5 que c'est une tribu, et c'est nécessairement la plus petite tribu sinon cela contredirait le fait que $\mathscr{S}$ est la plus petite classe monotone.

Pour $A \subset \mathsf{E}$, on pose $\mathscr{D}_A := \{B \in \mathscr{S},\ A \cap B \in \mathscr{S}\}$. Considérons le cas où $A \in \mathscr{S}$; nous allons montrer que dans ce cas, $\mathscr{D}_A$ est une classe monotone puisqu'elle vérifie les propriétés suivantes :

(i) $\mathsf{E} \in \mathscr{D}_A$. En effet, $\mathsf{E} \in \mathscr{S}$ et $\mathsf{E} \cap A = A \in \mathscr{S}$.

(ii) si $B_1, B_2 \in \mathscr{D}_A$ avec $B_1 \subset B_2$, alors $B_2 \setminus B_1 \in \mathscr{D}_A$. En effet, on sait que $A \cap B_i \in \mathscr{S}$ et que $(A \cap B_1) \subset (A \cap B_2)$; par propriété des classes monotones, il vient

$$A \cap (B_2 \setminus B_1) = (A \cap B_2) \setminus (A \cap B_1) \in \mathscr{S}.$$

Par suite, $B_2 \setminus B_1 \in \mathscr{D}_A$.

(iii) Soit une suite croissante $\{B_n,\ n \in \mathbb{N}\}$, d'élements de $\mathscr{D}_A$; alors $\bigcup_n B_n \in \mathscr{D}_A$. En effet, la famille d'ensembles $\{A \cap B_n, n \in \mathbb{N}\}$ est aussi croissante, et chaque ensemble est élément de $\mathscr{S}$. Donc, par propriété des classes monotones,

$$A \cap \bigcup_n B_n = \bigcup_n (A \cap B_n) \in \mathscr{S}.$$

Dans le cas où $A \in \mathscr{C}$, alors puisque pour tout $B \in \mathscr{C}$ on a $A \cap B \in \mathscr{C} \subset \mathscr{S}$, il vient que $B \in \mathscr{D}_A$; donc $\mathscr{C} \subset \mathscr{D}_A$.

En combinant ces résultats, nous voyons que pour $A \in \mathscr{C} \subset \mathscr{S}$, la famille $\mathscr{D}_A$ est une classe monotone qui contient $\mathscr{C}$. Par suite, on a $\mathscr{S} \subset \mathscr{D}_A$.

Soit $A \in \mathscr{C}$ et $B \in \mathscr{S}$. Puisque $\mathscr{S} \subset \mathscr{D}_A$, il vient que $B \in \mathscr{D}_A$ et donc que $A \in \mathscr{D}_B$. On en déduit que $\mathscr{C} \subset \mathscr{D}_B$. Mais $\mathscr{D}_B$ étant une classe monotone, alors $\mathscr{S} \subset \mathscr{D}_B$.

Par conséquent, soit $B, B' \in \mathscr{S}$. Vu ce que l'on vient d'établir, $B' \in \mathscr{D}_B$ et donc $B \cap B' \in \mathscr{S}$. Donc $\mathscr{S}$ est stable par intersection finie. □

A.1.3 Applications mesurables

Définition A.7 (Espace mesurable). *Un* espace mesurable *est un couple* $(\mathsf{E}, \mathscr{E})$ *où* $\mathscr{E}$ *étant une tribu sur* E.

Pour une application $f : \mathsf{E}_1 \to \mathsf{E}_2$ et pour $A \in \mathsf{E}_2$, on note $f^{-1}(A)$ l'image réciproque par f de A défini par

$$f^{-1}(A) := \{x \in \mathsf{E}_1 \ : f(x) \in A\}.$$

Définition A.8 (Applications mesurables). *Soient* $(\mathsf{E}_1, \mathscr{E}_1)$ *et* $(\mathsf{E}_2, \mathscr{E}_2)$ *deux espaces mesurables. Une application* f *de* E_1 *dans* E_2 *est dite* mesurable *si, pour tout* $A \in \mathscr{E}_2$, $f^{-1}(A) \in \mathscr{E}_1$.

Proposition A.9. *Soit* $f : (\mathsf{E}_1, \mathscr{E}_1) \to (\mathsf{E}_2, \mathscr{E}_2)$. *Lorsque* $\mathscr{E}_2$ *est la tribu engendrée par la famille de parties* $\mathscr{C} \subset \mathscr{P}(\mathsf{E})$, f *est mesurable si et seulement si* $f^{-1}(A) \in \mathscr{E}_1$ *pour tout* $A \in \mathscr{C}$.

Démonstration. L'implication directe est triviale puisque $\mathscr{C} \in \mathscr{E}_2$, et pour l'implication réciproque, il suffit de montrer que l'ensemble $\{A \in \mathsf{E}_2 \ : f^{-1}(A) \in \mathsf{E}_1\}$ des parties de E_2 est une tribu sur E_2 qui contient $\mathscr{C}$ et elle contient donc en partie ma tribu engendrée $\mathscr{E}_2$. □

Cette caractérisation entraîne que si f est continue de $\mathbb{R}^d$ dans $\mathbb{R}^m$, alors f est *borélienne* i.e. mesurable pour les tribus boréliennes.

De plus, la notion de *mesurabilité* est transitive : la composée de deux applications mesurables est mesurable.

Soit $(\mathsf{E}, \mathscr{E})$ un espace mesurable. Si pour tout $x \in \mathsf{E}$, la suite $\{f_n(x), n \geq 1\}$ croît (resp. décroît) vers $f(x)$, on écrira $f_n \uparrow f$ (resp. $f_n \downarrow f$). Comme les suites $\{\sup_{k \geq n} f_k, n \geq 0\}$ et $\{\inf_{k \geq n} f_k, n \geq 0\}$ sont respectivement décroissantes et croissantes, on définit leur limite que l'on note resp. limite supérieure et limite inférieure

$$\liminf f_n(x) = \lim_n \uparrow \inf_{k \geq n} f_k(x) \qquad \limsup f_n(x) = \lim_n \downarrow \sup_{k \geq n} f_k(x), ; \tag{A.1}$$

ces quantités sont à valeur dans $\overline{\mathbb{R}}$. Lorsqu'elles sont égales, on dit que la suite $\{f_n, n \geq 0\}$ converge vers cette valeur commune : $f = \lim_n f_n$ si et seulement si $\limsup f_n = \liminf f_n = f$.

Pour qu'une application numérique $f : \mathsf{E} \to \overline{\mathbb{R}}$ soit mesurable, il suffit que, pour tout $a \in \overline{\mathbb{R}}$, $\{f > a\} := \{x \ : f(x) > a\} \in \mathscr{E}$. On peut aussi considérer $\{f < a\}$, $\{f \leq a\}$ ou $\{f \geq a\}$. Ceci implique que, si f, g, f_n sont des fonctions numériques mesurables, il en est de même de $-f$, $\sup(f,g)$, $\inf(f,g)$, $f^+ := \sup(f, 0)$, $f^- := \sup(-f, 0)$, $\sup f_n$, $\inf f_n$, $\limsup f_n$, $\liminf f_n$, et, si elle existe : $\lim f_n$.

On appelle *fonction indicatrice de A*, la fonction notée $\mathbb{1}_A$ valant

$$\mathbb{1}_A(x) := \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

On a $\mathbb{1}_{A^c} = 1 - \mathbb{1}_A$. Si $\{A_n,\, n \in \mathbb{N}\}$ est une suite d'éléments de $\mathscr{P}(\mathsf{E})$, $\mathbb{1}_{A_n}$ est une application mesurable. Les relations

$$\mathbb{1}_{\bigcap_{n=0}^{\infty} A_n} = \prod_{n=0}^{\infty} \mathbb{1}_{A_n} = \inf_{n\geq 0} \mathbb{1}_{A_n}, \qquad \mathbb{1}_{\bigcup_{n=0}^{\infty} A_n} = \sup_{n\geq 0} \mathbb{1}_{A_n},$$

entraînent que

$$\liminf \mathbb{1}_{A_n} = \mathbb{1}_{\liminf A_n}, \qquad \limsup \mathbb{1}_{A_n} = \mathbb{1}_{\limsup A_n}.$$

Soient f, g des fonctions numériques mesurables définies sur l'espace mesurable $(\mathsf{E}, \mathscr{E})$. Alors $\phi : x \mapsto (f(x), g(x))$ est mesurable de $(\mathsf{E}, \mathscr{E})$ dans $\mathbb{R}^2$ puisque pour tout pavé $A \times B$ de $\mathbb{R}^2$, $\phi^{-1}(A \times B) = f^{-1}(A) \cap g^{-1}(B) \in \mathscr{E}$. Ceci implique que, si H est une application borélienne de $\mathbb{R}^2$ dans $\mathbb{R}$, $H(f,g)$ est mesurable. On en déduit la propriété suivante

Proposition A.10. *Soit $(\mathsf{E}, \mathscr{E})$ un espace mesurable et $f, g : \mathsf{E} \to \overline{\mathbb{R}}$ deux fonctions mesurables. Alors les fonctions $f+g$, fg, f/g (si elles existent), sont mesurables.*

Le prolongement de l'addition et de la multiplication à la droite achevée sont rappelées en section A.1 ; en particulier, les opérations $0/0$, $(+\infty) \times (-\infty)$ ne sont pas définies.

Le résultat clé pour définir la mesure d'une fonction, est que toute fonction mesurable positive est la limite croissante d'une suite de fonctions plus simples dites *fonctions étagées* pour lesquelles on sait aisément définir la mesure. La notion de fonction réelle étagée joue donc un rôle important similaire à celui des fonctions en escaliers en théorie de l'intégration. Nous introduisons ces fonctions puis prouvons le résultat de limite.

Définition A.11 (Fonction réelle étagée). *Soit $(\mathsf{E}, \mathscr{E})$ un ensemble mesurable. Une application $f : \mathsf{E} \to \mathbb{R}$ est dite* étagée *si elle s'écrit*

$$f = \sum_{k=1}^{p} a_k \mathbb{1}_{A_k}, \quad A_k \in \mathscr{E}.$$

Proposition A.12. *Toute fonction f mesurable à valeur dans $\overline{\mathbb{R}^+}$ est limite d'une suite croissante de fonctions étagées positives.*

Démonstration. Il suffit de considérer la fonction étagée

$$f_n(x) := \sum_{k=0}^{n2^n-1} 2^{-n} k \mathbb{1}_{\{x\,:\, 2^{-n}k \leq f(x) < 2^{-n}(k+1)\}} + n \mathbb{1}_{\{x\,:\, f(x)\geq n\}}. \tag{A.2}$$

□

Dans la suite, nous noterons :

$[\mathscr{E}]$ l'ensemble des fonctions réelles mesurables,
$b\mathscr{E}$ l'ensemble des fonctions réelles mesurables bornées,
$\mathscr{E}^+$ l'ensemble des fonctions mesurables à valeurs $\overline{\mathbb{R}^+}$,
$e\mathscr{E}^+$ l'ensemble des fonctions étagées positives.

