

Statistique et Apprentissage

Notes du cours MAP433

Gersende Fort, Matthieu Lerasle, Eric Moulines

27 novembre 2020

Notations

Ensembles

- $\mathbb{N}$: l'ensemble des entiers naturels, $\mathbb{N} = \{0, 1, 2, \dots\}$.
- $\mathbb{N}^*$: l'ensemble des entiers naturels, privé de 0, $\mathbb{N}^* = \{1, 2, \dots\}$.
- $\tilde{\mathbb{N}}$: l'ensemble des entiers naturels étendu, $\tilde{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$.
- $\mathbb{Z}$: l'ensemble des entiers relatifs, $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$.
- $\mathbb{R}$: l'ensemble des réels.
- $\mathbb{R}^d$: l'ensemble des vecteurs colonnes $\mathbf{x} = (x^{(1)}, \dots, x^{(d)})'$.
- $\bar{\mathbb{R}}$: l'ensemble des réels complétés, $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.
- $\lceil x \rceil$: 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 $A \subset X$, $d(x, A) = \inf\{d(x, y) : y \in A\}$.
- $B(x, r)$: la boule ouverte de rayon $r > 0$ centrée en x ,

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

- $\bar{U}$: la fermeture de $U \subset X$ (l'intersection de l'ensemble des fermés contenant U).
- U° : l'intérieur de U (l'union des ouverts inclus dans U).
- ∂U : la frontière de $U \subset X$, $\partial U = \bar{U} \setminus U^\circ$.

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 $\varepsilon_n \rightarrow 0$ telle que $(1 - \varepsilon_n)b_n \leq a_n \leq (1 + \varepsilon_n)b_n$.

Vecteurs, matrices

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

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

- $I_d : d \times d$ matrice identité.

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 , $\|f\|_\infty = \sup\{f(x) : x \in X\}$ est la norme-sup.

Espaces de fonctions

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

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

Mesures

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

- δ_x : mesure ponctuelle en x , i.e. $\delta_x(A) = 1$ si $x \in A$ et 0 autrement.
- λ_{Leb} : mesure de Lebesgue sur $\mathbb{R}^d$.
- $\mathbb{M}_s(X)$: l'ensemble des mesures signées sur $(X, \mathcal{X})$.
- $\mathbb{M}_+(X)$: l'ensemble des mesures positives sur $(X, \mathcal{X})$.
- $\mathbb{M}_1(X)$: l'ensemble des probabilités sur $(X, \mathcal{X})$.
- $\mu \ll \nu$: 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 $\mathcal{X} = \mathcal{B}(X)$ est la tribu borélienne, c'est-à-dire la tribu engendrée par la topologie de X .

Probabilité

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

- $\mathbb{E}(X)$, $\mathbb{E}[X]$: l'espérance (quand elle est définie) du vecteur aléatoire X par rapport à la probabilité $\mathbb{P}$, $\mathbb{E}[X] = \int X d\mathbb{P}$.
- $\text{Cov}(X, Y)$: la matrice de covariance (ou variance-covariance) des vecteurs aléatoires X et Y , $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T]$.
- $\mathcal{L}_{\mathbb{P}}(X)$: distribution de la variable aléatoire X sur $(X, \mathcal{X})$ sous $\mathbb{P}$, i.e. la probabilité image de $\mathbb{P}$ par X .
- $X_n \xrightarrow{\mathbb{P}} X$: la suite de variables aléatoires $\{X_n, n \in \mathbb{N}\}$ converge en loi vers X sous $\mathbb{P}$.
- $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$: la suite de variables aléatoires $\{X_n, n \in \mathbb{N}\}$ converge vers X en probabilité sous $\mathbb{P}$.
- $X_n \xrightarrow{\mathbb{P}\text{-p.s.}} X$: la suite de variables aléatoires $\{X_n, n \in \mathbb{N}\}$ converge vers X $\mathbb{P}$ -presque-sûrement.

Distributions usuelles

- $\text{Ber}(p)$: distribution de Bernoulli de probabilité de succès p .
- $\text{Bin}(n, p)$: distribution binômiale ; distribution du nombre de succès après n tirages indépendants de Bernoulli de probabilité de succès p .
- $N(\mu, \sigma^2)$: distribution gaussienne (ou *normale*) de moyenne μ et variance σ^2 .
- $\text{Unif}(a, b)$: distribution uniforme sur $[a, b]$.
- χ_n^2 : loi du chi-2 à n degrés de liberté
- $\text{Poi}(\lambda)$: loi de Poisson d'intensité λ .
- $t(n)$: loi de Student à n degrés de liberté.

Table des matières

I	Inférence statistique	11
I-1	Modèles statistiques	13
I-1.1	Exemples introductifs	13
I-1.2	Formulation mathématique	15
I-2	Estimation ponctuelle	23
I-2.1	Méthode des moments	23
I-2.2	Z-estimateurs	25
I-2.3	Maximum de vraisemblance	27
I-2.4	M-estimateurs	34
I-3	Tests et régions de confiance	39
I-3.1	Tests d'hypothèse	39
I-3.2	p -valeur	45
I-3.3	Régions de confiance	47
I-3.4	Dualité entre régions de confiance et tests d'hypothèse de base simple	52
I-3.5	Méthode du pivot	53
I-3.6	Utilisation d'inégalités de déviation	55
I-4	Bases de la théorie de la décision	59
I-4.1	Règles de décision, pertes et risques	59
I-4.2	Le cas du risque quadratique : estimation non biaisée	63
I-5	Optimalité en théorie des tests	75
I-5.1	Tests uniformément plus puissants	75
I-5.2	Rapport de vraisemblance monotone	83
II	Statistiques asymptotiques	91
II-1	Introduction aux statistiques asymptotiques	93
II-1.1	Consistance d'une suite d'estimateurs	95
II-1.2	Vitesse de convergence et Normalité Asymptotique	96
II-1.3	Etude asymptotique des estimateurs de moments	100
II-1.4	Régions de confiance asymptotiques	104
II-1.5	Transformations de stabilisation de la variance	105
II-1.6	Tests asymptotiques	110
II-2	Théorie asymptotique des (M,Z)-estimateurs	115
II-2.1	Consistance des Z- et des M-estimateurs	115
II-2.2	Loi limite des Z- et M-estimateurs	123

II-3	M.V., information statistique et optimalité	129
II-3.1	Consistance de l'estimateur du M.V.	129
II-3.2	Loi limite de l'estimateur du M.V.	131
II-3.3	Efficacité asymptotique	132
II-3.4	Méthode du score de Fisher	136
II-4	Tests asymptotiques	141
II-4.1	Introduction	141
II-4.2	Tests d'adéquation	143
II-4.3	Les tests de Wald, de Rao et du rapport de vraisemblance	146
III	Fondamentaux de l'apprentissage statistique	159
III-1	Classification supervisée	161
III-1.1	La classification binaire	161
III-1.2	Classification bayésienne	166
III-1.3	Risque moyen, excès de risque	168
III-1.4	Sur-apprentissage	169
III-1.5	Consistance	174
III-2	Apprentissage PAC	175
III-2.1	Minimiseur du risque empirique	176
III-2.2	Une borne PAC élémentaire	176
III-2.3	Une borne PAC agnostique	178
III-2.4	Une application : classification par histogramme	180
III-2.5	Conclusion partielle	181
III-3	Théorie de Vapnik-Chervonenkis	183
III-3.1	Inégalité de McDiarmid	183
III-3.2	Inégalité de Vapnik-Chervonenkis	185
IV	Outils probabilistes	195
IV-1	Éléments de la théorie de la mesure et de l'intégration	197
IV-1.1	Tribus et Mesurabilité	197
IV-1.2	Mesures	206
IV-1.3	Intégration	209
IV-2	Modes de convergence et théorèmes limites	219
IV-2.1	Convergence en probabilité	219
IV-2.2	Convergence presque-sûre	221
IV-2.3	Loi des grands nombres	223
IV-2.4	Convergence en loi	228
IV-2.5	Théorème de la limite centrale	234
IV-2.6	Convergence des moments	244
IV-2.7	Symboles o et O stochastiques	247
IV-3	Inégalités de déviations	249
IV-3.1	Inégalités de Markov et de Bienayme-Tchebychev	249
IV-3.2	Inégalité de Jensen	250
IV-3.3	Inégalités de Chernoff	250
IV-3.4	Inégalité de Hoeffding	252
IV-3.5	Inégalité de Pisier	254

IV-4	Fonction de répartition, quantiles et statistiques d'ordre	257
IV-4.1	Fonction de répartition	257
IV-4.2	Quantiles	261
IV-4.3	Statistiques d'ordre	263
IV-5	Famille de distributions	269
IV-5.1	Loi gaussienne	269
IV-5.2	Loi gaussienne multivariée	270
IV-5.3	Loi Gamma	272
IV-5.4	La loi du χ^2 à k degrés de liberté	272
IV-5.5	Loi de Cauchy	274
IV-5.6	Loi de Student	275
IV-5.7	Loi de Fisher	277
IV-6	Famille Exponentielle	279

Première partie

Inférence statistique

Chapitre I-1

Modèles statistiques

I-1.1 Exemples introductifs

L'objectif de cette section est de dégager la notion de modèle statistique. Avant de procéder à la construction mathématique, 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 observe une fonction $t \mapsto f(t)$ à des instants multiples de T_e , sur un intervalle de temps $[0, T]$: on observe donc $f(kT_e)$, $k = 1, \dots, n$, où $n := \lfloor T/T_e \rfloor$ - la partie entière inférieure de T/T_e . Chaque observation est affectée par une erreur de mesure, de sorte que l'on ne collecte pas directement les valeurs $\{f(kT_e), k = 1, \dots, n\}$ mais

$$y_k = f(kT_e) + e_k, \quad k = 1, \dots, n.$$

On dispose de n données. Dans ce problème, on est typiquement intéressé par "estimer" la fonction f . Ceci signifie que nous allons chercher à construire une fonction $t \mapsto \hat{f}(t; \{y_k\}_{k=1}^n)$ ne dépendant que des échantillons mesurés $\{y_k\}_{k=1}^n$, 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 $\{e_k\}_{k=1}^n$, 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 les $n = N^2$ données

$$y_{k,\ell} = f(k/N, \ell/N) + e_{k,\ell}, \quad 1 \leq k, \ell \leq N.$$

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

- 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.