A.1.4 Mesurabilité pour une tribu engendrée par des applications

Théorème A.13. *Soient $\mathscr{H}$ un espace vectoriel de fonctions numériques bornées définies sur E vérifiant*

$$1 \in \mathscr{H}, \ \ f_n \in \mathscr{H} \text{ et } \quad 0 \leq f_n \uparrow f \text{ bornée} \Rightarrow f \in \mathscr{H}. \tag{A.3}$$

Soit $\mathscr{C}$ un ensemble de parties de E stable par intersection finie et tel que $A \in \mathscr{C} \Rightarrow \mathbb{1}_A \in \mathscr{H}$.
Alors $\mathscr{H}$ contient toutes les fonctions $\sigma(\mathscr{C})$-mesurables bornées.

Démonstration. Soit $\mathscr{M} := \{A \subset \mathsf{E} \ : \mathbb{1}_A \in \mathscr{H}\}$. Alors $\mathscr{M}$ est une classe monotone sur E ; on a en effet $\mathbb{1}_E = 1 \in \mathscr{H}$; $\mathbb{1}_{B\setminus A} = \mathbb{1}_B - \mathbb{1}_A$ et $\mathscr{H}$ est un espace vectoriel ; pour une suite croissante $\{A_n,\ n \in \mathbb{N}\}$ d'éléments de $\mathscr{M}$, $\mathbb{1}_{\bigcup_n A_n} = \mathbb{1}_{\lim A_n} = \lim \uparrow \mathbb{1}_{A_n} \in \mathscr{H}$.

Puisque $\mathscr{C} \subset \mathscr{M}$, on peut appliquer le Théorème A.6 et l'on a $\sigma(\mathscr{C}) \subset \mathscr{M}$.

Soit $f : \mathsf{E} \to \mathbb{R}$ une fonction $\sigma(\mathscr{C})$-mesurable, étagée. Donc c'est une combinaison linéaire (finie) d'indicatrices d'éléments de $\sigma(\mathscr{C})$ et donc indicatrices d'éléments de $\mathscr{M}$. Par suite, $f \in \mathscr{H}$. Considérons maintenant le cas où f est une fonction positive bornée $\sigma(\mathscr{C})$-mesurable ; alors c'est la limite croissante de fonctions positives étagées $\sigma(\mathscr{C})$-mesurable (et donc de fonctions dans $\mathscr{H}$) ce qui entraîne que $f \in \mathscr{H}$. Enfin, pour toute fonction f bornée $\sigma(\mathscr{C})$-mesurable, on a $f \in \mathscr{H}$. □

Définition A.14 (Tribu engendrée par une application). *Soit X une application de Ω dans un espace mesurable $(\mathsf{E}, \mathscr{E})$. On appelle* tribu engendrée par *X la plus petite tribu sur Ω rendant X mesurable ; on la note $\sigma(X)$.*

Si $\{X_i,\ i \in I\}$ est une famille d'applications $X_i : \Omega \to (\mathsf{E}_i, \mathscr{E}_i)$, on appelle tribu engendrée par les applications *X_i la plus petite tribu sur Ω rendant toutes les applications X_i mesurables ; on la note $\sigma(X_i, i \in I)$.*

On a donc

$$\sigma(X) = \left\{X^{-1}(A)\,,\, A \in \mathscr{E}\right\}\ .$$

Soit

$$\mathscr{C} := \left\{ A \subset \Omega, A = \bigcap_{k=1}^{n} X_{i_k}^{-1}(\Gamma_k),\ \Gamma_k \in \mathscr{E}_{i_k},\ i_1, \ldots, i_n \in I \right\}.$$

$\mathscr{C}$ est stable par intersection finie et l'on a $\sigma(\mathscr{C}) = \sigma(X_i,\ i \in I)$.

Proposition A.15. *Pour tout $i \in I$, soit $(\mathsf{E}_i, \mathscr{E}_i)$ un espace mesurable et $X_i : \Omega \to \mathsf{E}_i$ une application. Une application $\Phi : (A, \mathscr{A}) \to (\Omega, \sigma(X_i,\ i \in I))$ est mesurable si et seulement si pour tout $i \in I$, $X_i \circ \Phi$ est mesurable de $(A, \mathscr{A})$ dans $(\mathsf{E}_i, \mathscr{E}_i)$.*

Démonstration. L'implication directe est une conséquence de la mesurabilité de la composée de deux applications mesurables. Pour le sens réciproque, il suffit (voir Proposition A.9) de vérifier que $\Phi^{-1}(C) \in \mathscr{A}$ pour tout C de la forme $X_i^{-1}(\Gamma_i)$ où $\Gamma_i \in \mathscr{E}_i$. Cela est vrai puisque

$$\Phi^{-1}(C) = \{a \in A \ : X_i \circ \Phi(a) \in \Gamma_i\},$$

et que l'ensemble de droite est un élément de $\mathscr{E}_i$ par hypothèse. □

Le Théorème A.13 implique alors :

Corollaire A.16. *Soient $\mathscr{H}$ un espace vectoriel de fonctions numériques bornées définies sur Ω. Pour tout $i \in I$, soit $(\mathsf{E}_i, \mathscr{E}_i)$ un espace mesurable et une application $X_i : \Omega \to \mathsf{E}_i$. On suppose que $\mathscr{H}$ vérifie (A.3) et que, pour tout $i_1, \dots, i_n \in I$ et tout $\Gamma_k \in \mathscr{E}_{i_k}$,*

$$\prod_{k=1}^{n} \mathbb{1}_{\Gamma_k} \circ X_{i_k} \in \mathscr{H}.$$

Alors $\mathscr{H}$ contient toutes les fonctions $\sigma(X_i, i \in I)$-mesurables bornées.

On suppose que, pour tout $i \in I$, $(\mathsf{E}_i, \mathscr{E}_i) = (\mathsf{E}, \mathscr{E})$. On note $F := \mathsf{E}^{\mathbb{N}}$. Pour $x = \{x_n,\ n \in \mathbb{N}\} \in F$, on définit $\xi_n : F \to \mathsf{E}$ par $\xi_n(x) := x_n$ et on pose $\mathscr{F} = \sigma(\xi_n, n \in \mathbb{N})$.

Corollaire A.17. *Soient, pour tout $i \in I$, $X_i : \Omega \to (\mathsf{E}, \mathscr{E})$ et $Y : \Omega \to \mathbb{R}$ (resp. $\Omega \to \overline{\mathbb{R}^+}$). Alors Y est $\sigma(X_i, i \in I)$-mesurable si et seulement si il existe $i_1, \dots, i_n, \dots \in I$ et $h : F \to \mathbb{R}$ (resp. $h : F \to \mathbb{R}^+$) $\mathscr{F}$-mesurable telle que $Y = h(X_{i_1}, \dots, X_{i_n}, \dots)$.*

Démonstration. Vu Proposition A.15, si $h \in b\mathscr{F}$, $h(X_{i_1}, \dots, X_{i_n}, \dots) \in b\sigma(X_i, i \in I)$. Dans l'autre sens, soit

$$\mathscr{H} = \{Z : \Omega \to \mathbb{R};\ Z = h(X_{i_1}, \dots, X_{i_n}, \dots),\ i_k \in I,\ h \in b\mathscr{F}\}.$$

On vérifie (assez) facilement que $\mathscr{H}$ est un espace vectoriel de fonctions numériques bornées vérifiant (A.3) et contenant $\prod_{k=1}^{n} \mathbb{1}_{\Gamma_k}(X_{i_k})$. Appliquant le Corollaire A.16, $\mathscr{H} \supset b\sigma(X_i, i \in I)$. □

Corollaire A.18. *Soit $\mathscr{H}$ un espace vectoriel de fonctions numériques bornées définies sur $\mathbb{R}^d$. On suppose que $\mathscr{H}$ vérifie (A.3) et contient toutes les fonctions continues à support compact. Alors $\mathscr{H} \supset b\mathscr{B}(\mathbb{R}^d)$.*

Démonstration. En effet, pour tout U ouvert borné, on a $\mathbb{1}_U = \lim \uparrow f_n$ avec f_n continue à support compact ; il suffit de prendre $f_n(x) := 1 \wedge n\mathrm{d}(x, U^c)$ où $\mathrm{d}(x, U^c)$ désigne la distance de x au fermé U^c. Donc $\mathbb{1}_U \in \mathscr{H}$ et on applique le Théorème A.13. □

Enfin combinant les Corollaires A.16 and A.18, on obtient, notant $C_k(\mathbb{R}^d)$ l'espace des fonctions continues à support compact sur $\mathbb{R}^d$,

Corollaire A.19. *Soient $\mathscr{H}$ un espace vectoriel de fonctions numériques bornées définies sur Ω et $\{X_i,\ i \in I\}$ une famille d'applications de Ω dans $\mathbb{R}^d$. On suppose que $\mathscr{H}$ vérifie (A.3) et que, pour tout $i_1, \dots, i_n \in I$ et toute fonction $f_j \in C_k(\mathbb{R}^d)$,*

$$\prod_{j=1}^{n} f_j \circ X_{i_j} \in \mathscr{H}.$$

Alors $\mathscr{H}$ contient toutes les fonctions $\sigma(X_i, i \in I)$-mesurables bornées.

A.1.5 Espaces produits

Étant donné deux ensembles arbitraires E_1 et E_2, on note

$$\mathsf{E}_1 \times \mathsf{E}_2 := \{(e_1, e_2)\ : e_1 \in \mathsf{E}_1, e_2 \in \mathsf{E}_2\}.$$

L'application $X_i : \mathsf{E}_1 \times \mathsf{E}_2 \to \mathsf{E}_i$ ($i = 1,2$) qui à (e_1, e_2) fait correspondre e_i s'appelle *l'application coordonnée i*.

Si A est un sous-ensemble arbitraire de $\mathsf{E}_1 \times \mathsf{E}_2$, on note

$$A_{e_1} := \{e_2 \in \mathsf{E}_2 \ : (e_1, e_2) \in A\},$$

dite la *section* de A en e_1. Pour tout $e_1 \in \mathsf{E}_1$ (fixé), l'application $A \mapsto A_{e_1}$ de $\mathscr{P}(\mathsf{E}_1 \times \mathsf{E}_2) \to \mathscr{P}(\mathsf{E}_2)$ est un homomorphisme pour les opérations de réunion, d'intersection et de complémentation : en effet, si $\{A^\alpha\}$ est une famille d'éléments de $\mathscr{P}(\mathsf{E}_1 \times \mathsf{E}_2)$, on a

$$\left(\bigcup_\alpha A^\alpha\right)_{e_1} = \bigcup_\alpha A^\alpha_{e_1}, \qquad \left(\bigcap_\alpha A^\alpha\right)_{e_1} = \bigcap_\alpha A^\alpha_{e_1}, \qquad (A^c)_{e_1} = (A_{e_1})^c.$$

Si f est une application arbitraire de $\mathsf{E}_1 \times \mathsf{E}_2$ dans un espace quelconque E, on note

$$f_{e_1} : e_2 \mapsto f(e_1, e_2),$$

dite la *section* de f en e_1. Pour justifier cette terminologie, on remarquera que $(\mathbb{1}_A)_{e_1} = \mathbb{1}_{A_{e_1}}$. La transformation $f \to f_{e_1}$ (e_1 fixé) préserve les opérations habituelles sur les fonctions (additions, multiplication, ratio), y compris la convergence simple.