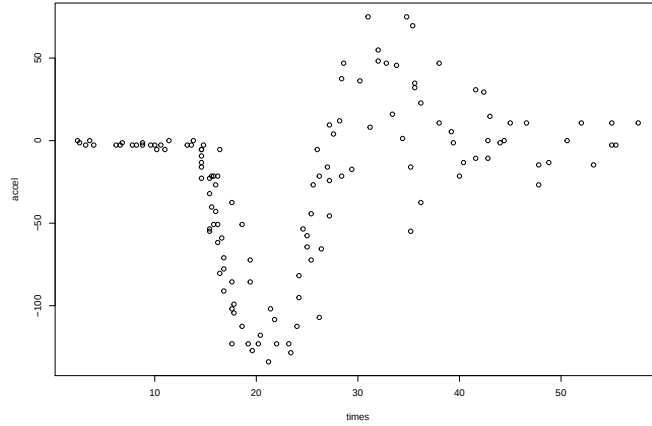


FIGURE I-1.1 – Accélération de la tête en fonction du temps suite à un impact

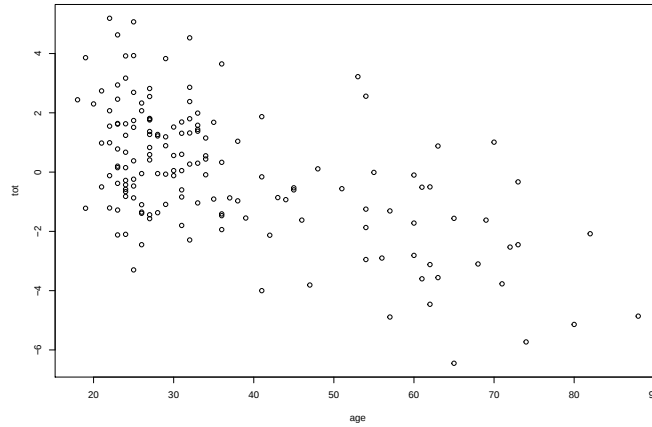


FIGURE I-1.2 – Qualité d’un greffon de rein en fonction de l’âge du donneur (données du laboratoire de néphrologie du Dr. B. Myers, Université de Stanford)

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 $x_i^* = \min\{x_i, \tau\}$, 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. $\diamond$

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 $f : \mathbb{R} \rightarrow \mathbb{R}$ telle que $y = f(x)$. Toutefois, il est raisonnable de penser que y_i soit en partie expliquée par x_i , ce qui revient à postuler le modèle

$$y_i = f(x_i) + e_i, \quad i \in \{1, \dots, n\}, \quad (\text{I-1.1})$$

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, la représentation (I-1.1) devient $y_i = f(\mathbf{x}_i) + e_i$ où maintenant f est une fonction $\mathbb{R}^k \rightarrow \mathbb{R}$.

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 $f(\mathbf{x}) = t(\theta^T \mathbf{x})$, où $t : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction connue et $\theta = (\theta_1, \dots, \theta_k)^T \in \mathbb{R}^k$ sont des paramètres inconnus. Dans les cas les plus élémentaires, $t(z) = z$, le modèle (I-1.1) s'écrit

$$y_i = \theta^T \mathbf{x}_i + e_i, \quad i \in \{1, \dots, n\}, \quad (\text{I-1.2})$$

et est appelé le modèle 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. $\diamond$

I-1.2 Formulation mathématique

Construire un modèle statistique consiste à identifier trois éléments distincts :

Des données $z = (x_1, \dots, x_n)$ où pour tout $i \in \{1, \dots, n\}$, $x_i \in \mathbb{R}^d$ sont des vecteurs, mais on peut imaginer des situations plus complexes¹. 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 $Z = (X_1, \dots, X_n)$ dont les mesures $z = (x_1, \dots, x_n)$ 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 les paramètres inconnus, ou à prédire la valeur d'un signal en un point non observé. Il faut alors contrôler la qualité de cette estimation. On va aussi chercher 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 $(Z, \mathcal{Z})$, l'espace des observations,*
- *une famille de probabilités $\mathcal{C}$ sur $(Z, \mathcal{Z})$.*

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.

Nous notons $(Z, \mathcal{Z}, \mathcal{C})$ le modèle statistique.

Le modèle est dit paramétrique lorsque $\mathcal{C}$ 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 $\theta \in \Theta \subset \mathbb{R}^d$ un élément $\mathbb{P}_\theta \in \mathcal{C}$.

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

Exemple I-1.6 (Modèle de sondage). Considérons tout d'abord l'exemple I-1.1. On dispose d'une population de N individus (N est typiquement très grand). Dans cette population, $N\theta$ individus votent pour le candidat A, où $\theta \in \Theta = \{k/N, k \in \{0, \dots, N\}\}$. Supposons tout d'abord qu'on procède à un tirage uniforme de n individus, avec remplacement 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 canonique associé est alors donné par

$$\left(\{0, 1\}^n, \mathcal{P}(\{0, 1\}^n), \{\text{Ber}_\theta^{\otimes n}, \theta \in \Theta\} \right),$$

où $\mathcal{P}(\{0, 1\}^n)$ est l'ensemble des parties de $\{0, 1\}^n$ et $\text{Ber}_\theta^{\otimes n}$ est le produit de n lois de Bernoulli de paramètre θ (voir Section IV-1.1.5 et Section IV-1.3.5), i.e.

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

Les tirages sont indépendants et le résultat de chaque tirage est une variable de Bernoulli de paramètre $\theta \in \Theta$.

Une autre manière de procéder est de considérer qu'on n'observe que le nombre de votants n_A pour le candidat A, ce qui détermine aussi n_B puisque $n_A + n_B = n$. Dans ce cas, l'espace des observations est $Z = \{0, \dots, n\}$ qui est muni de $\mathcal{P}(Z)$ l'ensemble des parties de Z . La famille $\mathcal{C}$ 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 canonique s'écrit

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

Supposons maintenant que les tirages sont sans remplacement. L'espace des observations est encore $Z = \{0, \dots, n\}$ et $\mathcal{Z} = \mathcal{P}(Z)$. 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 $\text{Hyper}(N\theta, N, n)$. $\diamond$

Exemple I-1.7. Dans le problème de reconstruction d'un signal (Exemple I-1.2), les observations sont clairement à valeurs dans $Z = \mathbb{R}^n$ avec $n = \lfloor T/T_e \rfloor$. Il est naturel de munir cet ensemble de sa tribu borélienne (voir Définition IV-1.3) : $\mathcal{Z} = \mathcal{B}(\mathbb{R}^n)$. Pour construire un modèle statistique, il faut se donner un ensemble $\mathcal{C}$ 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 d'une suite $(\xi_1, \dots, \xi_n)$ 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 $\theta \in \Theta$, la loi du vecteur d'observations est la loi Gaussienne sur $\mathbb{R}^p$, 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 $\sigma^2 \mathbf{I}_{n \times n}$ dont la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ est donnée par :

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

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

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

Pour toute loi $\mathbb{P} \in \mathcal{C}$, la statistique T comme application de $(Z, \mathcal{Z})$ dans $(T, \mathcal{T})$ 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

$$\mathcal{C}^T = \{\mathbb{P}^T, \mathbb{P} \in \mathcal{C}\},$$

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

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

Pour servir de support à l'intuition, il est souvent pratique d'introduire des statistiques associées aux observations individuelles.

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

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

Notons X_i la statistique correspondant au résultat du i -ème sondage, $i \in \{1, \dots, n\}$: pour tout $\mathbf{x} = (x_1, \dots, x_n) \in \{0, 1\}^n$, $X_i(\mathbf{x}) = x_i$. En notant $\mathbb{P}_{\theta} = \text{Ber}_{\theta}^{\otimes n}$, 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 $X_1, \dots, X_n$ sont indépendantes : pour tout $(x_1, \dots, x_n) \in \{0, 1\}^n$, nous avons

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

où $Z = (X_1, \dots, X_n)$. Pour le modèle canonique $(Z, \mathcal{Z}, \mathcal{C})$, les statistiques $(X_1, \dots, X_n)$ (que nous appelons aussi les "observations") sont indépendantes et identiquement distribuées (i.i.d.) : pour tout $\theta \in \Theta$ et $i \in \{1, \dots, n\}$, l'observation X_i est distribuée suivant une loi de Bernoulli de paramètre θ . $\diamond$

Exemple I-1.11. Considérons le modèle canonique

$$(Z, \mathcal{Z}, \mathcal{C}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_{\theta}, \theta \in \Theta = \mathbb{R} \times \mathbb{R}_+^*\})$$

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

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

Notons X_i la i -ème observation (dite aussi "statistique associée à la i -ème donnée collectée"), $i \in \{1, \dots, n\}$. Cette statistique est définie, pour $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ 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

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

Les statistiques $Z = (X_1, \dots, X_n)$ sont i.i.d. de densité gaussienne de moyenne $\mu \in \mathbb{R}$ et variance $\sigma^2 \in \mathbb{R}_+^*$. En effet,

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

$\diamond$

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

*Remarque * I-1.13.* Cette discussion est plus avancée et peut-être omise en première lecture. Le théorème de Radon-Nykodim (Théorème IV-1.48) montre que le modèle $(Z, \mathcal{Z}, \mathcal{C})$ est dominé si et seulement si toute loi de $\mathbb{P} \in \mathcal{C}$ est absolument continue par rapport à μ (voir Définition IV-1.47).

La mesure de domination peut être choisie de multiples façons. Soient μ et ν sont deux mesures σ -finies distinctes sur $(Z, \mathcal{Z})$. Supposons que le modèle statistique canonique paramétrique $(Z, \mathcal{Z}, \mathcal{C}) = \{(Z, \mathcal{Z}), \{\mathbb{P}_\theta, \theta \in \Theta\}\}$ 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 $B \in \mathcal{Z}$, $\lambda(B) = \mu(B) + \nu(B) = 0$ implique que $\mu(B) = 0$ et $\nu(B) = 0$. En appliquant le théorème IV-1.48, 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 $\theta \in \Theta$, $p_\theta m = q_\theta n$ λ -p.p. Posons $Z_0 = \{z \in Z, m(z) > 0\}$. Notons que

$$\mu(Z_0^c) = \int_{Z_0^c} m(z) \lambda(dz) = 0,$$

et donc, pour tout $\theta \in \Theta$, $\mathbb{P}_\theta(Z_0^c) = 0$. Pour tout $z \in Z_0$ 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-1.14. 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-1.15. Un exemple plus subtil est donné par le modèle statistique canonique dont la famille de loi est donnée par

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

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. ◇

I-1.2.2 Modèle statistique produit et n -échantillon

Soient $(X, \mathcal{X}, \mathcal{C})$ et $(X', \mathcal{X}', \mathcal{C}')$ deux modèles statistiques. Nous allons construire le produit de ces deux modèles. Un élément z de l'espace des observations $Z = X \times X'$ est donc un couple $z = (x, x')$ avec $x \in X$ et $x' \in X'$. Nous munissons cet espace de la tribu produit $\mathcal{Z} = \mathcal{X} \otimes \mathcal{X}'$ qui est par définition la plus petite tribu qui contient les pavés mesurables $A \times A'$ où $A \in \mathcal{X}$ et $A' \in \mathcal{X}'$ (voir Section IV-1.1.5). Pour $\mathbb{Q} \in \mathcal{C}$ et $\mathbb{Q}' \in \mathcal{C}'$, $\mathbb{Q} \otimes \mathbb{Q}'$ est la mesure produit (voir Section IV-1.3.5) définie, pour tout $(A, A') \in \mathcal{X} \times \mathcal{X}'$ par :

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

Considérons le modèle statistique

$$(Z, \mathcal{Z}, \mathcal{C} \otimes \mathcal{C}') \quad \text{où} \quad \mathcal{C} \otimes \mathcal{C}' := \{ \mathbb{Q} \otimes \mathbb{Q}', \mathbb{Q} \in \mathcal{C}, \mathbb{Q}' \in \mathcal{C}' \} . \quad (\text{I-1.3})$$

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

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

Pour tout $\mathbb{P} = \mathbb{Q} \otimes \mathbb{Q}' \in \mathcal{C} \otimes \mathcal{C}'$ et $(A, A') \in \mathcal{X} \times \mathcal{X}'$, 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 $(X, \mathcal{X}, \mathcal{C})$ et $(X', \mathcal{X}', \mathcal{C}')$. Considérer des produits de modèles statistiques correspond, dans la pratique, à étudier des systèmes d'observations indépendantes.

Définition I-1.16 (Produit de modèles statistiques). Soient $(X, \mathcal{X}, \mathcal{C})$ et $(X', \mathcal{X}', \mathcal{C}')$ deux modèles statistiques. On appelle produit de ces deux modèles et on note $(X, \mathcal{X}, \mathcal{C}) \otimes (X', \mathcal{X}', \mathcal{C}')$ le modèle statistique $(X \times X', \mathcal{X} \otimes \mathcal{X}', \mathcal{C} \otimes \mathcal{C}')$ où :

$$\mathcal{C} \otimes \mathcal{C}' := \{ \mathbb{P} = \mathbb{Q} \otimes \mathbb{Q}', \mathbb{Q} \in \mathcal{C}, \mathbb{Q}' \in \mathcal{C}' \} .$$

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 $(X, \mathcal{X}, \mathcal{C})$ un modèle statistique et $n \in \mathbb{N}^*$. Posons

$$Z = X^n, \quad \mathcal{Z} = \mathcal{X}^{\otimes n} .$$

Un élément z de l'espace des observations Z est donc un n -uplet $z = (x_1, \dots, x_n)$ où, pour tout $i \in \{1, \dots, n\}$, $x_i \in X$. Soit $\mathcal{X}^{\otimes n} = \mathcal{X} \otimes \dots \otimes \mathcal{X}$ la tribu produit (voir Section IV-1.1.5) i.e. la plus petite tribu contenant les pavés mesurables $A_1 \times \dots \times A_n$, où $(A_1, \dots, A_n) \in \mathcal{X}^n$. Pour tout $\mathbb{Q} \in \mathcal{C}$, notons $\mathbb{Q}^{\otimes n}$ la probabilité produit sur $\mathcal{X}^{\otimes n}$ (voir Section IV-1.3.5) définie, pour tout $(A_1, \dots, A_n) \in \mathcal{X}^n$ par

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

Considérons le modèle statistique

$$(X, \mathcal{X}, \mathcal{C})^n := (X^n, \mathcal{X}^{\otimes n}, \{ \mathbb{Q}^{\otimes n}, \mathbb{Q} \in \mathcal{C} \}) , \quad (\text{I-1.4})$$

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

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

alors que

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

Définition I-1.17 (n -échantillon). Soient $(X, \mathcal{X}, \mathcal{C})$ un modèle statistique et $n \in \mathbb{N} \setminus \{0\}$. On appelle n -échantillon de $(X, \mathcal{X}, \mathcal{C})$ le modèle statistique

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

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

Lemme I-1.18. Soit $(X, \mathcal{X}, \mathcal{C})^n$ un n -échantillon du modèle $(X, \mathcal{X}, \mathcal{C})$.

- (i) Les observations $Z = (X_1, \dots, X_n)$ sont indépendantes.
- (ii) Pour tout $i \in \{1, \dots, n\}$, le modèle induit par la statistique X_i est $(X, \mathcal{X}, \mathcal{C})$.

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

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

Nous prendrons quelques libertés avec la terminologie dans la suite de l'exposé pour éviter d'alourdir les énoncés des résultats. Nous dirons par exemple, conformément à un usage très répandu dans la littérature statistique

“($X_1, \dots, X_n$) est un n -échantillon du modèle $(X, \mathcal{X}, \mathcal{C})$ ”
ou
“Soient $(X_1, \dots, X_n)$ n variables indépendantes et identiquement distribuées de loi $\mathbb{Q} \in \mathcal{C}$ ”

Ces deux expressions synthétisent plusieurs informations :

- On dispose d'un n -échantillon $(X, \mathcal{X}, \mathcal{C})^n$ d'un modèle statistique $(X, \mathcal{X}, \mathcal{C})$ (l'espace des observations est le plus souvent implicite tout comme la tribu des événements).
- Les observations $(X_1, \dots, X_n)$ définies pour $i \in \{1, \dots, n\}$ par $X_i : X^n \rightarrow X, X_i(x_1, \dots, x_n) = x_i$ sont des statistiques indépendantes et de même loi, et le modèle induit par X_i est $(X, \mathcal{X}, \mathcal{C})$.

Soit $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$ un modèle paramétrique. Supposons qu'il existe une mesure σ -finie μ sur $(X, \mathcal{X})$ 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 $\mathbb{P}_\theta = \mathbb{Q}_\theta^n$ est dominée par $\mu^{\otimes n}$ et admet pour densité

$$(x_1, \dots, x_n) \mapsto p(\theta, x_1, \dots, x_n) = \prod_{i=1}^n q(\theta, x_i). \quad (\text{I-1.5})$$

Exemple I-1.19. 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 $\theta \in \mathbb{R}_+ \setminus \{0\}$, de densité par rapport à la mesure de Lebesgue λ_{Leb} donnée par

$$q_\theta(x) = q(\theta, x) = \theta e^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x). \quad (\text{I-1.6})$$

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

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

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

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

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 et pour tout i , l'expérience statistique induite par X_i est

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{q_{\theta} \cdot \lambda_{\text{Leb}}, \theta \in \Theta = \mathbb{R}_+^*\}) .$$

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

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

Notons que la loi $\mathbb{Q}_{\theta}^*$ n'est pas absolument continue par rapport à la mesure de Lebesgue. En revanche, la famille $\{\mathbb{Q}_{\theta}^*, \theta \in \Theta\}$ est dominée par $\mu = \lambda_{\text{Leb}} + \delta_{\tau}$, 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

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

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

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

C'est un n -échantillon du modèle statistique

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

◇

Exemple I-1.20 (Regression logistique). Une nouvelle molécule anti-cancéreuse est en cours de développement². 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 $\text{Bin}_{n, \pi}$, $\pi \in]0, 1[$ est la loi sur $\{0, \dots, n\}$ de densité par rapport à la mesure de comptage donnée par

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

L'application

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

est appelée “application logit”. C'est une fonction croissante, $\lim_{\pi \downarrow 0} \text{logit}(\pi) = -\infty$ et $\lim_{\pi \uparrow 1} \text{logit}(\pi) = \infty$. 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

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

où

$$\pi(\theta, x) = \frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)} .$$

On vérifie aisément que

$$\text{logit}(\pi(\theta, x)) = \theta_0 + \theta_1 x .$$

Considérons le modèle statistique

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

Pour $i \in \{1, \dots, n\}$ on note Y_i la i -ème observation qui est la statistique définie, pour tout $z = (y_1, \dots, y_q) \in Z$ par $Y_i(z) = y_i$. Les observations $(Y_1, \dots, Y_n)$ sont indépendantes et le modèle induit par chaque observation est le modèle binomial $\mathcal{E}_i$. 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].

Chapitre I-2

Estimation ponctuelle

Nous étudierons les procédures d'estimation ponctuelle de façon systématique dans la section II-1.6.1. Nous ne donnons ici qu'une introduction rapide de l'estimation, centrée sur quelques méthodes élémentaires classiques, afin de comprendre les enjeux.

Définition I-2.1 (Estimateur ponctuel). Soit $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ 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 $(X_1, \dots, X_n)$ un n -échantillon d'un modèle statistique paramétrique $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$. Pour tout $\theta \in \Theta \subseteq \mathbb{R}^d$, nous notons $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$. Le paramètre $\theta = (\theta^{(1)}, \dots, \theta^{(d)})$ 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 $T_i : X \rightarrow \mathbb{R}$, $i \in \{1, \dots, d\}$ telles que, pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[|T_i(X_1)|] < \infty$. On pose alors, pour tout $\theta \in \Theta$ et tout $i = 1, \dots, d$,

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

Les "moments" $\mathbb{E}_\theta[T_i(X_1)]$ ne sont pas connus, mais on peut les estimer par les moments empiriques $n^{-1} \sum_{j=1}^n T_i(X_j)$. En supposant que $\mathbb{E}_\theta[|T_i^2(X_1)|] < \infty$, l'inégalité de Bienayme-Tchebychev (rappelée au lemme IV-3.2) montre que, pour tout $\varepsilon > 0$ et pour tout i ,

$$\mathbb{P}_\theta \left(\left| n^{-1} \sum_{j=1}^n T_i(X_j) - \mathbb{E}_\theta[T_i(X_1)] \right| \geq \varepsilon \right) \leq \frac{\text{Var}_\theta(T_i(X_1))}{n\varepsilon^2} .$$

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

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta \left(\left| n^{-1} \sum_{j=1}^n T_i(X_j) - \mathbb{E}_\theta[T_i(X_1)] \right| \geq \varepsilon_n \right) = 0 .$$

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

$$n^{-1} \sum_{j=1}^n T_i(X_j) = e_i(\theta) .$$

En supposant que ce système admette une solution unique $\hat{\theta}_n$, on appelle $\hat{\theta}_n$ l'*estimateur des moments* de θ associé aux fonctions T_i , $i = 1 \dots d$. 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

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

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

$$q(\theta, x) = \theta e^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x) . \quad (\text{I-2.1})$$

Considérons les fonctions $T(x) = x$ et $\tilde{T}(x) = x^2$. Chaque fonction est associée à un estimateur des moments de θ . Déterminons tout d'abord les fonctions e et $\tilde{e}$ associées. Un calcul élémentaire montre que

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

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, qui définissent les estimateurs des moments suivants :

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

◇

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

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

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

$$q(\theta, x) = \frac{1}{\pi(1 + (x - \theta)^2)}, \quad x \in \mathbb{R}.$$

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 $T(x) = \text{signe}(x)$, avec

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

On a

$$\mathbb{E}_\theta [T(X_1)] = \int \text{signe}(x) q(\theta, x) dx = 1 - 2F(-\theta) ,$$

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

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

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

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

◇

Exemple I-2.4. Soit $(X_1, \dots, X_n)$ un n -échantillon du modèle $(\mathbf{X}, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ où

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

où q est une densité sur $\mathbb{R}$ vérifiant

$$\int xq(x)dx = 0, \quad \int x^2q(x)dx = m_2 > 0, \quad \int x^4q(x)dx = m_4 < \infty.$$

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 $m_4 = 4! = 24$.