Un *pavé* de $\mathsf{E}_1 \times \mathsf{E}_2$ est un sous-ensemble de la forme

$$A_1 \times A_2 := \{(e_1, e_2) \in \mathsf{E}_1 \times \mathsf{E}_2 \,,\ e_1 \in A_1, e_2 \in A_2\} \ ;$$

un pavé est vide si et seulement si un de ses facteurs A_1 ou A_2 l'est. La section d'un pavé vaut : $(A_1 \times A_2)_{e_1} = A_2$ ou $\emptyset$ selon que $e_1 \in A_1$ ou que $e_1 \notin A_1$.

Définition A.20 (Pavé mesurable, Tribu produit, Produit d'espaces mesurables). *Soient* $(\mathsf{E}_1, \mathscr{E}_1)$ *et* $(\mathsf{E}_2, \mathscr{E}_2)$ *deux espaces mesurables.*

Un pavé mesurable $A_1 \times A_2$ *de* $(\mathsf{E}_1, \mathscr{A}_1) \times (\mathsf{E}_2, \mathscr{A}_2)$ *est un pavé de* $\mathsf{E}_1 \times \mathsf{E}_2$ *tel que* $A_1 \in \mathscr{E}_1$ *et* $A_2 \in \mathscr{E}_2$.

On appelle tribu produit *de* $\mathscr{E}_1$ *et de* $\mathscr{E}_2$*, la plus petite tribu sur* $\mathsf{E}_1 \times \mathsf{E}_2$ *qui contient les pavés mesurables de* $(\mathsf{E}_1, \mathscr{A}_1) \times (\mathsf{E}_2, \mathscr{A}_2)$*. Elle est notée* $\mathscr{E}_1 \otimes \mathscr{E}_2$.

L'espace mesurable $(\mathsf{E}_1 \times \mathsf{E}_2,\ \mathscr{A}_1 \otimes \mathscr{A}_2)$ *est appelé* le produit des espaces mesurables $(\mathsf{E}_1,\ \mathscr{A}_1)$ *et* $(\mathsf{E}_2, \mathscr{A}_2)$.

L'ensemble des pavés mesurables n'est pas une tribu, c'est la raison pour laquelle on a défini la plus petite tribu qui contient cette famille. On peut en effet facilement observer que la réunion de deux pavés n'est pas un pavé en considérant la réunion de $\{e_1\} \times E_2$ et $\{e_1'\} \times E_2$ pour $e_1 \neq e_1'$. L'ensemble des pavés mesurables a toutefois la structure d'une *semi-algèbre* i.e. $\emptyset$ et $\mathsf{E}_1 \times \mathsf{E}_2$ sont des pavés mesurables ; l'intersection de deux pavés mesurables est encore un pavé mesurable :

$$(A_1 \times A_2) \cap (A_1' \times A_2') = (A_1 \cap A_1') \times (A_2 \cap A_2');$$

et le complémentaire d'un pavé mesurable s'écrit comme la réunion finie de pavés mesurables disjoints :

$$(A_1 \times A_2)^{\mathrm{c}} = (A_1^c \times \mathsf{E}_2) \cup (\mathsf{E}_1 \times A_2^c) = (A_1^c \times A_2^c) \cup (A_1 \times A_2^c) \cup (A_1^c \times A_2)\,.$$

Le lemme suivant considère le cas de l'espace produit $\mathbb{R}^d$. Il établit que la tribu borélienne de $\mathbb{R}^d$ est aussi la tribu produit sur le produit cartésien $\mathbb{R}^d$.

Lemme A.21. $\mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R}) \otimes \ldots \otimes \mathscr{B}(\mathbb{R}) = \mathscr{B}(\mathbb{R}^d)$

Démonstration. On pose $\mathscr{E}^{\otimes d} := \mathscr{B}(\mathbb{R})\otimes\mathscr{B}(\mathbb{R})\otimes\ldots\otimes\mathscr{B}(\mathbb{R})$. Si U un ouvert de $\mathbb{R}^d$, $U=\cup_n P_n$ où $\{P_n,\, n\in\mathbb{N}\}$ est une famille de pavés ouverts de $\mathbb{R}^d$; pour construire cette famille, on peut considérer les pavés de la forme $\prod_{i=1}^d]r_i-1/n_i, r_i+1/n_i[$, l'union étant prise sur toutes les valeurs $(n_1,\cdots,n_d)\in\mathbb{N}^d$ et sur tous les centres $(r_1,\cdots,r_d)\in\mathbb{Q}^d$ tels que le pavé est inclu dans U. Donc $U\in\mathscr{E}^{\otimes d}$ et $\mathscr{B}(\mathbb{R}^d)\subset\mathscr{E}^{\otimes d}$.

Réciproquement, soient $X_1, X_2, \ldots, X_d$ les applications coordonnées de $\mathbb{R}^d$ sur $\mathbb{R}$. Les X_k sont des fonctions continues donc mesurables de $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ dans $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ d'où $\mathscr{E}^{\otimes d} = \sigma(X_1,\ldots,X_d)\subset\mathscr{B}(\mathbb{R}^d)$. □

Nous terminons cette section par des propriétés sur la mesurabilité des sections qui seront utiles notamment pour justifier le théorème de Fubini (voir Théorème A.48 et Théorème A.48).

Proposition A.22. *Soit e_1 fixé. La section A_{e_1} de toute partie mesurable A de $(\mathsf{E}_1\times\mathsf{E}_2, \mathscr{A}_1\otimes\mathscr{A}_2)$, est une partie mesurable de $(\mathsf{E}_2,\mathscr{A}_2)$. De même, la section f_{e_1} d'une fonction numérique mesurable f sur $(\mathsf{E}_1\times\mathsf{E}_2, \mathscr{A}_1\otimes\mathscr{A}_2)$ est une fonction numérique mesurable sur $(\mathsf{E}_2,\mathscr{A}_2)$.*

Démonstration. Soit $\mathscr{C}_{e_1}$ la classe des parties A de $\mathsf{E}_1\times\mathsf{E}_2$ telles que $A_{e_1}\in\mathscr{A}_2$. On voit facilement que tout pavé mesurable appartient à $\mathscr{C}_{e_1}$ et que $\mathscr{C}_{e_1}$ est stable pour les opérations de complémentation et d'intersection dénombrable. Il s'en suit que $\mathscr{C}_{e_1}$ contient $\mathscr{A}_1\otimes\mathscr{A}_2$, ce qui démontre la première partie de la proposition. La seconde partie est alors une simple conséquence de l'identité : $(f_{e_1})^{-1}(B) = [f^{-1}(B)]_{e_1}$.□

Corollaire A.23. *Pour que le pavé non vide $B_1\times B_2$ de $\mathsf{E}_1\times\mathsf{E}_2$ appartienne à $\mathscr{A}_1\otimes\mathscr{A}_2$, il faut et il suffit que $B_1\in\mathscr{A}_1$ et que $B_2\in\mathscr{A}_2$.*

Pour que la fonction réelle non identiquement nulle $f_1(e_1)f_2(e_2)$ définie sur $\mathsf{E}_1\times\mathsf{E}_2$ soit $\mathscr{E}_1\otimes\mathscr{E}_2$ mesurable, il faut et il suffit que f_1 (resp. f_2) soit $\mathscr{E}_1$ (resp. $\mathscr{E}_2$) mesurable. (Ce corollaire justifie la terminologie "pavé mesurable" utilisée ci-dessus).

Démonstration. En effet, si $B_1\times B_2$ est non vide, l'ensemble B_1 ne peut être vide. Par suite, si $e_1\in B_1$, $B_2 = (B_1\times B_2)_{e_1}$ appartient à $\mathscr{A}_2$ d'après la proposition précédente. De même : $B_1\in\mathscr{A}_1$. Un raisonnement analogue s'applique aux fonctions. □

A.2 Mesures

A.2.1 Définitions

Avant de définir ce qu'est une mesure, nous commençons par un lemme qui donne un sens à la somme d'une famille dénombrable de réels à valeur dans $\overline{\mathbb{R}^+}$: la somme est indépendante de l'ordre dans lequel les éléments sont sommés.

Lemme A.24 (Sommation par paquets). *Soit I un ensemble dénombrable, $\{\mathrm{P}_j,\ j\in J\}$ une partition de I où J est un ensemble au plus dénombrable, et $\{a_i,\ i\in I\}$ une famille d'éléments de $\overline{\mathbb{R}^+}$. Alors*

$$\sum_{i\in I} a_i = \sum_{j\in J}\left(\sum_{i\in\mathrm{P}_j} a_i\right).$$

Démonstration. Soit ϕ une énumération de I i.e. une bijection de $\mathbb{N}$ sur I. On pose $S_n^\phi := \sum_{k=0}^n a_{\phi(k)}$. La suite $\{S_n^\phi, n\geq 0\}$ est croissante et $S^\phi := \lim\uparrow S_n^\phi$ existe dans $\overline{\mathbb{R}^+}$. Si ψ est une autre énumération de I,

on a, pour n fixé et m assez grand, $\{a_{\phi(0)},\dots,a_{\phi(n)}\} \subset \{a_{\psi(0)},\dots,a_{\psi(m)}\}$, d'où $S_n^\phi \le S_m^\psi$; puis $S^\phi \le S^\psi$. Permutant ϕ et ψ, on a $S^\psi \le S^\phi$ et $S^\phi = S^\psi$. On pose donc $\sum_{i\in I} a_i := \lim\uparrow S_n^\phi$, quantité qui ne dépend pas de l'énumération ϕ. □

Définition A.25 (Mesure, Espace mesuré). *On appelle* mesure *sur l'espace mesurable* $(\mathsf{E},\mathscr{E})$ *toute application* μ *de* $\mathscr{E}$ *dans* $\overline{\mathbb{R}^+}$ *telle que*

(i) $\mu(\emptyset) = 0$,

(ii) (σ*-additivité) pour toute suite* $\{A_n,\ n\in\mathbb{N}\}$ *d'éléments de* $\mathscr{E}$ *deux à deux disjoints,*

$$\mu\left(\bigcup_{n=0}^{\infty} A_n\right) = \sum_{n=0}^{\infty} \mu(A_n)\ .$$

Le triplet $(\mathsf{E},\mathscr{E},\mu)$ *s'appelle un* espace mesuré.

Les propriétés suivantes découlent de façon élémentaire de la définition.

Proposition A.26. *Soit* $(\mathsf{E},\mathscr{E},\mu)$ *un espace mesuré.*

(i) Si $A,B\in\mathscr{E}$ *sont disjoints, alors* $\mu(A\cup B) = \mu(A)+\mu(B)$.

(ii) si $A,B\in\mathscr{E}$ *et* $A\subset B$, $\mu(A)\le\mu(B)$,

(iii) si $\{A_n,\ n\in\mathbb{N}\}$ *est une suite d'éléments de* $\mathscr{E}$, $\mu\left(\bigcup_{n=0}^{\infty}A_n\right) \le \sum_{n=0}^{\infty}\mu(A_n)$,

(iv) si $\{A_n,\ n\in\mathbb{N}\}$ *est une suite d'éléments de* $\mathscr{E}$ *telle que, pour tout* $n\in\mathbb{N}$, $A_n\subset A_{n+1}$ *alors* $\lim_{n\to\infty}\uparrow \mu(A_n) = \mu(A)$ *où* $A := \bigcup_{n=0}^{\infty}A_n$.

(v) si $\{A_n,\ n\in\mathbb{N}\}$ *est une suite d'éléments de* $\mathscr{E}$ *telle que, pour tout* $n\in\mathbb{N}$, $A_{n+1}\subset A_n$ *et si, pour un* n_0, $\mu(A_{n_0})<+\infty$, *alors* $\lim_{n\to\infty}\downarrow \mu(A_n) = \mu(A)$ *où* $A := \bigcap_{n=0}^{\infty}A_n$.

Démonstration. Pour la première propriété, on écrit $A\cup B$ comme l'union dénombrable $A\cup B\cup\emptyset\cup\emptyset\cup\cdots$ puis on applique la définition d'une mesure puisque par définition, $\emptyset$ est disjoint avec tout ensemble. Les autres relations s'établissent en écrivant que $A\cup B = A\cup(B\setminus A)$ lorsque $A\subset B$. □

Définition A.27 (Mesure σ-finie, Mesure bornée, Probabilité, Espace de probabilité).

1. *Une* mesure μ *sur* $(\mathsf{E},\mathscr{E})$ *est dite* σ-finie *s'il existe une suite* $\{E_n,\ n\in\mathbb{N}\}$ *d'éléments de* E *telle que* $\mathsf{E} = \bigcup_{n=1}^{\infty}\mathsf{E}_n$ *et* $\mu(\mathsf{E}_n)<+\infty$ *pour tout* $n\in\mathbb{N}$.
2. *La* mesure μ *est* bornée *si* $\mu(\mathsf{E})<+\infty$. *Si* $\mu(\mathsf{E})=1$, *la mesure* μ *est appelée une* probabilité. *Lorsque* μ *est une probabilité, le triplet* $(\mathsf{E},\mathscr{E},\mu)$ *est un* espace de probabilité.

La proposition suivante donne des conditions suffisantes pour établir que deux mesures sont égales. En particulier, lorsque μ et ν sont deux probabilités sur un même espace mesurable $(\mathsf{E},\mathscr{E})$, elles sont égales si et seulement si elles coïncident sur une classe stable par intersection finie engendrant $\mathscr{E}$.

Proposition A.28. *Soient* μ *et* ν *deux mesures sur* $(\mathsf{E},\mathscr{E})$ *et* $\mathscr{C}\subset\mathscr{E}$ *une classe d'ensembles stable par intersection finie. On suppose que, pour tout* $A\in\mathscr{C}$, $\mu(A)=\nu(A)<+\infty$ *et que* $\mathsf{E} = \lim\uparrow \mathsf{E}_n$ *avec* $\mathsf{E}_n\in\mathscr{C}$. *Alors* $\mu(A)=\nu(A)$ *pour tout* $A\in\sigma(\mathscr{C})$.

Démonstration. Observons tout d'abord que $\mu(E) = \nu(E)$ puisque

$$\mu(E) = \lim \uparrow \mu(E_n) = \lim \uparrow \nu(E_n) = \nu(E).$$

Supposons d'abord $\mu(\mathsf{E}) = \nu(\mathsf{E}) < +\infty$. Soit $\mathscr{M} := \{A \in \mathscr{E} \,:\, \mu(A) = \nu(A)\}$. On vérifie facilement que $\mathscr{M}$ est une classe monotone qui contient $\mathscr{C}$. Par le Théorème A.6, on a $\sigma(\mathscr{C}) \subset \mathscr{M}$. Le cas $\mu(E) = \nu(E) = +\infty$ se traite en appliquant ce résultat aux mesures $\mu_n(A) := \mu(A \cap \mathsf{E}_n)$ et $\nu_n(A) := \nu(A \cap \mathsf{E}_n)$, puis en utilisant la propriété $\mu(A) = \lim_n \uparrow \mu_n(A)$. □

A.2.2 Ensembles négligeables

Définition A.29 (Ensemble négligeable, presque-partout, presque-sûrement). *Soit $(\mathsf{E}, \mathscr{E}, \mu)$ un espace mesuré. Un sous-ensemble A de E est dit* négligeable *(ou μ-négligeable s'il y a ambiguïté) si $A \subset B$ avec $B \in \mathscr{E}$ et $\mu(B) = 0$.*

Une propriété est vraie presque-partout *(en abrégé p.p.) si elle est vraie en dehors d'un ensemble négligeable. Si μ est une probabilité, on dit aussi* presque-sûrement *(en abrégé p.s.) au lieu de presque-partout.*

Par exemple "$f = g$ p.p." signifie que $\{x \in \mathsf{E},\ f(x) \neq g(x)\}$ est négligeable.

Définition A.30 (Espace mesuré complet). *L'*espace mesuré $(\mathsf{E}, \mathscr{E}, \mu)$ *est dit* complet *si $\mathscr{E}$ contient la classe des ensembles négligeables $\mathscr{N}$.*

On peut toujours "compléter" un espace mesuré. Pour ce faire, on définit $\overline{\mathscr{E}} = \sigma(\mathscr{E}, \mathscr{N})$. Alors $A \in \overline{\mathscr{E}}$ si et seulement si $A = B \cup N$ avec $B \in \mathscr{E}$ et $N \in \mathscr{N}$. On peut prolonger μ à $\overline{\mathscr{E}}$ en posant $\mu(A) = \mu(B)$ (il est facile de voir que ceci ne dépend pas de l'écriture de A). L'espace $(\mathsf{E}, \overline{\mathscr{E}}, \mu)$ est alors complet et s'appelle le complété de $(\mathsf{E}, \mathscr{E}, \mu)$. Enfin on vérifie aisément que $f : \mathsf{E} \to \overline{\mathbb{R}}$ est $\overline{\mathscr{E}}$-mesurable si et seulement si il existe $g, h : \mathsf{E} \to \overline{\mathbb{R}}$ $\mathscr{E}-$mesurables telles que $g \leq f \leq h$ et $g = h$ p.p.

A.2.3 Construction d'un espace mesuré

Dans la suite, la plupart du temps, on se donnera un espace mesuré ou un espace de probabilité sans se soucier de sa construction. Il est néanmoins indispensable de s'assurer de l'existence de tels objets.

Nous commençons par discuter de la construction d'une mesure sur l'espace mesurable $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Observons d'abord que $\mathscr{C} := \{]a,b],\ -\infty < a \leq b < +\infty\}$ est une classe stable par intersection finie et que $\sigma(\mathscr{C}) = \mathscr{B}(\mathbb{R})$ (voir Appendice A.1). Il résulte alors de la Proposition A.28 qu'une mesure μ sur $\mathscr{B}(\mathbb{R})$ finie sur les intervalles bornés est déterminée par les valeurs $\mu(]a,b])$. Comment peut-on spécifier la mesure de tous ces intervalles ?

Etant donnée une mesure sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, on définit la fonction F par

$$\begin{aligned} &F(0) = 0, \\ &F(x) = \mu(]0,x]), \qquad x > 0, \\ &F(x) = -\mu(]x,0]), \qquad x < 0. \end{aligned}$$

Alors F est une fonction continue à droite et croissante et l'on a $\mu(]a,b]) = F(b) - F(a)$.

Réciproquement, soit F une application de $\mathbb{R}$ dans $\mathbb{R}$ continue à droite et croissante, existe-t-il une mesure μ sur $\mathscr{B}(\mathbb{R})$ telle que $\mu(]a,b]) = F(b) - F(a)$? Il est facile de décrire l'algèbre $\mathscr{A}$ engendrée par $\mathscr{C}$. On a

$$\mathscr{A} = \{A = \cup_{k=1}^{n}]a_k, b_k],\ -\infty \leq a_1 < b_1 < a_2 < \ldots < b_{n-1} < a_n < b_n \leq +\infty\}$$

en convenant que, si $b_n = +\infty$, $]a_n, b_n] =]a_n, +\infty[$. On définit μ sur $\mathscr{A}$ par $\mu(A) = \sum_{k=1}^{n} F(b_k) - F(a_k)$ où $F(+\infty) = \lim_{x\to+\infty} F(x)$, $F(-\infty) = \lim_{x\to-\infty} F(x)$. Il est facile de montrer que μ est additive sur $\mathscr{A}$, un peu plus délicat de montrer que μ est σ-additive sur $\mathscr{A}$ mais cela se fait. On a donc construit une mesure μ sur $\mathscr{A}$ telle que $\mu(]a,b]) = F(b) - F(a)$. Pour passer à $\mathscr{B}(\mathbb{R})$, on utilise le théorème de Carathéodory :

Théorème A.31 (Théorème de Carathéodory). *Soit μ une mesure sur une algèbre $\mathscr{A}$, alors μ se prolonge en une mesure sur $\sigma(\mathscr{A})$. De plus, si μ est σ-finie, ce prolongement est unique.*

► **Mesure de Lebesgue sur $\mathbb{R}$**

Si on choisit $F(x) = x$, on obtient l'existence et l'unicité d'une mesure λ_{Leb} sur $\mathscr{B}(\mathbb{R})$ vérifiant, pour tout intervalle I, $\lambda_{\text{Leb}}(I) = |I|$. C'est la *mesure de Lebesgue* sur $\mathbb{R}$. Si $\mathscr{N}$ est la classe des ensembles λ_{Leb}-négligeables, $\overline{\mathscr{B}(\mathbb{R})} = \sigma(\mathscr{E}, \mathscr{N})$ s'appelle la tribu des ensembles Lebesgue-mesurables (elle est beaucoup plus "grosse" que $\mathscr{B}(\mathbb{R})$) et λ_{Leb} se prolonge sans peine à $\overline{\mathscr{B}(\mathbb{R})}$.