Il est facile de vérifier que pour tout θ

$$\mathbb{E}_\theta[X_1] = \mu, \quad \text{Var}_\theta(X_1) = \sigma^2 m_2.$$

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

$$\mu = \frac{1}{n} \sum_{i=1}^n X_i, \quad (\text{I-2.5})$$

$$\sigma^2 m_2 + \mu^2 = \frac{1}{n} \sum_{i=1}^n X_i^2. \quad (\text{I-2.6})$$

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

$$J_\theta(\mu, \sigma^2) = \begin{bmatrix} 1 & 0 \\ 2\mu & m_2 \end{bmatrix};$$

elle est donc inversible et son inverse est donnée par

$$\left\{ J_\theta(\mu, \sigma^2) \right\}^{-1} = \frac{1}{m_2} \begin{bmatrix} m_2 & 0 \\ -2\mu & 1 \end{bmatrix}.$$

Cette transformation est aussi injective, car

$$\begin{bmatrix} \mu_1 \\ \mu_1^2 + m_2 \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \mu_2 \\ \mu_2^2 + m_2 \sigma_2^2 \end{bmatrix} \Leftrightarrow \begin{bmatrix} \mu_1 \\ \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \mu_2 \\ \sigma_2^2 \end{bmatrix}.$$

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

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

◇

I-2.2 Z-estimateurs

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

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

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

où nous avons défini la fonction

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

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

$$\psi_i : (\theta, x) \mapsto \psi_i(\theta, x), \quad i = 1, \dots, d,$$

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

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

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

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

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 ψ tout estimateur $\hat{\theta}_n$ satisfaisant $\Psi_n(\hat{\theta}_n) = 0$, où

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

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

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{Q}_\theta, \theta \in \Theta = \mathbb{R}\})$$

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 : $\int |x|F(dx) < \infty$. Comme la loi F est symétrique, $\int xF(x)dx = 0$. On a donc

$$\mathbb{E}_\theta[X] = \int xF_\theta(dx) = \int (x + \theta)F(dx) = \theta,$$

ce qui suggère de construire l'estimateur des moments $\hat{\theta}_n$ qui est solution de l'équation

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

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 $\int \psi(x)F(dx) = 0$ où $\psi(x) = \text{signe}(x)$. Un Z-estimateur du paramètre de translation associé à la fonction ψ est une solution de l'équation

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

où la fonction $\text{signe}(\cdot)$ est définie par (I-2.3). En particulier, la *médiane empirique* de l'échantillon, définie par

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

où $X_{m:n}$ est la m -ème statistique d'ordre (voir Définition IV-4.14). La médiane empirique est un Z-estimateur. Ce n'est par contre pas un estimateur de moment. $\diamond$

I-2.3 Maximum de vraisemblance

vraisemblance,
fonction de
vraisemblance,
estimateur
du
maximum de

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

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta =]0, 1[\}) .$$

Pour tout $\theta \in \Theta$, nous notons $\mathbb{P}_\theta = \text{Ber}^{\otimes n}(\theta)$. 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

$$p(\theta, x_1, \dots, x_n) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \prod_{i=1}^n q(\theta, x_i) .$$

Pour une réalisation donnée $(X_1, \dots, X_n)$, on considère la fonction

$$\theta \mapsto \ell_n(\theta) = n^{-1} \sum_{i=1}^n \{X_i \log(\theta) + (1 - X_i) \log(1 - \theta)\} = n^{-1} \sum_{i=1}^n \log q(\theta, X_i) .$$

Cette fonction est appelée *log-vraisemblance*. Pour $\theta_0 \in \Theta$, nous avons $\mathbb{E}_{\theta_0}[X_i] = \theta_0$ pour tout $i \in \{1, \dots, n\}$ et donc, pour tout $\theta, \theta_0 \in \Theta$,

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

La dérivée seconde de la fonction $\theta \mapsto \mathbb{E}_{\theta_0}[\log q(\theta, X_1)]$ est donnée, pour tout $\theta, \theta_0 \in \Theta$, 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 Θ , elle admet donc un maximum unique au point θ_0 . En utilisant l'inégalité de Bienayme-Tchebychev (Lemme IV-3.2), pour tout $\theta, \theta_0 \in \Theta$ et $\delta > 0$,

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

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 $\theta, \theta_0 \in \Theta$, il vient

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

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 $\lim_{\theta \rightarrow 0} \ell_n(\theta) = -\infty$ et $\lim_{\theta \rightarrow 1} \ell_n(\theta) = -\infty$. 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 $T(x) = x$ et à la fonction $e(\theta) = \theta = \mathbb{E}_\theta[T(X_1)]$. Cette propriété sera vérifiée dans les modèles statistique de type exponentiel (voir Chapitre IV-6). $\diamond$

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

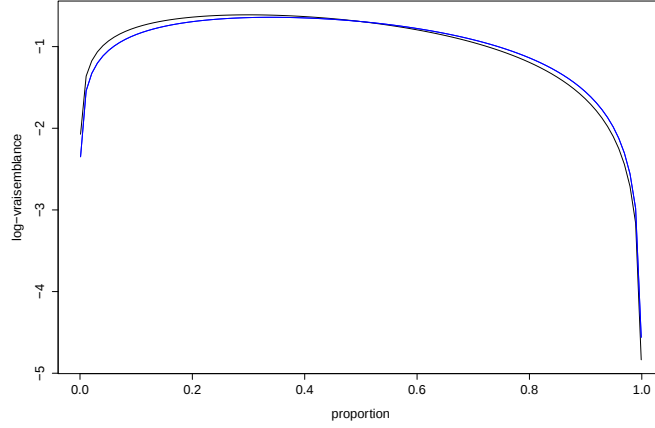


FIGURE I-2.1 – Log-vraisemblance et sa limite

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

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

$$\theta \in \Theta \rightarrow L_n(\theta, X_1, \dots, X_n) = \prod_{i=1}^n q(\theta, X_i), \quad q(\theta, x) = q_\theta(x).$$

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

$$L_n(\theta) = L_n(\theta, X_1, \dots, X_n).$$

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

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

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

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

La vraisemblance admet un maximum unique en

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

◇

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

$$X_i = \min\{Y_i, \tau\}, \quad i = 1, \dots, n$$

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

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\text{Expo}(\theta), \theta \in \Theta = \mathbb{R}_+^*\})$$

Notons que les variables $(Y_1, \dots, Y_n)$ ne sont pas observées. Posons $\mu = \lambda_{\text{Leb}} + \delta_\tau$, 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

$$q(\theta, x) = \theta e^{-\theta x} \mathbb{1}_{\{x < \tau\}} + c(\theta) \mathbb{1}_{\{x = \tau\}},$$

avec $c(\theta) = \int_\tau^{+\infty} \theta e^{-\theta t} dt = e^{-\theta \tau}$: $(X_1, \dots, X_n)$ est un n -échantillon du modèle

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

La vraisemblance s'écrit

$$L_n(\theta, X_1, \dots, X_n) = \theta^{\text{card } N_n^-} \exp \left(-\theta \sum_{i \in N_n^-} X_i \right) c(\theta)^{\text{card } N_n^+},$$

où $N_n^- = \{i \leq n : X_i < \tau\}$ et $N_n^+ = \{i \leq n : X_i = \tau\}$. 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

$$L_n(\theta, Y_1, \dots, Y_n) = \theta^n \exp \left(-\theta \sum_{i=1}^n Y_i \right).$$

◇

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

$$L_n(\hat{\theta}_n^{\text{MV}}, X_1, \dots, X_n) = \max_{\theta \in \Theta} L_n(\theta, X_1, \dots, X_n),$$

autrement dit

$$\hat{\theta}_n^{\text{MV}} \in \arg \max_{\theta \in \Theta} L_n(\theta, X_1, \dots, X_n). \quad (\text{I-2.14})$$

L'application

$$\theta \in \Theta \rightarrow \ell_n(\theta, X_1, \dots, X_n) = \frac{1}{n} \log L_n(\theta, X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n \log q(\theta, X_i), \quad (\text{I-2.15})$$

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

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

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

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

Ceci fournit un système de d équations si $\Theta \subset \mathbb{R}^d$ avec $d \geq 1$.

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 $\theta \rightarrow \ell_n(\theta)$, en particulier, tous ses maxima et minima locaux. On appelle racine de l'équation de vraisemblance tout (estimateur) $\hat{\theta}_n^{\text{rv}}$ solution de (I-2.16), c'est-à-dire tel que

$$\nabla_\theta \ell_n(\hat{\theta}_n^{\text{rv}}, X_1, \dots, X_n) = 0.$$

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

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

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

$$q(\theta, x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}.$$

La log-vraisemblance associée s'écrit

$$\ell_n((\mu, \sigma^2), X_1, \dots, X_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

Les équations de vraisemblance s'écrivent

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

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

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

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

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

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

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)

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

La vraisemblance s'écrit alors,

$$L_n(\theta, X_1, \dots, X_n) = \frac{1}{\pi^n} \prod_{i=1}^n \frac{1}{1+(X_i-\theta)^2}.$$

La log-vraisemblance est donnée par

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

dont l'équation de vraisemblance associée est :

$$\ell'_n(\theta, X_1, \dots, X_n) = \sum_{i=1}^n \frac{X_i - \theta}{1+(X_i-\theta)^2} = 0. \quad (\text{I-2.17})$$

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. La valeur à la $k+1$ -ième itération est donnée par

$$\theta^{(k+1)} = \theta^{(k)} + \gamma \ell'_n(\theta^{(k)}), \quad \diamond$$

où γ est un pas d'adaptation. Nous avons représenté Figure I-2.2 la log-vraisemblance et sa dérivée pour $n = 50$ observations d'une loi de Cauchy de paramètre $\theta = 2$.

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.

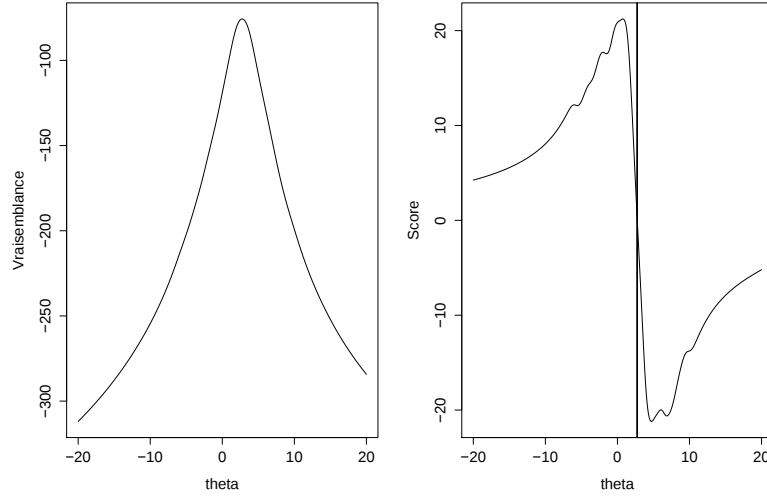


FIGURE I-2.2 – Vraisemblance et score pour $n = 50$ observations d'un modèle de Cauchy de paramètre $\theta = 2$

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

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\text{Unif}([0, \theta])\}, \theta \in \Theta = \mathbb{R}_+ \setminus \{0\})$$

où pour $\theta \in \Theta$, $\text{Unif}([0, \theta])$ 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

$$L_n(\theta, X_1, \dots, X_n) = \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{[0, \theta]}(X_i) = \theta^{-n} \mathbb{1}_{[0, \theta]}(X_{n:n}),$$

où $X_{n:n} = \max_{i=1, \dots, n} X_i$ (voir Définition IV-4.14). La valeur maximale de $L_n(\theta, X_1, \dots, X_n)$ est obtenue pour $\theta = X_{n:n}$ et donc $\hat{\theta}_n^{\text{MV}} = X_{n:n}$. Par contre, la vraisemblance n'est pas dérivable et la fonction de log-vraisemblance n'est pas définie pour toutes les valeurs de $\theta \in \Theta = \mathbb{R}^* \setminus \{0\}$. $\diamond$

L'estimateur du maximum de vraisemblance n'est pas toujours défini.

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

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

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

$$L_n(\theta, X_1, \dots, X_n) = \prod_{i=1}^n q_0(X_i - \theta).$$

On a $\lim_{\theta \rightarrow X_i} L_n(\theta, X_1, \dots, X_n) = +\infty$ pour tout $i = 1, \dots, n$. Pour cette expérience statistique, l'estimateur du maximum de vraisemblance n'est pas défini. $\diamond$

Il n'y a pas nécessairement unicité de l'estimateur du maximum de vraisemblance.