► **Mesure de Lebesgue sur $\mathbb{R}^d$**

Soit λ_{Leb} la mesure de Lebesgue sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. On définit alors, sur $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, $\lambda_{\text{Leb}}^{\otimes d} = \lambda_{\text{Leb}} \otimes \lambda_{\text{Leb}} \otimes \ldots \otimes \lambda_{\text{Leb}}$. On peut appliquer la proposition A.28 à

$$\mathscr{C} := \left\{ A, A = \prod_{i=1}^{d}]a_i, b_i[\,:\, -\infty < a_i < b_i < +\infty \right\}.$$

On obtient que $\lambda_{\text{Leb}}^{\otimes d}$ est l'unique mesure sur $\mathscr{B}(\mathbb{R}^d)$ telle que, pour tous $-\infty < a_i < b_i < +\infty$,

$$\lambda_{\text{Leb}}^{\otimes d}\left(\prod_{i=1}^{d}]a_i, b_i[\right) = \prod_{i=1}^{d} (b_i - a_i).$$

On appelle $\lambda_{\text{Leb}}^{\otimes d}$ la *mesure de Lebesgue sur* $\mathbb{R}^d$.

A.3 Intégration

Soit $(\mathsf{E}, \mathscr{E}, \mu)$ un espace mesuré. L'objectif de cette section est de construire l'intégrale d'une fonction mesurable $f : \mathsf{E} \to \mathbb{R}$ par rapport à la mesure μ. Cette intégrale sera notée $\mu(f)$ ou $\int f \, \mathrm{d}\mu$.

Nous commençons par la définir pour des fonctions positives étagées, puis pour des fonctions positives et enfin des fonctions signées.

A.3.1 Intégration des fonctions positives

Si f est une fonction positive étagée, elle s'écrit $f = \sum_{k=1}^{p} a_k \mathbb{1}_{A_k}$ où $A_k \in \mathscr{E}$. On pose

$$\int f \, \mathrm{d}\mu := \sum_{k=1}^{p} a_k \mu(A_k).$$

La première propriété à vérifier est que $\int f \, \mathrm{d}\mu$ est bien définie, i.e. que cette quantité ne dépend pas du choix de la représentation de f. La preuve (élémentaire) est laissée au lecteur.

On peut aussi montrer (toujours en utilisant des arguments élémentaires) que

Proposition A.32. *Soient $f, g \in e\mathscr{E}^+$.*

(i) Si $\mu\{f \neq g\} = 0$, alors $\int f \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu$.

(ii) Pour $a, b \in \mathbb{R}^+$, on a $\int (af + bg) \,\mathrm{d}\mu = a \int f \,\mathrm{d}\mu + b \int g \,\mathrm{d}\mu$.

(iii) Si $f \leq g$, alors $\int f \,\mathrm{d}\mu \leq \int g \,\mathrm{d}\mu$.

On a aussi le résultat plus technique suivant qui est la clé de la construction :

Lemme A.33. *Si $\{f_n,\, n \in \mathbb{N}\}, \{g_n,\, n \in \mathbb{N}\}$ sont des suites croissantes d'éléments de $e\mathscr{E}^+$ et si $\lim\uparrow f_n = \lim\uparrow g_n$, alors on a $\lim\uparrow \int f_n \,\mathrm{d}\mu = \lim\uparrow \int g_n \,\mathrm{d}\mu$.*

Démonstration. Il suffit de montrer que

$$\lim\uparrow f_n \geq g = \sum_{k=1}^{p} \alpha_k \mathbb{1}_{A_k} \in e\mathscr{E}^+ \quad \Longrightarrow \quad \lim\uparrow \int f_n \,\mathrm{d}\mu \geq \int g \,\mathrm{d}\mu \ .$$

Soient $c \in]0,1[$ et $\mathsf{E}_n := \{f_n \geq cg\}$. On a $\mathsf{E}_n \in \mathscr{E}$, $\mathsf{E}_n \uparrow \mathsf{E}$ et $f_n \geq cg\mathbb{1}_{\mathsf{E}_n}$ d'où

$$\int f_n \,\mathrm{d}\mu \geq c \int g \mathbb{1}_{\mathsf{E}_n} \,\mathrm{d}\mu = c \sum_{k=1}^{p} \alpha_k \mu(A_k \cap \mathsf{E}_n).$$

On obtient, lorsque $n \to +\infty$,

$$\lim\uparrow \int f_n \,\mathrm{d}\mu \geq c \sum_{k=1}^{p} \alpha_k \mu(A_k) = c \int g \,\mathrm{d}\mu.$$

Puisque $c < 1$ est arbitraire, cela entraine $\lim\uparrow \int f_n \,\mathrm{d}\mu \geq g$ et conclut la preuve. □

On peut maintenant définir l'intégrale d'une fonction $f \in \mathscr{E}^+$. Par la Proposition A.12, il existe une suite $\{f_n,\, n \in \mathbb{N}\}$ de fonctions de $e\mathscr{E}^+$ telles que $f = \lim\uparrow f_n$. On a alors $\int f_n \,\mathrm{d}\mu \uparrow$ et on pose

$$\int f \,\mathrm{d}\mu := \lim\uparrow \int f_n \,\mathrm{d}\mu.$$

Le point important est que, d'après le Lemme A.33, cette limite ne dépend pas de la suite $\{f_n,\, n \in \mathbb{N}\}$ choisie. On dira que $f \in \mathscr{E}^+$ est *intégrable* si $\int f \,\mathrm{d}\mu < +\infty$.

On a les propriétés suivantes

Proposition A.34. *Soient $f, g \in \mathscr{E}^+$.*

(i) Pour $a, b \in \mathbb{R}^+$, on a $\int (af + bg) \,\mathrm{d}\mu = a \int f \,\mathrm{d}\mu + b \int g \,\mathrm{d}\mu$.

(ii) Si $f \leq g$, alors $\int f \,\mathrm{d}\mu \leq \int g \,\mathrm{d}\mu$.

(iii) Si $\int f \,\mathrm{d}\mu < +\infty$, alors $f < +\infty$ p.p.

(iv) Si $\int f \,\mathrm{d}\mu = 0$, alors $f = 0$ p.p.

Démonstration. Pour établir les deux premières propriétés, on utilise qu'une fonction positive mesurable est la limite croissante d'une suite de fonctions positives étagées. Pour la troisième propriété, il suffit de remarquer que $f = f\mathbb{1}_{f=+\infty} + f\mathbb{1}_{f<\infty}$ et que chaque terme de la somme est un élément de $\mathscr{E}^+$; il vient que $f \geq f\mathbb{1}_{f=+\infty}$ dont on déduit par la relation ((ii)), que si l'intégrale de f est finie, c'est que l'ensemble $\{f = +\infty\}$ est de mesure nulle pour la mesure μ. Enfin, on procède de même pour établir la dernière relation en écrivant $f = f\mathbb{1}_{f=0} + f\mathbb{1}_{f>0} \geq f\mathbb{1}_{f>0}$. □

A.3.2 Intégration des fonctions réelles ou complexes

On pose

$$\mathrm{L}^1(\mathsf{E},\mathscr{E},\mu) := \left\{ f \in [\mathscr{E}],\ \int |f|\,\mathrm{d}\mu < +\infty \right\}. \tag{A.4}$$

Quand il n'y a pas d'ambiguïté sur l'espace sur lequel on intègre, on posera $\mathrm{L}^1(\mu) := \mathrm{L}^1(\mathsf{E},\mathscr{E},\mu)$.

Si $f \in \mathrm{L}^1(\mu)$, $f^+ := \max(f,0)$ et $f^- := \max(-f,0)$ sont intégrables et on pose

$$\int f\,\mathrm{d}\mu = \int f^+\,\mathrm{d}\mu - \int f^-\,\mathrm{d}\mu.$$

Il est facile de voir (vu que $|f+g| \le |f|+|g|$) que $\mathrm{L}^1(\mu)$ est un espace vectoriel et que $f \to \int f\,\mathrm{d}\mu$ est une forme linéaire positive sur $\mathrm{L}^1(\mu)$.

On a aussi les propriétés :

Proposition A.35. *Soient $f,g \in \mathrm{L}^1(\mu)$.*

(i) Si $f \le g$ p.p., alors $\int f\,\mathrm{d}\mu \le \int g\,\mathrm{d}\mu$.

(ii) $|\int f\,\mathrm{d}\mu| \le \int |f|\,\mathrm{d}\mu$.

(iii) Si $A \in \mathscr{E}$, alors $f\mathbb{1}_A \in \mathrm{L}^1(\mu)$. On pose $\int_A f\,\mathrm{d}\mu := \int f\mathbb{1}_A\,\mathrm{d}\mu$.

(iv) Si pour tout $A \in \mathscr{E}$, $\int_A f\,\mathrm{d}\mu \ge 0$ alors $f \ge 0$ p.p.

(v) Si, pour tout $A \in \mathscr{E}$, $\int_A f\,\mathrm{d}\mu \le \int_A g\,\mathrm{d}\mu$, alors $f \le g$ p.p.

Démonstration. La première relation est une conséquence de la même propriété pour les fonction positives, appliquée à $g-f$. Pour la seconde, on revient à la définition de l'intégrale d'une fonction signée et on utilise le fait que l'intégrale de la somme de fonctions positives mesurables est la somme des intégrales :

$$|\int f\,\mathrm{d}\mu| = \left|\int f^+\mathrm{d}\mu - \int f^-\mathrm{d}\mu\right| \le \int f^+\mathrm{d}\mu + \int f^-\mathrm{d}\mu = \int (f^+ + f^-)\mathrm{d}\mu = \int |f|\mathrm{d}\mu.$$

La troisième propriété découle de la majoration $|f|\mathbb{1}_A \le |f|$. L'avant-dernière se prouve en prenant $A = \{f < 0\}$ qui est mesurable puisque f l'est. Enfin, la dernière se démontre en combinant la positivité de l'intégrale et l'hypothèse avec $A = \{g - f < 0\}$. □

Si f est $\mathscr{E}$-mesurable à valeurs $\mathbb{C}$, on pose ($|f|$ désignant le module),

$$\mathrm{L}^1_{\mathbb{C}} = \mathrm{L}^1_{\mathbb{C}}(\mathsf{E},\mathscr{E},\mu) := \left\{ f\ \mathscr{E}\text{-mesurable complexe},\ \int |f|\,\mathrm{d}\mu < +\infty \right\}. \tag{A.5}$$

On définit alors, pour $f \in \mathrm{L}^1_{\mathbb{C}}$,

$$\int f\,\mathrm{d}\mu := \int \mathrm{Re}(f)\,\mathrm{d}\mu + i\int \mathrm{Im}(f)\,\mathrm{d}\mu.$$

$\mathrm{L}^1_{\mathbb{C}}$ est un espace vectoriel sur $\mathbb{C}$ et $f \to \int f\,\mathrm{d}\mu$ une forme linéaire sur $\mathrm{L}^1_{\mathbb{C}}$. On a aussi

Proposition A.36. *Pour toute fonction $f \in \mathrm{L}^1_{\mathbb{C}}$, $|\int f\,\mathrm{d}\mu| \le \int |f|\,\mathrm{d}\mu$.*

Démonstration. On a $\int f\,\mathrm{d}\mu = r\mathrm{e}^{\mathrm{i}\theta}$ et

$$\left|\int f\,\mathrm{d}\mu\right| = r = \mathrm{Re}\left(\mathrm{e}^{-\mathrm{i}\theta}\int f\,\mathrm{d}\mu\right) = \mathrm{Re}\left(\int(\mathrm{e}^{-\mathrm{i}\theta}f)\,\mathrm{d}\mu\right) = \int \mathrm{Re}(\mathrm{e}^{-\mathrm{i}\theta}f)\,\mathrm{d}\mu$$
$$\leq \int |\mathrm{e}^{-\mathrm{i}\theta}f|\,\mathrm{d}\mu = \int |f|\,\mathrm{d}\mu.$$

L'inégalité est une conséquence de la positivité de l'intégrale pour des fonctions de $\mathrm{L}^1(\mu)$ (voir Proposition A.35). □

A.3.3 Permutation limite et intégrale

Il nous reste à énoncer les résultats concernant les passages à la limite. Le premier d'où découlent facilement les autres s'appelle théorème de convergence monotone ou théorème de Beppo-Levi.

Théorème A.37 (Convergence monotone ou Beppo-Levi). *Soit* $(\mathsf{E},\mathscr{E},\mu)$ *un espace mesuré et* $\{f_n,\ n\in\mathbb{N}\}$ *une suite croissante de fonctions de* $\mathscr{E}^+$. *Alors*

$$\lim\uparrow \int f_n\,\mathrm{d}\mu = \int \lim\uparrow f_n\,\mathrm{d}\mu.$$

Démonstration. Pour tout n, il existe une famille $\{f_{n,k}, k\in\mathbb{N}\}$ de fonction de $e\mathscr{E}^+$ telles que $f_n = \lim\uparrow_k f_{n,k}$. On pose $g_k := \max_{n\leq k} f_{n,k}$. On a $g_k \in e\mathscr{E}^+$, la suite $\{g_k,\ k\in\mathbb{N}\}$ est croissante et pour $n\leq k$,

$$f_{n,k}\leq g_k\leq f_n, \qquad \int f_{n,k}\,\mathrm{d}\mu \leq \int g_k\,\mathrm{d}\mu \leq \int f_n\,\mathrm{d}\mu.$$

On pose $f := \lim\uparrow f_n$. On a, pour $k\to+\infty$,

$$f_n \leq \lim\uparrow g_k \leq f_n, \qquad \int f_n\,\mathrm{d}\mu \leq \lim\uparrow \int g_k\,\mathrm{d}\mu = \int \lim\uparrow g_k\,\mathrm{d}\mu \leq \int f_n\,\mathrm{d}\mu,$$

puis pour $n\to+\infty$,

$$f\leq \lim\uparrow g_n \leq f, \qquad \lim\uparrow \int f_n\,\mathrm{d}\mu \leq \int \lim\uparrow g_k\,\mathrm{d}\mu \leq \lim\uparrow \int f_n\,\mathrm{d}\mu.$$ □

On en déduit $f = \lim\uparrow g_n$ et $\int f\,\mathrm{d}\mu = \lim\uparrow \int f_n\,\mathrm{d}\mu$.

En appliquant le théorème précédent avec $f_n \leftarrow \sum_{k=0}^n g_k$, on obtient le corollaire suivant :

Corollaire A.38. *Soit* $(\mathsf{E},\mathscr{E},\mu)$ *un espace mesuré et* $\{g_n,\ n\in\mathbb{N}\}$ *une suite d'éléments de* $\mathscr{E}^+$. *Nous avons*

$$\sum_n \int g_n\,\mathrm{d}\mu = \int \sum_n g_n\,\mathrm{d}\mu.$$

Le second résultat de passage à la limite est valable pour toute suite de fonctions mesurables positives ; il ne requiert pas d'hypothèses de monotonie, mais donne un résultat plus faible (une inégalité plutôt qu'une égalité de permutation limite/intégrale).

Proposition A.39. *(Lemme de Fatou) Soit* $\{f_n,\, n\in\mathbb{N}\}$ *une suite d'éléments de* $\mathscr{E}^+$. *Alors*

$$\int \liminf f_n \,\mathrm{d}\mu \le \liminf \int f_n \,\mathrm{d}\mu.$$

Démonstration. On a $\liminf f_n = \lim_n \uparrow \inf_{k\ge n} f_k$, qui est la limite d'une suite croissante de fonctions de $\mathscr{E}^+$. Par application du Théorème A.37 et en utilisant l'inégalité $\inf_{k\ge n} f_k \le f_k$ pour tout $k\ge n$, on a

$$\int \liminf f_n \,\mathrm{d}\mu = \lim_n \uparrow \int \inf_{k\ge n} f_k \,\mathrm{d}\mu \le \lim_n \uparrow \inf_{k\ge n} \int f_k \,\mathrm{d}\mu = \liminf \int f_n \,\mathrm{d}\mu. \qquad \square$$

Théorème A.40 (Théorème de convergence dominée ou de Lebesgue). *Soit* $(\mathsf{E},\mathscr{E},\mu)$ *un espace mesuré et* $\{f_n,\, n\in\mathbb{N}\}$ *une suite d'éléments de* $\mathrm{L}^1_{\mathbb{C}}$ *telle que*

(i) il existe une fonction $f\in[\mathscr{E}]$ *telle que* $\lim_{n\to\infty} f_n = f$ *p.p.*

(ii) il existe une fonction positive $g\in \mathrm{L}^1(\mu)$ *telle que* $|f_n|\le g$ *p.p.*

Alors

$$\lim \int f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu.$$

Démonstration. Remarquons tout d'abord qu'il suffit de considérer la cas réel ; ainsi, dans la suite, $f_n \in \mathrm{L}^1(\mu)$ pour tout n. Appliquant la Proposition A.39 aux fonctions positives $g+f_n$ et $g-f_n$, puis en retranchant $\int g\mathrm{d}\mu$ qui est finie par hypothèse, on a

$$\int \liminf f_n \,\mathrm{d}\mu \le \liminf \int f_n \,\mathrm{d}\mu \le \limsup \int f_n \,\mathrm{d}\mu \le \int \limsup f_n \,\mathrm{d}\mu.$$

On termine la démonstration en utilisant l'hypothèse, $\liminf f_n = \limsup f_n = f$. $\square$

Ce théorème a une version "continu" très utile, qui permet de permuter intégrale et limite lorsque la famille de fonctions est indexée par t, pour t à valeur dans un ouvert de $\mathbb{R}^d$. La démonstration du corollaire suivant repose sur l'application de Théorème A.40 en se souvenant que $\lim_{t\to t_0}\int f_t\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$ si et seulement si pour toute suite $\{t_n,\, n\in\mathbb{N}\}$ tendant vers t_0, $\lim_{t_n\to t_0}\int f_{t_n}\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$.

Corollaire A.41. *Soit* $(\mathsf{E},\mathscr{E},\mu)$ *un espace mesuré, U un ouvert de* $\mathbb{R}^d$ *et* $\{f_t,\, t\in U\}$ *une famille d'éléments de* $\mathrm{L}^1_{\mathbb{C}}$. *On suppose que* $\lim_{t\to t_0} f_t = f$ *p.p. et qu'il existe une fonction* $g\in\mathrm{L}^1(\mu)$ *telle que pour tout* $t\in U$, $|f_t|\le g$ *p.p. Alors* $\lim_{t\to t_0}\int f_t\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$.

Donnons un exemple d'utilisation de ce corollaire.

Proposition A.42. *Soient* $(\mathsf{E},\mathscr{E},\mu)$ *un espace mesuré, I un intervalle ouvert et* $\{f(t,\cdot),\, t\in I\}$ *une famille d'éléments de* $\mathrm{L}^1_{\mathbb{C}}(\mu)$. *On pose, pour tout* $t\in I$,

$$\phi(t) := \int f(t,x)\,\mu(\mathrm{d}x)\,.$$

On suppose qu'il existe $A\in\mathscr{E}$ *et une fonction* $g\in\mathrm{L}^1(\mu)$ *tels que*

(i) $\mu(A^c) = 0$,

(ii) pour tout $x \in A$, $t \mapsto f(t,x)$ *est dérivable sur* I,

(iii) pour tout $x \in A$ *et* $t \in I$, $|\frac{\partial f}{\partial t}(t,x)| \leq g(x)$.

Alors ϕ *est dérivable sur* I *et*

$$\phi'(t) = \int \frac{\partial f}{\partial t}(t,x)\,\mu(\mathrm{d}x)\ .$$

Démonstration. Soit $t \in I$. Puisque $\mu(A^c) = 0$, pn a pour tout $h \in \mathbb{R}$,

$$h^{-1}\{\phi(t+h) - \phi(t)\} = \int_A h^{-1}\left(f(t+h,x) - f(t,x)\right)\mu(\mathrm{d}x).$$

D'après la formule des accroissements finis, on a, pour $x \in A$, et h suffisamment petit (tel que $t+h \in I$)

$$|h^{-1}(f(t+h,x) - f(t,x))| = \left|\frac{\partial f}{\partial t}(s,x)\right|$$

pour $s \in [0,t]$. De plus, par dérivabilité de $t \mapsto f(t,x)$ pour tout $x \in A$, il vient

$$\lim_{h\to 0} h^{-1}(f(t+h,x) - f(t,x)) = \frac{\partial f}{\partial t}(t,x)\ .$$

On peut appliquer le Corollaire A.41 et encore une fois $\mu(A^c) = 0$ pour obtenir

$$\lim_{h\to 0}\int_A h^{-1}(f(t+h,x) - f(t,x))\,\mu(\mathrm{d}x) = \int_A \frac{\partial f}{\partial t}(t,x)\,\mu(\mathrm{d}x) = \int \frac{\partial f}{\partial t}(t,x)\,\mu(\mathrm{d}x)\ .$$ □

A.3.4 Exemples

► Mesure de Lebesgue - Liens avec l'intégrale de Riemann

Dans cette section, μ désigne la mesure de Lebesgue sur $\mathbb{R}$. Soit f une fonction réelle continue sur $[a,b]$ et posons, pour $a \leq x \leq b$,

$$F(x) := \int_a^x f(t)\,\mathrm{d}t, \qquad G(x) := \int \mathbb{1}_{[a,x[} f\,\mathrm{d}\mu;$$

F est l'intégrale au sens de Riemann. On sait que $F(a) = 0$, F est continue sur $[a,b]$ et que, sur $]a,b[$, F est dérivable avec $F' = f$. Il est facile de vérifier que G a les mêmes propriétés. Ceci implique que $F = G$ sur $[a,b]$ et, en particulier, que

$$\int_a^b f(t)\,\mathrm{d}t = \int \mathbb{1}_{[a,b[} f\,\mathrm{d}\mu.$$

Par additivité, cette formule est encore vraie si f est continue par morceaux sur $[a,b]$.