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

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{q_\theta \cdot \lambda_{\text{Leb}}\}, \theta \in \Theta = \mathbb{R})$$

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

$$q(\theta, x) = \frac{1}{2} \exp(-|x - \theta|).$$

La log-vraisemblance est donnée par

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

Maximiser $L_n(\theta, X_1, \dots, X_n)$ revient à minimiser la fonction $\theta \rightarrow \sum_{i=1}^n |X_i - \theta|$. Cette fonction est dérivable presque partout, de dérivée $-\sum_{i=1}^n \text{sign}(X_i - \theta)$. La dérivée (définie presque partout) est constante par morceaux. Si n est impair, elle s'annule en un point unique $X_{1:(n+1)/2}$ (sous réserve que les X_i soient deux-à-deux distincts, ce qui est vrai presque sûrement car la loi de $(X_1, \dots, X_n)$ est à densité par rapport à la mesure de Lebesgue), où $X_{1:n} \leq \dots \leq X_{n:n}$ désigne la statistique d'ordre associée à l'échantillon (voir Définition IV-4.14). Si n est pair, il y a une infinité de solutions : tout point de l'intervalle $(X_{1:n/2}, X_{1:n/2+1})$ est un estimateur du maximum de vraisemblance. $\diamond$

I-2.3.1 Invariance par reparamétrisation

L'estimateur du maximum de vraisemblance est invariant par changement de paramétrisation. Cela signifie que, si $\hat{\theta}_n^{\text{MV}}$ est l'estimateur du maximum de vraisemblance pour θ , alors $G(\hat{\theta}_n^{\text{MV}})$ 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 $\{\mathbb{P}_\theta, \theta \in \Theta\}$ et une bijection

$$G : \Theta \rightarrow G(\Theta)$$

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

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

Proposition I-2.18. *Si $G : \Theta \rightarrow \tilde{\Theta}$ est une bijection et si $\hat{\theta}_n^{\text{MV}}$ désigne l'estimateur du maximum de vraisemblance pour l'expérience statistique associée à la famille de lois $\{\mathbb{P}_\theta, \theta \in \Theta\}$, alors $G(\hat{\theta}_n^{\text{MV}})$ est l'estimateur du maximum de vraisemblance de $G(\theta)$, c'est-à-dire pour l'expérience statistique associée à la famille de lois $\{\mathbb{P}_{G^{-1}(\tau)}, \tau \in G(\Theta)\} = \{\mathbb{Q}_\tau, \tau \in G(\Theta)\}$.*

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

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

On a donc

$$\widetilde{L}_n(\tau, X_1, \dots, X_n) = L_n(\theta, X_1, \dots, X_n) \leq L_n(\hat{\theta}_n^{\text{MV}}, X_1, \dots, X_n) = \widetilde{L}_n(\hat{\tau}_n, X_1, \dots, X_n). \quad \square$$

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

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\text{Expo}(\theta), \theta \in \Theta = \mathbb{R}_+^*\})$$

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

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

La log-vraisemblance s'écrit

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

donc $\ell'_n(\theta) = 0$ si et seulement si $\theta = 1/\bar{X}_n$ où $\bar{X}_n$ est la moyenne empirique. On vérifie que c'est un maximum global, donc $\hat{\theta}_n^{\text{MV}} = \frac{1}{\bar{X}_n}$. 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 $\tau = 1/\theta$, $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ est $\hat{\tau}_n = \bar{X}_n$. $\diamond$

Exemple I-2.20 (Régression linéaire). Nous avons défini la vraisemblance à des observations indépendantes et identiquement distribuées. 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 ??

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{p_\theta \dots \lambda_{\text{Leb}}[n], \theta \in \Theta\})$$

où

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

$$f(\beta, \mathbf{x}) = \sum_{\ell=0}^p \beta_\ell \varphi_\ell(\mathbf{x}),$$

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

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

L'espace des observations est ici $Z = \mathbb{R}^n$ et $\mathcal{Z} = \mathcal{B}(\mathbb{R}^n)$. Le modèle est dominé par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ et la densité conjointe des observations est donnée par

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

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

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

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

On obtient ainsi un système d'équations linéaires

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

En notant Φ la matrice de régression $n \times (p+1)$ définie par

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

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

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

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

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

Cet estimateur des paramètres de régression β est appelé *estimateur des moindres carrés*. Pour estimer la variance σ^2 , nous pouvons par exemple considérer

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

◇

\$M\$-
estimateur
contraste,
estimateur
de

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 $(X_1, \dots, X_n)$ un n -échantillon d'un modèle statistique $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$. Nous posons pour tout $\theta \in \Theta$, $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$. Considérons une fonction $m : \Theta \times X \rightarrow \mathbb{R}$, $(\theta, x) \mapsto m(\theta, x)$. Supposons 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 $M_{\theta_0} : \Theta \rightarrow \mathbb{R}$, définie par

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

Supposons 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 $(X_1, \dots, X_n)$ et la fonction M_{θ_0} n'est donc pas connue. Par contre, nous pouvons l'estimer par une moyenne empirique. Soit $M_n : \Theta \rightarrow \mathbb{R}$ 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 $\delta > 0$,

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

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

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} M_n(\theta) . \quad (\text{I-2.20})$$

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 pour tout $x \in X$ la fonction $\theta \rightarrow m_\theta(x) = m(\theta, x)$ est différentiable, en posant

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

et si l'ensemble des paramètres Θ est ouvert, nous avons

$$\frac{1}{n} \sum_{i=1}^n \nabla m(\hat{\theta}_n, X_i) = \frac{1}{n} \sum_{i=1}^n \psi(\hat{\theta}_n, X_i) = 0 \quad (\text{I-2.21})$$

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'éq. (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, $\sigma^2 = \int x^2 F(dx) < \infty$. 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 est donc la valeur du paramètre $\hat{\theta}_n$ qui maximise la fonction

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

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

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

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, car elle est solution du problème d'optimisation

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

◇

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 $(X, \mathcal{X})$. 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 IV-1.48).

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

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

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

Remarque I-2.23. Nous montrerons (Théorème I-2.24) que la définition de la divergence de Kullback-Leibler $\text{KL}(\mathbb{P}_0, \mathbb{P}_1)$ 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 :

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

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

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

Notons que $\text{KL}(\mathbb{P}_0, \mathbb{P}_1)$ 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-3.4).

Théorème I-2.24. Soient $(X, \mathcal{X})$ un espace mesurable et $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités.

- (i) $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) < +\infty$ implique $\mathbb{P}_0 \ll \mathbb{P}_1$.
- (ii) $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) \in [0, \infty]$ et $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$ si et seulement si $\mathbb{P}_0 = \mathbb{P}_1$.
- (iii) $\text{KL}(\mathbb{P}_0, \mathbb{P}_1)$ ne dépend pas du choix de la mesure de domination.

Démonstration. (i) Remarquons en effet que, s'il existe $A \in \mathcal{X}$ 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

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

Ceci montre en fait que $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) < \infty$ 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 $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$. Il nous suffit donc de montrer $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) \geq 0$ et $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$ implique $p_0(x) = p_1(x)$ $\mathbb{P}_0$ -p.s. On applique l'inégalité de Jensen (Lemme IV-3.4) à la variable aléatoire p_1/p_0 , et on obtient, par convexité de $-\log$,

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

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 IV-1.48, 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,

$$\text{KL}(\mathbb{P}_0, \mathbb{P}_1) = \int \tilde{f}_0(x) \log(\tilde{f}_0(x)) \mathbb{P}_1(dx), \quad (\text{I-2.26})$$

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 $\mathbb{P}_0 = \tilde{f}_0 p_1 \cdot \mu$. Comme la densité est unique à une μ -équivalence près par le théorème IV-1.48, nous avons $p_0 = \tilde{f}_0 p_1$ μ -presque partout. Il s'en suit

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

L'assertion (i) et (I-2.26) montre que $\text{KL}(\mathbb{P}_0, \mathbb{P}_1)$ 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 $\text{KL}(\mathbb{P}_0, \mathbb{P}_1) = \infty$. $\square$

Nous allons maintenant montrer que l'estimateur du maximum de vraisemblance est un M -estimateur. Soit $(X_1, \dots, X_n)$ un n -échantillon d'un modèle statistique paramétrique dominé $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$. 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 $(\theta, x) \in \Theta \times X$, $q(\theta, x) > 0$ et posons

$$m(\theta, x) = \log q(\theta, x). \quad (\text{I-2.27})$$

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 $M_{\theta_0} : \theta \rightarrow \mathbb{R}$ définie par

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

Notons que

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

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

$$\text{Ent}(\mathbb{Q}_{\theta_0}) = \int q(\theta_0, x) \log q(\theta_0, x) \mu(dx) . \quad (\text{I-2.28})$$

Par Théorème I-2.24, la fonction $\theta \mapsto \text{KL}(\mathbb{Q}_{\theta_0}, \mathbb{Q}_{\theta})$ est finie et positive. Nous avons $\text{KL}(\mathbb{Q}_{\theta_0}, \mathbb{Q}_{\theta}) = 0$ si et seulement si $\mathbb{Q}_{\theta_0} = \mathbb{P}_{\theta}$ ce qui équivaut 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

I-3.1 Tests d'hypothèse

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. Commençons par un exemple élémentaire.

Exemple I-3.1. Considérons le modèle statistique $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\theta, 1), \theta \in \Theta\})$ où $\Theta \subset \mathbb{R}$. On cherche à déterminer si θ appartient à un ensemble $\Theta_0 \cup \Theta_1$, où Θ_0 et Θ_1 sont deux ensembles disjoints de $\Theta = \mathbb{R}$. On veut savoir si $\theta \in \Theta_0$ ou $\theta \in \Theta_1$. On appelle Θ_0 l'hypothèse (ou hypothèse nulle) et Θ_1 l'alternative (ou hypothèse alternative). $\diamond$

Plus généralement, étant donnée une expérience statistique $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ 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*

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

Pour effectuer un test, une première idée consiste à construire une statistique prenant deux valeurs 0 ou 1.

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 $\phi : Z \rightarrow \{0, 1\}$ 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

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

est la région de rejet du test. L'ensemble $\mathcal{A} = Z \setminus \mathcal{R}$ est la région d'acceptation du test.

Exemple I-3.3. Reprenons Exemple I-3.1 et supposons que $\Theta_0 = \{0\}$ et $\Theta_1 = \{\vartheta\}$ où $0 < \vartheta$. Une première idée est de rejeter l'hypothèse nulle H_0 si Z est plus proche de ϑ que de 0, c'est à dire si $Z > \vartheta/2$. La région de rejet du test est alors définie par

$$\mathcal{R} = \{z \in \mathbb{R} : z > \vartheta/2\}$$

et la région d'acceptation est

$$\mathcal{A} = \{z \in \mathbb{R} : z \leq \vartheta/2\} .$$

$\diamond$

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 simples.

Exemple I-3.4 (Sondage). Considérons l'exemple I-1.1. Supposons que nous cherchions à tester $\theta \leq \theta_0 = 1/2$ où θ est la probabilité de voter pour le candidat B . De façon plus formelle, soit $(X_1, \dots, X_n)$ un n -échantillon d'une expérience statistique de Bernoulli, $(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta =]0, 1[\})$. Pour $\theta_0 \in \Theta$, considérons la partition de l'espace des paramètres $\Theta_0 = \{\theta \in \Theta : \theta \leq \theta_0\}$ et $\Theta_1 = \{\theta \in \Theta : \theta > \theta_0\}$. Contrairement aux problèmes d'estimation, le problème n'est pas d'approcher la valeur numérique de θ mais de répondre à la question $\theta \leq \theta_0$ en étant capable d'évaluer la "fiabilité" de la réponse.

Pour $t \geq 0$, il est naturel de considérer un test de la forme suivante

$$T(X_1, \dots, X_n) = \mathbb{1} \left\{ \sum_{i=1}^n X_i \geq t \right\}.$$

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

$$\mathcal{R} = \left\{ (x_1, \dots, x_n) \in \{0, 1\}^n : \sum_{i=1}^n x_i \geq t \right\}.$$

◇

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

- $H_0 : \theta = \theta_0$, contre $H_1 : \theta = \theta_1$, comme dans l'exemple I-3.3. 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 $H_1 : \theta > \theta_0$, qui est un test d'une hypothèse composite contre une alternative composite.

Il est fréquent que la région de rejet $\mathcal{R}$ se mette sous la forme $\mathcal{R} = \{z \in Z : T(z) > c\}$, pour une certaine statistique $T : Z \rightarrow \mathbb{R}$ et un réel c . La variable aléatoire T est alors appelée la *statistique de test* et c la *valeur critique*. 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 , $T(Z) > c$.

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

Définition I-3.5 (Fonction puissance). Soit $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique paramétrique. Notons par Z l'observation canonique. La fonction puissance d'un test pur ϕ est définie par

$$\beta_\phi : \theta \in \Theta \mapsto \mathbb{P}_\theta(\phi(Z) = 1) = \mathbb{P}_\theta(\mathcal{R}) \in [0, 1].$$

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 1ère espèce* : rejeter H_0 alors que cette hypothèse est satisfaite.
- *Erreur de 2ème 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). La taille du test ϕ (ou risque de première espèce), est définie par

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

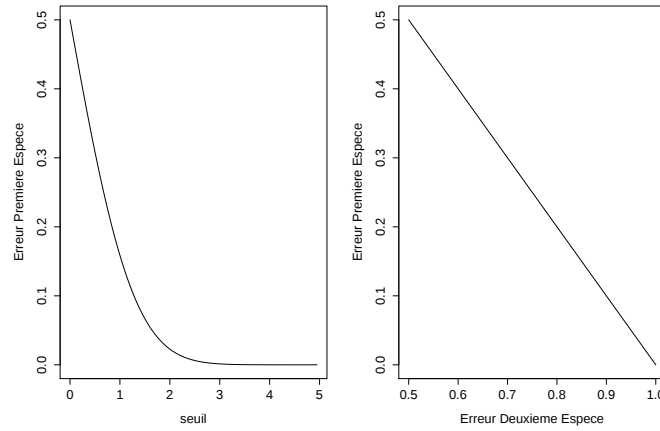


FIGURE I-3.1 – Figure de gauche : erreur de première espèce en fonction du seuil. Figure de droite : erreur de première espèce en fonction de l'erreur de seconde espèce

Le risque de première espèce est la probabilité (maximale) de rejeter H_0 alors qu'elle était vraie.

Il est naturel de se poser la question du choix optimal de la région critique. Il est clair que si ϕ_1 et ϕ_2 sont deux tests et si

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

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 $\Theta_0 = \{\theta \leq 0\}$ et $\Theta_1 = \{\theta > 0\}$. Donnons nous deux réels $c_1 > c_2$ et définissons deux tests $\phi_1(Z) = \mathbb{1}_{\{Z > c_1\}}$ et $\phi_2(Z) = \mathbb{1}_{\{Z > c_2\}}$; la statistique de test est ici $T(Z) = Z$ et c_1 et c_2 sont deux seuils. Les régions de rejet de ces tests sont données par $\mathcal{R}_1 := \{z \in \mathbb{R} : Z > c_1\}$, $\mathcal{R}_2 := \{z \in \mathbb{R} : Z > c_2\}$. Il est clair que $\mathcal{R}_1 \subset \mathcal{R}_2$, ce qui implique que, pour tout $\theta \in \Theta$,

$$\beta_{\phi_1}(\theta) = \mathbb{P}_{\theta}(Z \in \mathcal{R}_1) < \beta_{\phi_2}(\theta) = \mathbb{P}_{\theta}(Z \in \mathcal{R}_2),$$

Plus précisément,

$$\mathbb{P}_{\theta}(Z \in \mathcal{R}_1) = \mathbb{P}_{\theta}(Z > c_1) = 1 - \Phi(c_1 - \theta).$$

et donc

$$\sup_{\theta \in \Theta_0} \mathbb{P}_{\theta}(Z \in \mathcal{R}_1) = 1 - \inf_{\theta \leq 0} \Phi(c_1 - \theta) = 1 - \Phi(c_1).$$

La taille du test ϕ_1 est égale à $1 - \Phi(c_1)$. Si ϕ_1 a un risque de première espèce $1 - \Phi(c_1)$ strictement inférieur au risque de première espèce $1 - \Phi(c_2)$ de ϕ_2 . Par contre, pour tout $\theta \in \Theta_1$, nous avons $\beta_{\phi_1}(\theta) < \beta_{\phi_2}(\theta)$.

Considérons maintenant le test d'hypothèse simple exemple I-3.3

$$H_0 : \theta = 0, \quad \text{contre} \quad H_1 : \theta = \vartheta$$

Ici encore, prenons comme statistique de test $T(Z) = Z$ et un seuil critique c . La taille du test est donnée par

$$\mathbb{P}_0(Z > c) = 1 - \Phi(c).$$

Nous avons représenté dans la fig. I-3.2 le risque de première espèce en fonction du seuil c du test. ◇

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 les erreurs de première et seconde espèce. 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

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

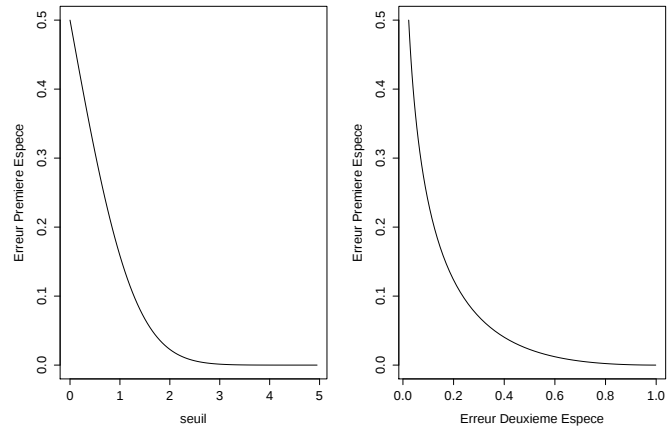


FIGURE I-3.2 – Figure de gauche : erreur de première espèce en fonction du seuil. Figure de droite : erreur de première espèce en fonction de l’erreur de seconde espèce pour $\vartheta = 2$

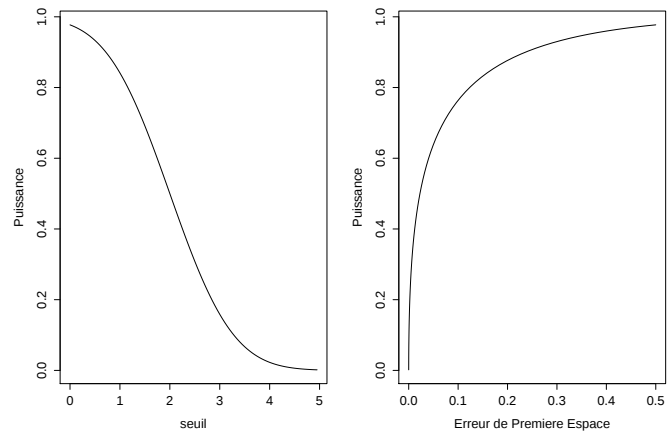


FIGURE I-3.3 – Figure de gauche : erreur de première espèce en fonction du seuil. Figure de droite : erreur de première espèce en fonction de l’erreur de seconde espèce pour $\vartheta = 2$

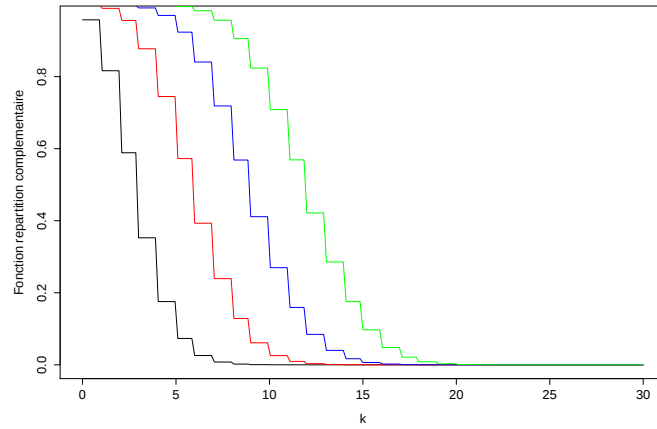


FIGURE I-3.4 – Fonction de répartition complémentaire de la loi binômiale pour $n = 30$ et $\theta = 0.1, 0.2, 0.3, 0.4$

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

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

Ceci est illustré dans fig. I-3.4. 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,

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

Ainsi, pour tout c , nous avons

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

Pour un seuil c donné, nous pouvons maintenant aisément calculer les erreurs de première et de deuxième espèce. En effet, l'erreur de première espèce est donnée par

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

De même, l'erreur de deuxième espèce est donnée par

$$\inf_{\theta \in \Theta_1} \{1 - \mathbb{P}_\theta(A_n \leq c)\} = \mathbb{P}_{\theta_0}(A_n \leq c).$$

Nous avons visualisé dans la fig. I-3.5 l'évolution de l'erreur de première espèce en fonction du seuil du test et l'erreur de première espèce en fonction de l'erreur de deuxième espèce.

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

$$\mathbb{P}_\theta(A_n > c_\alpha) \geq \mathbb{P}_{\theta_0}(A_n > c_\alpha) \geq \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(A_n > c_\alpha).$$

◇

De manière plus générale, la minimisation du risque de première espèce nous conduit à choisir la région de rejet la plus petite possible. À l'inverse, la maximisation de la puissance du test nous conduit à choisir la région de rejet la plus grande possible. La détermination de la zone 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, comme pour les régions de confiance, un réel $\alpha \in]0, 1[$ typiquement petit, de l'ordre de 5%, 1/1000 ou 1/10⁶ selon les applications et on construit le test de façon à garantir

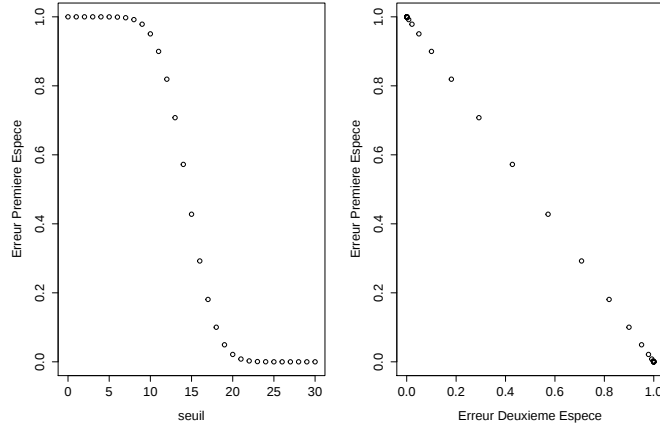


FIGURE I-3.5 – Figure de gauche : Erreur de première espèce en fonction du seuil ; Figure de droite : Erreur de première espèce en fonction de l’erreur de deuxième espèce

que sa taille $\overline{\alpha}(\phi)$ soit inférieure à α . On dit alors que le test est de *niveau* α . Ayant borné le risque de première espèce par α , il est naturel de chercher à maximiser la puissance du test $\beta_\phi(\theta)$ pour tout $\theta \in \Theta_1$.