Considérons maintenant une application f de $\mathbb{R}$ dans $\mathbb{R}$ continue par morceaux telle que $\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t$ soit absolument convergente. Lorsque $a \downarrow -\infty$ et $b \uparrow +\infty$,

— d'une part, par définition, $\int_a^b |f(t)|\,\mathrm{d}t \to \int_{-\infty}^{+\infty} |f(t)|\,\mathrm{d}t < +\infty$ et $\int_a^b f(t)\,\mathrm{d}t \to \int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t$;

— d'autre part, $\int \mathbb{1}_{[a,b[}|f|\,\mathrm{d}\mu \to \int |f|\,\mathrm{d}\mu$ (par application du théorème de convergence monotone) ce qui implique que $f \in \mathrm{L}^1(\mu)$ puis $\int \mathbb{1}_{[a,b[} f\,\mathrm{d}\mu \to \int f\,\mathrm{d}\mu$ (théorème de Lebesgue) puisque $|\mathbb{1}_{[a,b[} f| \leq |f| \in \mathrm{L}^1(\mu)$.

Donc

$$\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t = \int f\,\mathrm{d}\mu.$$

L'intégrale de Lebesgue permet d'intégrer beaucoup de fonctions qui ne sont pas intégrables au sens de Riemann (car pas assez régulières) : par exemple, la fonction $\mathbb{1}_{\mathbb{Q}}$ n'est pas Riemann-intégrable (toute fonction en escaliers la majorant resp. la minorant est supérieure ou égale à 1, resp. inférieure ou égale à 0), mais elle est intégrable au sens de Lebesgue et d'intégrale nulle.

Un autre avantage de l'intégrale de Lebesgue est qu'elle est valable dans des espaces mesurables quelconques (ce qui permet de donner un fondement rigoureux à la théorie des probabilités), tandis que la construction de Riemann se limite à l'intégration de fonctions $f : \mathbb{R}^k \to \mathbb{R}$.

Enfin, les preuves des théorèmes fondamentaux (convergence monotone, convergence dominée) sont nettement plus simples et transparentes dans le formalisme de Lebesgue. La raison principale est qu'une limite croissante de fonctions intégrables au sens de Lebesgue est intégrable, alors qu'une limite croissante de fonctions intégrables au sens de Riemann n'est pas nécessairement intégable au sens de Riemann.

Par contre, si $\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t$ est convergente mais pas absolument convergente (par exemple $f(x) = \sin x/x$), alors $f \notin \mathrm{L}^1(\mu)$.

► Mesure de comptage

Soient E un ensemble dénombrable et $\mu : \mathsf{E} \to \overline{\mathbb{R}^+}$. On pose pour tout $A \in \mathsf{E}$,

$$\mu(A) := \sum_{x\in A} \mu(x).$$

Le Lemme A.24 implique que μ est une mesure sur l'espace mesurable $(\mathsf{E}, \mathscr{P}(\mathsf{E}))$. On a alors

$$\mathrm{L}^1(\mu) = \left\{ f\,,\ \sum_{x\in\mathsf{E}} |f(x)|\mu(x) < +\infty \right\}$$

et, pour $f \in \mathrm{L}^1(\mu)$,

$$\int f\,\mathrm{d}\mu = \sum_{x\in\mathsf{E}} f(x)\mu(x)\,.$$

En particulier si on prend pour μ la mesure de comptage i.e. $\mu(x) = 1$ pour tout $x \in \mathsf{E}$, on a

$$\mathrm{L}^1(\mu) = \left\{ f\,,\ \sum_{x\in\mathsf{E}} |f(x)| < +\infty \right\} \quad \text{et} \quad \int f\,\mathrm{d}\mu = \sum_{x\in\mathsf{E}} f(x)\,.$$

Il est intéressant d'énoncer dans ce cadre les théorèmes de convergence/permutations limite-intégrale, vus précédemment. On a

(i) (Beppo-Levi) Si $0 \le f_n \uparrow f$, alors $\sum_x f_n(x) \uparrow \sum_x f(x)$.

(ii) (Fatou) Si $0 \le f_n$, alors $\sum_x \liminf_n f_n(x) \le \liminf_n \sum_x f_n(x)$.

(iii) (Convergence dominée) Si $f_n \to f$ et si $|f_n| \le g$ avec $\sum_x g(x) < +\infty$, alors $\lim_n \sum_x f_n(x) = \sum_x f(x)$.

► Mesures images

A toute mesure μ sur un espace mesurable $(\mathsf{E}, \mathscr{E})$, on peut associer une application I de $\mathscr{E}^+$ dans $\overline{\mathbb{R}^+}$ en posant

$$I(f) := \int f\,\mathrm{d}\mu, \qquad \forall f \in \mathscr{E}^+.$$

L'application I a alors les propriétés suivantes :

$$\begin{aligned} &I(f+g) = I(f)+I(g),\\ &I(af) = aI(f), \qquad a \in \mathbb{R}^+,\\ &I(f_n) \uparrow I(f) \ \text{ si } f_n \uparrow f. \end{aligned}$$

Réciproquement on a,

Proposition A.43. *Soient $(\mathsf{E}, \mathscr{E})$ un espace mesurable et I une application de $\mathscr{E}^+$ dans $\overline{\mathbb{R}^+}$ telle que*

(i) si $f, g \in \mathscr{E}^+$, $I(f+g) = I(f)+I(g)$;

(ii) si $f \in \mathscr{E}^+$ et $a \in \mathbb{R}^+$, $I(af) = aI(f)$,

(iii) si $f_n \in \mathscr{E}^+$ et si $f_n \uparrow f$, $I(f_n) \uparrow I(f)$.

Alors l'application $\mu : \mathscr{E} \to \overline{\mathbb{R}^+}$ définie par

$$\mu(A) := I(\mathbb{1}_A)$$

est une mesure sur $\mathscr{E}$ et on a, pour toute fonction $f \in \mathscr{E}^+$, $I(f) = \int f \, \mathrm{d}\mu$.

Démonstration. Soient $\{A_n,\ n \in \mathbb{N}\}$ une suite d'éléments de $\mathscr{E}$ deux à deux disjoints, d'union notée A. D'une part, comme les ensembles sont disjoints, $\mathbb{1}_A = \sum_n \mathbb{1}_{A_n}$ et d'autre part, $\sum_n \mathbb{1}_{A_n} = \lim\uparrow \sum_{k=1}^n \mathbb{1}_{A_k}$. Ainsi, par les propriétés (i) et (iii) de l'application I,

$$\mu(A) = I(\mathbb{1}_A) = I\left(\lim\uparrow \sum_{k=1}^n \mathbb{1}_{A_k}\right) = \lim\uparrow I\left(\sum_{k=1}^n \mathbb{1}_{A_k}\right) = \lim\uparrow \sum_{k=1}^n I(\mathbb{1}_{A_k}) = \sum_n \mu(A_n)\,.$$

Enfin, par (ii) appliquée avec $a = 0$, on a $\mu(\emptyset) = 0$. Ce qui conclut la preuve que μ est une mesure. On a alors, pour toute $f \in e\mathscr{E}^+$, $I(f) = \int f \, \mathrm{d}\mu$. On conclut facilement en utilisant la Proposition A.12. □

Théorème A.44. *Soient h une application mesurable de $(\mathsf{E}, \mathscr{E})$ dans $(\mathsf{Y}, \mathscr{Y})$ et μ une mesure sur $(\mathsf{E}, \mathscr{E})$. La formule*

$$\nu(A) = \mu(h^{-1}(A))\,, \quad A \in \mathscr{Y}\,,$$

définit une mesure sur $(\mathsf{Y}, \mathscr{Y})$ appelée mesure image de μ par h et notée $\mu \circ h^{-1}$ ou μ^h. On a, pour toute $f \in \mathscr{Y}^+$,

$$\int f \, \mathrm{d}\mu \circ h^{-1} = \int f \circ h \, \mathrm{d}\mu. \tag{A.6}$$

De plus $f \in [\mathscr{Y}]$ est $\mu \circ h^{-1}$-intégrable si et seulement si $f \circ h$ est μ-intégrable et dans ce cas, f vérifie (A.6).

Démonstration. On considère la fonctionnelle $I(f) = \int f \circ h \, \mathrm{d}\mu$, $f \in \mathscr{Y}^+$ et on applique la Proposition A.43. La mesure associée à I est

$$\nu(A) = \int \mathbb{1}_A \circ h \, \mathrm{d}\mu = \int \mathbb{1}_{h^{-1}(A)} \, \mathrm{d}\mu = \mu(h^{-1}(A)).$$

On conclut facilement. □

► **Mesures à densité**

Théorème A.45. *Soient $(\mathsf{E}, \mathscr{E}, \mu)$ un espace mesuré et $h \in \mathscr{E}^+$. La formule*

$$\nu(A) := \int_A h \, \mathrm{d}\mu, \qquad A \in \mathscr{E},$$

définit une mesure sur $\mathscr{E}$, notée $h \cdot \mu$. On a, pour toute $f \in \mathscr{E}^+$,

$$\int f \, \mathrm{d}\nu = \int fh \, \mathrm{d}\mu. \tag{A.7}$$

De plus $f \in [\mathscr{E}]$ est ν-intégrable si et seulement si fh est μ-intégrable et dans ce cas, f vérifie la relation (A.7). Enfin, si $\mu(\{h \neq h'\}) = 0$ alors les mesures $h \cdot \mu$ et $h' \cdot \mu$ coïncident.

Démonstration. On considère la fonctionnelle $I(f) = \int fh \, \mathrm{d}\mu$, $f \in \mathscr{E}^+$, et on applique la Proposition A.43. La dernière assertion s'établit en écrivant $f = f^+ - f^-$. □

Définition A.46 (densité d'une mesure, absolue continuité, domination, mesure étrangère). *Soient $(\mathsf{E}, \mathscr{E})$ un espace mesurable et μ une mesure sur $\mathscr{E}$.*

(i) Pour $f \in \mathscr{E}^+$, la mesure $\nu : B \mapsto \int_B f \mathrm{d}\mu$ sur $\mathscr{E}$ est appelée mesure de densité f par rapport à μ. *Elle est notée $\nu = f \cdot \mu$ et la densité (unique à un ensemble μ-négligeable près) est notée $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.*

(ii) Une mesure ν est dite absolument continue par rapport à μ *si, pour tout $B \in \mathscr{E}$ tel que $\mu(B) = 0$, nous avons $\nu(B) = 0$. Lorsque ν est absolument continue par rapport à μ, on dit aussi que la mesure μ* domine *ν.*

(iii) Une mesure ν est dite étrangère à μ *s'il existe $N \in \mathscr{E}$ tel que $\mu(N) = 0$ et $\nu(N^c) = 0$. Lorsque ν est étrangère à μ, on dit aussi que ν est* μ-singulière.