Exemple I-3.9. Dans l’exemple I-3.3, on garantit un niveau α si on choisit de rejeter H_0 lorsque $Z > z_{1-\alpha}$, où $z_{1-\alpha}$ est le $(1 - \alpha)$ -quantile de la loi $N(0, 1)$. En effet, la région de rejet du test est alors $\mathcal{R} = \{z \in \mathbb{R} : z > z_{1-\alpha}\}$ et on a bien

$$\mathbb{P}_0(Z \in \mathcal{R}) = \mathbb{P}_0(Z > z_{1-\alpha}) = \alpha .$$

Imaginons maintenant qu’on permute les hypothèses H_0 et H_1 , c’est à dire qu’on fasse le test

$$H_0 : \theta = \vartheta, \quad \text{contre} \quad H_1 : \theta = 0 .$$

Un calcul élémentaire nous montre qu’alors, on obtient un test de niveau α en choisissant pour région de rejet $\mathcal{R}' = \{z \in \mathbb{R} : z < \vartheta - z_{1-\alpha}\}$. Pour comprendre la dissymétrie entre les hypothèses H_0 et H_1 , prenons $0 < \vartheta < z_{1-\alpha}$. Dans ce cas, on peut découper l’ensemble des valeurs prises par Z en trois régions.

- $\mathcal{R}_0 = \{Z < \vartheta - z_{1-\alpha}\}$ est la région dans laquelle les deux tests choisissent l’hypothèse H_0 ,
- $\mathcal{R}_1 = \{Z > z_{1-\alpha}\}$ est la région dans laquelle les deux tests choisissent l’hypothèse H_1 ,
- $\mathcal{R}_2 = \{\vartheta - z_{1-\alpha} \leq Z \leq z_{1-\alpha}\}$ est la région dans laquelle la région choisie dépend de l’hypothèse nulle.

En fait ces 3 régions correspondent à la situation suivante. Si on cherche un intervalle de confiance $\hat{I}_\alpha$ de niveau $1 - \alpha$ pour θ en sachant que $\theta \in \{0, \vartheta\}$, on vérifie que,

- lorsque $Z \in \mathcal{R}_0$, l’intervalle de confiance ne contient que 0,
- lorsque $Z \in \mathcal{R}_1$, $\hat{I}_\alpha$ ne contient que ϑ
- lorsque $Z \in \mathcal{R}_2$, $\hat{I}_\alpha$ contient 0 et ϑ .

Ainsi, la troisième région décrit une expérience dans laquelle le point de vue des intervalles de confiance consiste à répondre que les deux paramètres sont acceptables, autrement dit, les données ne permettent pas de trancher entre $\theta = 0$ et $\theta = \vartheta$. Le point de vue de Neyman-Pearson consiste à décider H_0 dans ce cas litigieux. $\diamond$

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. La p -valeur, ou valeur critique permet de quantifier plus précisément le niveau d'incertitude de l'hypothèse.

Définition I-3.10 (p -valeur). Soit $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ 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 ,

$$\bar{\alpha}(\phi_\alpha) = \sup_{\theta \in \Theta_0} \beta_{\phi_\alpha}(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi_\alpha(Z) = 1) \leq \alpha,$$

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

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

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

Ainsi, étant 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-3.11. Nous avons déjà vu qu'on obtient un test de niveau α dans l'exemple I-3.3 lorsqu'on choisit pour région de rejet $\mathcal{R}_\alpha = \{Z > z_{1-\alpha}\}$. Ainsi, ayant observé Z , on va rejeter H_0 pour tout α tel que $z_{1-\alpha} < Z$ et accepter H_0 pour tout α tel que $z_{1-\alpha} \geq Z$. Pour $0 \leq \alpha < \alpha' \leq 1$, nous avons $z_{1-\alpha} \geq z_{1-\alpha'}$ et donc $\mathcal{R}_\alpha \subset \mathcal{R}_{\alpha'}$. 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

$$Z = z_{1-\hat{\alpha}(Z)} \Rightarrow \Phi(Z) = 1 - \hat{\alpha}(Z)$$

où Φ est la fonction de répartition de la loi normale centrée réduite. Par conséquent

$$\hat{\alpha}(Z) = 1 - \Phi(Z),$$

Remarquons que, sous l'hypothèse de base H_0 , la distribution de $\hat{\alpha}(Z)$ est donnée par

$$\mathbb{P}_0(\hat{\alpha}(Z) \leq u) = \mathbb{P}_0(1 - \Phi(Z) \leq u) = \mathbb{P}_0(\Phi(Z) \geq 1 - u) = u,$$

car $\Phi(Z)$ est, sous l'hypothèse nulle, distribuée suivant une loi uniforme sur $[0, 1]$ (voir proposition IV-4.9); donc, la p -valeur $1 - \Phi(Z)$ est uniformément distribuée sur $[0, 1]$. $\diamond$

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

Proposition I-3.12. Considérons $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique paramétrique. Considérons le test

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

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.

$$\bar{\alpha}(\phi_\alpha) = \sup_{\theta \in \Theta_0} \beta_{\phi_\alpha}(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi_\alpha(Z) = 1) \leq \alpha,$$

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

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

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

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

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

alors

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

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 $\{Z \in \mathcal{R}_v\}$ pour tout $u < v$ ce qui implique

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

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

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

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

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

La proposition suivante généralise l'exemple I-3.11 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). Supposons que, pour tout $\alpha \in [0, 1]$, $\phi_\alpha(Z) = \mathbb{1}_{\{T(Z) \geq c_\alpha\}}$ et pour tout $\alpha \in]0, 1]$, $\bar{\alpha}(\phi_\alpha) = \alpha$. Supposons de plus que la fonction $c_\alpha : \alpha \mapsto c_\alpha$ soit décroissante sur $[0, 1]$. Alors

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

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

Exemple I-3.14. Soient $(X_1, \dots, X_n)$ un n -échantillon du modèle gaussien

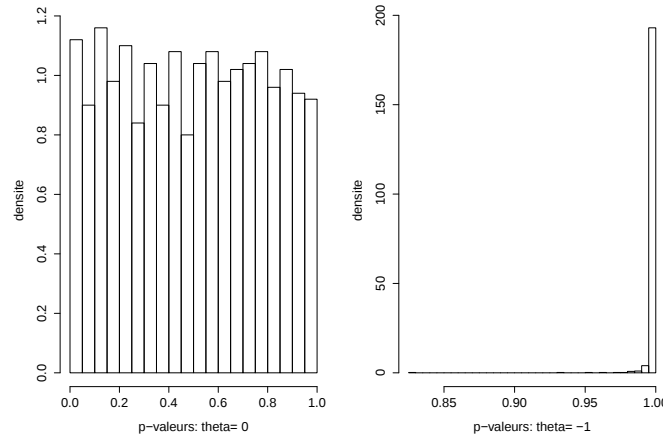
$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\theta, 1), \theta \in \mathbb{R}\}).$$

Nous notons pour tout $\theta \in \Theta$, $\mathbb{P}_\theta = N^{\otimes n}(\theta, 1)$. On cherche à tester l'hypothèse $H_0 = \theta \in \Theta_0$ où $\Theta_0 = \mathbb{R}_-$. Pour tout $\theta \in \Theta$, la moyenne empirique $\bar{X}_n$ est distribuée suivant une loi gaussienne $N(\theta, n^{-1})$. Pour tout $\alpha \in [0, 1]$, considérons le test

$$\phi_\alpha(X_1, \dots, X_n) = \mathbb{1}_{\{\sqrt{n}\bar{X}_n \leq z_{1-\alpha}\}}. \quad (\text{I-3.4})$$

Pour tout $\theta \in \Theta_0$, nous avons

$$\begin{aligned} \mathbb{P}_\theta(\sqrt{n}\bar{X}_n \geq z_{1-\alpha}) &= \mathbb{P}_\theta(\sqrt{n}\{\bar{X}_n - \theta\} \geq z_{1-\alpha} - \sqrt{n}\theta) \\ &= 1 - \Phi(z_{1-\alpha} - \sqrt{n}\theta) \leq 1 - \Phi(z_{1-\alpha}) = \alpha, \end{aligned}$$

FIGURE I-3.6 – p -valeurs de la famille de tests (I-3.4)

ce qui montre que pour tout $\alpha \in]0, 1[$,

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\sqrt{n}\bar{X}_n \leq z_{1-\alpha}) = \alpha.$$

Pour $0 \leq \alpha \leq \alpha' \leq 1$, nous avons $z_{1-\alpha} \geq z_{1-\alpha'}$ et donc $\mathcal{R}_\alpha \subset \mathcal{R}_{\alpha'}$. La proposition I-3.13 montre que

$$\begin{aligned} \hat{\alpha}(X_1, \dots, X_n) &= \inf \{ \alpha \in]0, 1[: \sqrt{n}\bar{X}_n \geq z_{1-\alpha} \} \\ &= \inf \{ \alpha \in]0, 1[: \Phi(\sqrt{n}\bar{X}_n) \geq 1 - \alpha \} = 1 - \Phi(\sqrt{n}\bar{X}_n). \end{aligned} \quad \diamond$$

Nous avons visualisé (voir exemple I-3.14) pour 1000 expériences indépendantes les p -valeurs calculées pour $n = 20$ et $\theta = 0$ et $\theta = -1$.

I-3.3 Régions de confiance

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 $T(X_1, \dots, X_n)$ 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 $T : Z \rightarrow \mathbb{R}^p$ est abandonnée en faveur d’une *région d’estimation*, c’est-à-dire une fonction $\mathcal{C}$, définie sur l’espace des observations Z , à valeurs dans l’ensemble des parties de $\mathbb{R}^p$. Pour des raisons de mesurabilité, nous supposons que, pour tout $\theta \in \Theta$,

$$\{z \in Z : g(\theta) \in \mathcal{C}(z)\} \in \mathcal{Z}.$$

Étant donnée l’observation de Z , la partie $\mathcal{C}(Z)$ doit être interprétée comme la “région” à laquelle la fonction du paramètre $g(\theta)$ appartient. Pour que $\mathcal{C}(Z)$ puisse être appelée région de confiance, on doit être capable de montrer une borne inférieure sur la probabilité que $g(\theta) \in \mathcal{C}(Z)$ (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 $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique, $g : \Theta \rightarrow \mathbb{R}^p$ une fonction et $\alpha \in]0, 1[$. Une fonction $\mathcal{C} : Z \rightarrow \mathcal{B}(\mathbb{R}^p)$ est une *région de confiance* au niveau $1 - \alpha$ pour $g(\theta)$ si, pour tout $\theta \in \Theta$, $\{z \in Z : g(\theta) \in \mathcal{C}(z)\} \in \mathcal{Z}$ et

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta(\{z \in Z : g(\theta) \in \mathcal{C}(z)\}) \geq 1 - \alpha.$$

Borne de
confiance

Intervalle de
confiance bilatéral
confiance,
intervalle de

Si $\mathcal{C}(Z)$ 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 $\mathcal{C}(Z)$ ". 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 (une statistique à valeurs ensemble) qui sous $\mathbb{P}_\theta$ a une probabilité $1 - \alpha$ 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 très grand nombre de fois (pour la même valeur de $\theta \in \Theta$), la proportion des régions de confiance obtenues qui contiendraient la valeur $g(\theta)$ du paramètre d'intérêt serait d'au moins $1 - \alpha$.