Par exemple, la mesure de Lebesgue sur $\mathbb{R}$ et la mesure Gaussienne $\mathrm{N}(0,1)$ sont absolument continues l'une par rapport à l'autre. En revanche, la mesure de Lebesgue λ_{Leb} sur $\mathbb{R}$ et la mesure de comptage μ sur $\mathbb{N}$ sont étrangères puisque $\lambda_{\mathrm{Leb}}(\mathbb{N}) = 0$ et $\mu(\mathbb{N}^c) = 0$.

Nous admettrons le résultat très important suivant :.

Théorème A.47 (Radon-Nikodym). *Soit μ une mesure σ-finie sur un espace mesurable $(\mathsf{E}, \mathscr{E})$. Soit ν une mesure absolument continue par rapport à μ. Il existe une fonction mesurable positive f, unique à un ensemble μ-négligeable près, telle que l'on ait $\nu = f \cdot \mu$. De plus,*

(i) la fonction f est μ-presque partout finie si et seulement si la mesure ν est σ-finie.

(ii) la fonction f est μ-intégrable si et seulement si la mesure ν est bornée.

A.3.5 Intégration par rapport à une mesure produit

Théorème A.48 (Théorème de Fubini). *Soient $(\mathsf{E}_1, \mathscr{E}_1, \mu_1)$ et $(\mathsf{E}_2, \mathscr{E}_2, \mu_2)$ deux espaces mesurés avec μ_1 et μ_2 σ-finies.*

Il existe une unique mesure sur $\mathscr{E}_1 \otimes \mathscr{E}_2$, notée $\mu_1 \otimes \mu_2$ et appelée mesure produit de μ_1 et μ_2, telle que,

$$\text{pour tout } A_1 \in \mathscr{E}_1,\ A_2 \in \mathscr{E}_2,\ \ \mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1)\,\mu(A_2).$$

De plus, pour toute fonction $f \in (\mathscr{E}_1 \otimes \mathscr{E}_2)^+$,

$$\int f \,\mathrm{d}\mu_1 \otimes \mu_2 = \int \left[\int f(x_1,x_2)\,\mathrm{d}\mu_1(x_1) \right] \mathrm{d}\mu_2(x_2) = \int \left[\int f(x_1,x_2)\,\mathrm{d}\mu_2(x_2) \right] \mathrm{d}\mu_1(x_1).$$

Soit $f \in \mathrm{L}^1_{\mathbb{C}}(\mu_1 \otimes \mu_2)$. Alors,

(i) $\int |f(x_1,x_2)|\,\mathrm{d}\mu_2(x_2) < +\infty$ *μ_1 p.p. et* $x_1 \mapsto \int f(x_1,x_2)\,\mathrm{d}\mu_2(x_2) \in \mathrm{L}^1_{\mathbb{C}}(\mu_1)$,

(ii) $\int |f(x_1,x_2)|\,\mathrm{d}\mu_1(x_1) < +\infty$ *μ_2 p.p. et* $x_2 \mapsto \int f(x_1,x_2)\,\mathrm{d}\mu_1(x_1) \in \mathrm{L}^1_{\mathbb{C}}(\mu_2)$, *et*

$$\int f \,d\mu_1 \otimes \mu_2 = \int \left[\int f(x_1,x_2)\,\mathrm{d}\mu_1(x_1) \right] \mathrm{d}\mu_2(x_2) = \int \left[\int f(x_1,x_2)\,\mathrm{d}\mu_2(x_2) \right] \mathrm{d}\mu_1(x_1).$$

Démonstration. (i) On applique la Proposition A.28 à

$$\mathscr{C} := \{A,\ A = A_1 \times A_2,\ A_1 \in \mathscr{E}_1,\ A_2 \in \mathscr{E}_2\,,\ \mu(A_1) < +\infty,\ \mu(A_2) < +\infty\}\ .$$

(ii) On applique la proposition A.43 à $I_1(f) := \int [\int f(x_1,x_2)\,\mathrm{d}\mu_1(x_1)]\,\mathrm{d}\mu_2(x_2)$ ce qui donne l'existence. Mais on peut aussi appliquer la proposition A.43 à

$$I_2(f) := \int \left[\int f(x_1,x_2)\,\mathrm{d}\mu_2(x_2) \right] \mathrm{d}\mu_1(x_1)$$

et, vu l'unicité, on a $I_1(f) = I_2(f)$.

(iii) Si $f \in \mathrm{L}^1_{\mathbb{C}}(\mu_1 \otimes \mu_2)$, on applique le résultat précèdent à $[\mathrm{Re}(f)]^+$, $[\mathrm{Re}(f)]^-$, $[\mathrm{Im}(f)]^+$ et $[\mathrm{Im}(f)]^-$. □

Tout ceci s'étend sans (trop de) peine au cas de n espaces mesurables. Il y a quelques vérifications fastidieuses à faire du type $\mu_1 \otimes (\mu_2 \otimes \mu_3) = (\mu_1 \otimes \mu_2) \otimes \mu_3$. De plus dans la formule d'intégrations successives, les variables peuvent être intégrées dans tous les ordres possibles. A ce sujet, le grand principe est : si f est positive, tout est permis, et l'intégrale est à valeur dans $\overline{\mathbb{R}^+}$; si f est de signe quelconque ou complexe, on applique d'abord le théorème à $|f|$, et si l'intégrale est finie (i.e. $|f|$ intégrable), on peut appliquer le théorème à la fonction f.

Annexe B

Inversion Locale et Globale

Soient $\mathscr{U}$ et $\mathscr{V}$ deux ensembles ouverts de $\mathbb{R}^d$. Nous dirons que la fonction $\mathbf{e} : \mathscr{U} \mapsto \mathscr{V}$ est un *différomorphisme* de classe C^1 de $\mathscr{U}$ sur $\mathscr{V}$ si et seulement si

(i) $\mathbf{e}$ est continûment différentiable sur $\mathscr{U}$,

(ii) $\mathbf{e}$ est une bijection de $\mathscr{U}$ sur $\mathscr{V}$,

(iii) $\mathbf{e}^{-1}$ est continûment différentiable sur $\mathscr{V}$.

Il résulte de la définition que

$$\mathbf{e}^{-1} \circ \mathbf{e} = \mathrm{I}_{\mathscr{U}} \quad \text{et} \quad \mathbf{e} \circ \mathbf{e}^{-1} = \mathrm{I}_{\mathscr{V}} \ .$$

Pour tout $\theta \in \mathscr{U}$, nous avons

$$\mathrm{J}_{\mathbf{e}^{-1}}(\mathbf{e}(\theta)) \cdot \mathrm{J}_{\mathbf{e}}(\theta) = \mathrm{I}_{\mathbb{R}^d} \ ,$$

où $\mathrm{J}_{\mathbf{e}^{-1}}(\vartheta)$ est la matrice jacobienne de $\mathbf{e}^{-1}$ au point ϑ et $\mathrm{J}_{\mathbf{e}}(\theta)$ est la matrice jacobienne de $\mathbf{e}$ au point θ. De même, pour tout $\theta \in \mathscr{V}$, nous avons

$$\mathrm{J}_{\mathbf{e}}(\mathbf{e}^{-1}(\theta)) \cdot \mathrm{J}_{\mathbf{e}^{-1}}(\theta) = \mathrm{I}_{\mathbb{R}^d} \ .$$

Les matrices jacobiennes $\mathrm{J}_{\mathbf{e}}(\theta)$ et $\mathrm{J}_{\mathbf{e}^{-1}}(\theta)$ sont donc inversibles et

$$[\mathrm{J}_{\mathbf{e}}(\theta)]^{-1} = \mathrm{J}_{\mathbf{e}^{-1}}(\mathbf{e}(\theta)) \ .$$

Le théorème d'inversion locale (que nous admettrons) montre que si une fonction $\mathbf{e}$ est continûment différentiable en un point et si sa différentielle en ce point est inversible alors, localement, $\mathbf{e}$ est inversible et son inverse est différentiable.

Théorème B.1 (Inversion locale). *Soit Θ un sous-ensemble ouvert de $\mathbb{R}^d$, $\theta \in \Theta$ et $\mathbf{e} : \Theta \to \mathbb{R}^d$ une application continûment différentiable. Si la matrice jacobienne $\mathrm{J}_{\mathbf{e}}(\theta)$ est inversible, alors il existe un voisinage ouvert $\mathscr{U}$ de θ, $\mathscr{V}$ un voisinage ouvert de $\mathbf{e}(\theta)$ telle que $\mathbf{e}$ soit un difféomorphisme de $\mathscr{U}$ sur $\mathscr{V}$. De plus, pour tout $\theta \in \mathscr{U}$,*

$$[\mathrm{J}_{\mathbf{e}}(\theta)]^{-1} = \mathrm{J}_{\mathbf{e}^{-1}}(\mathbf{e}(\theta)) \ . \tag{B.1}$$

Si, pour tout $\theta \in \Theta$, la matrice jacobienne $\mathrm{J}_{\mathbf{e}}(\theta)$ est inversible, on peut appliquer le théorème d'inversion locale en chaque point de $\theta \in \Theta$, ce qui assure que $\mathbf{e}(\Theta)$ contient un voisinage $\mathscr{V}$ de $\mathbf{e}(\theta)$. Par conséquent, $\mathbf{e}(\Theta)$ est un ensemble ouvert.

Corollaire B.2. *Soient Θ un ouvert de $\mathbb{R}^d$ et $\mathbf{e} : \Theta \to \mathbb{R}^d$ une application continûment différentiable. Si, pour tout $\theta \in \Theta$, la matrice jacobienne $\mathrm{J}_{\mathbf{e}}(\theta)$ est inversible, alors l'application $\mathbf{e}$ est ouverte, i.e. l'image de tout ouvert de Θ est ouvert dans $\mathbb{R}^d$. En particulier, $\mathbf{e}(\Theta)$ est un ouvert.*

On en déduit le théorème d'inversion globale

Théorème B.3 (Inversion globale). *Soient Θ un ouvert de $\mathbb{R}^d$ et $\mathbf{e} : \Theta \to \mathbb{R}^d$ une application injective continûment différentiable. Si la matrice jacobienne $\mathrm{J}_{\mathbf{e}}(\theta)$ est inversible pour tout $\theta \in \Theta$, alors $\mathscr{V} = \mathbf{e}(\Theta)$ est un voisinage ouvert de $\mathbb{R}^d$ et $\mathbf{e}$ est un difféomorphisme de Θ sur $\mathscr{V}$.*

Index