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

- (i) $\mathcal{C}(Z) = [m(Z), \infty[$: $m(Z)$ est une *borne inférieure de confiance*.
- (ii) $\mathcal{C}(Z) =]-\infty, M(Z)]$: $M(Z)$ est une *borne supérieure de confiance*.
- (iii) $\mathcal{C}(Z) = [m(Z), M(Z)]$ 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 $(X_1, \dots, X_n)$ un n -échantillon du modèle de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta = [0, 1]\})$$

Pour tout $\theta \in \Theta$, nous posons $\mathbb{P}_\theta = \text{Ber}^{\otimes n}(\theta)$. Si $\hat{\theta}_n = n^{-1} \sum_{i=1}^n X_i$ est le nombre moyen de succès, nous avons pour tout $\theta \in \Theta$,

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

ce qui implique que

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta [\{\sqrt{n}(\hat{\theta}_n - \theta)\}^2] = 1/4. \quad (\text{I-3.5})$$

Pour construire un intervalle de confiance à partir de la borne (I-3.5), on procède de la façon suivante. Pour tout $\delta > 0$ et $\theta \in \Theta$, par l'inégalité de Bienayme-Tchebychev (Lemme IV-3.2),

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

Choisissons $\alpha \in (0, 1)$ et 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

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

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

$$\mathbb{P}_\theta (\theta \in \mathcal{J}_{n,\alpha}) \geq 1 - \alpha.$$

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

L'interprétation de $\mathcal{J}_{n,\alpha}$ est claire. On garantit, avec probabilité $1 - \alpha$, que la quantité inconnue d'intérêt θ appartient à l'ensemble aléatoire $\mathcal{J}_{n,\alpha}$ que l'on observe. La qualité de cet intervalle se mesure à sa longueur, notée $|\mathcal{J}_{n,\alpha}|$, qui est ici égale à

$$|\mathcal{J}_{n,\alpha}| = \frac{1}{\sqrt{n\alpha}}.$$

L'ordre de grandeur de $\mathcal{J}_{n,\alpha}$, $1/\sqrt{n}$, est le même que celui de la perte quadratique de l'estimateur $\hat{\theta}_n$. On a aussi $|\mathcal{J}_{n,\alpha}| \rightarrow +\infty$ lorsque $\alpha \rightarrow 0$. Il s'agit d'un compromis inévitable entre précision d'estimation (vouloir $|\mathcal{J}_{n,\alpha}|$ petit) et risque (vouloir α petit) qui sont antagonistes.

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

On peut raffiner ce résultat en utilisant les inégalités de déviation exponentielle (Section IV-3.4). Comme établi au corollaire IV-3.10, pour tout $\delta > 0$,

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

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 $\alpha > 0$,

$$\mathcal{J}_{n,\alpha}^* = \left[\hat{\theta}_n \pm \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \right].$$

Par construction $\mathcal{J}_{n,\alpha}^*$ est un intervalle de confiance pour θ de niveau $1 - \alpha$. On a

$$\frac{|\mathcal{J}_{n,\alpha}^*|}{|\mathcal{J}_{n,\alpha}|} = \sqrt{2\alpha \log(2/\alpha)} \rightarrow 0 \quad \text{lorsque } \alpha \rightarrow 0.$$

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

$$\frac{|\mathcal{J}_{n,\alpha}^*|}{|\mathcal{J}_{n,\alpha}|} = 0,61.$$

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 $\mathcal{J}_{n,\alpha}$ et $\mathcal{J}_{n,\alpha}^*$ sont comparables en n . $\diamond$

Nous introduisons maintenant un objet important pour la construction de régions de confiance.

Définition I-3.17 (Fonction Pivotal). Soit $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique paramétrique où $\Theta \in \mathcal{B}(\mathbb{R}^d)$. On dit qu'une fonction mesurable

$$G : (Z \times \Theta, \mathcal{Z} \otimes \mathcal{B}(\Theta)) \rightarrow (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$$

$$(z, \theta) \mapsto G(z, \theta)$$

est *pivotal* si, pour tout $\theta \in \Theta$, la fonction $z \mapsto G(z, \theta)$ est mesurable et 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. pour tout $\theta, \vartheta \in \Theta$,

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

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

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

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 $A_\alpha \in \mathcal{B}(\mathbb{R}^p)$ ainsi choisi, la région définie pour tout $z \in Z$ par

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

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 $A_\alpha =]-q_{1-\alpha/2}, q_{1-\alpha/2}]$. 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 $A_\alpha =]q_{p_1(\alpha)}, q_{p_2(\alpha)}]$. 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-3.18 (Intervalle de confiance pour un échantillon gaussien de variance connue). Soit $(X_1, \dots, X_n)$ un n -échantillon d'un modèle gaussien

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

où la variance $\sigma^2 > 0$ est connue. Nous cherchons à construire un intervalle de confiance

$$[m(X_1, \dots, X_n), M(X_1, \dots, X_n)]$$

pour la moyenne μ , de niveau de confiance $1 - \alpha$:

$$\mathbb{P}_\mu (\mu \in [m(X_1, \dots, X_n), M(X_1, \dots, X_n)]) = 1 - \alpha. \quad (\text{I-3.7})$$

Pour tout $\mu \in \mathbb{R}$,

$$G(X_1, \dots, X_n, \mu) = \sqrt{n}(\bar{X}_n - \mu) / \sigma,$$

où $\bar{X}_n$ est la moyenne empirique, est distribué sous $\mathbb{P}_\mu$ suivant une loi gaussienne centrée réduite, c'est donc un pivot. Notons Φ la fonction de répartition d'une loi gaussienne standard et, pour tout $\alpha \in]0, 1[$, z_α le α -quantile de cette loi, défini par

$$\Phi(z_\alpha) = \alpha.$$

En particulier, nous avons $z_{1-\alpha/2} = 1,96$ si $\alpha = 0,05$ et $z_{1-\alpha/2} = 2,57$ si $\alpha = 0,01$. L'intervalle symétrique $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$ vérifie alors

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

Autrement dit $m(X_1, \dots, X_n) = \bar{X}_n - \sigma z_{1-\alpha/2} / \sqrt{n}$ et $M(X_1, \dots, X_n) = \bar{X}_n + \sigma z_{1-\alpha/2} / \sqrt{n}$ sont les extrémités d'un intervalle de confiance de niveau $1 - \alpha$ pour μ i.e. (I-3.7) est valide pour tout $\mu \in \mathbb{R}$.

On peut aussi construire une borne de confiance inférieure de niveau $1 - \alpha$ donné, $\bar{X}_n - \sigma z_{1-\alpha} / \sqrt{n}$ ou une borne de confiance supérieure de niveau $1 - \alpha$ donné, $\bar{X}_n + \sigma z_{1-\alpha} / \sqrt{n}$.

L'intervalle de confiance bilatéral a pour diamètre $2\sigma z_{1-\alpha/2} / \sqrt{n}$ qui tend vers 0 quand $n \rightarrow \infty$, pour tout niveau de confiance $1 - \alpha$ donné. $\diamond$

Exemple I-3.19 (Intervalle de confiance pour la moyenne à variance inconnue). Nous allons maintenant déterminer des intervalles de confiance lorsque la variance est inconnue.

Soit $(X_1, \dots, X_n)$ un n -échantillon du modèle gaussien

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

Considérons la fonction

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

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

$$S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (\text{I-3.8})$$

Le théorème IV-5.24 montre que $G(X_1, \dots, X_n, \mu)$ est une fonction pivotale distribuée suivant une loi de Student à $(n-1)$ degrés de liberté : pour tout $A \in \mathbb{R}$ et $\theta = (\mu, \sigma^2) \in \Theta$,

$$\mathbb{P}_\theta (G(X_1, \dots, X_n, \mu) \in A) = t_{n-1}(A).$$

Notons $t_\alpha^{(n-1)}$ le α -quantile de la loi t_{n-1} . La densité de Student est une fonction paire ce qui justifie de considérer des intervalles symétriques. Nous avons

$$\mathbb{P}_\theta \left(-t_{1-\alpha/2}^{(n-1)} \leq G(X_1, \dots, X_n, \mu) \leq t_{1-\alpha/2}^{(n-1)} \right) = 1 - \alpha.$$

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

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

avec

$$m(X_1, \dots, X_n) = \bar{X}_n - S_n t_{1-\alpha/2}^{(n-1)} / \sqrt{n} \quad \text{et} \quad M(X_1, \dots, X_n) = \bar{X}_n + S_n t_{1-\alpha/2}^{(n-1)} / \sqrt{n}. \quad \diamond$$

Nous avons représenté dans la fig. I-3.7 les quantiles d'ordre $p = 0.95$ de la loi de Student à $(n-1)$ degrés de liberté et de la loi gaussienne centrée réduite. Nous voyons sur ce graphique que, dès que la taille de l'échantillon dépasse $n \geq 100$, les valeurs $t_{0.95}^{(n-1)}$ et $z_{0.95}$ sont très proches.

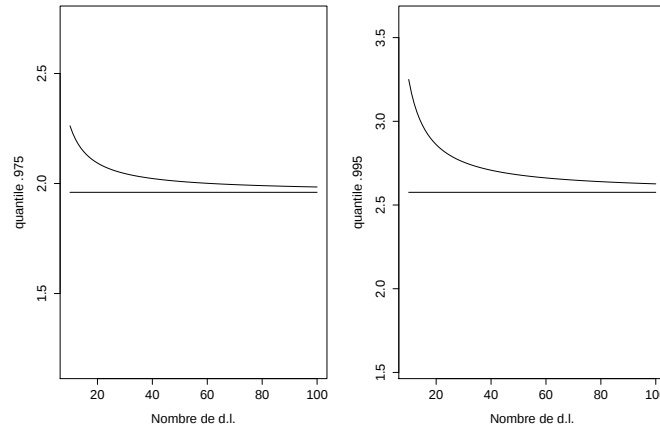


FIGURE I-3.7 – Quantile à .975 et .995 pour la loi gaussienne standard et la loi de Student en fonction du nombre de degrés de liberté de la loi de Student

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

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

La fonction

$$G(X_1, \dots, X_n, \sigma^2) = (n-1)S_n^2 / \sigma^2$$

où S_n^2 est la variance de l'échantillon ((I-3.8)) est distribuée suivant une loi de χ^2 à $n-1$ degrés de liberté (voir Section IV-5.4). Cette fonction est donc pivotale. Si nous notons $\chi_{\alpha}^2(n-1)$ 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) ,

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

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

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

est un intervalle de confiance pour σ^2 de niveau de confiance $1 - \alpha$. 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, $\alpha_1^* \simeq \alpha_2^* \simeq \alpha/2$. $\diamond$

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

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

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

$$I_1(X_1, \dots, X_n) = \left[\bar{X}_n - S_n t_{1-\alpha/4}^{(n-1)} / \sqrt{n}, \bar{X}_n + S_n t_{1-\alpha/4}^{(n-1)} / \sqrt{n} \right],$$

pour l'intervalle de confiance pour la moyenne μ de niveau de confiance $1 - \alpha/2$ et :

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

Test statistique pour l'intervalle de confiance pour la variance σ^2 de niveau de confiance $1 - \alpha/2$. Nous avons :
 Région d'acceptation

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

et donc $I(X_1, \dots, X_n) = I_1(X_1, \dots, X_n) \times I_2(X_1, \dots, X_n)$ est une région de confiance de niveau de confiance supérieur à $1 - \alpha$. En fait, on peut montrer en utilisant le théorème IV-5.24 que le niveau de confiance de cette région est exactement égal à $(1 - \alpha/2)^2$. $\diamond$

I-3.4 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-3.22 (Test bilatéral pour la moyenne d'une gaussienne). Soient $(X_1, \dots, X_n)$ un n -échantillon d'un modèle gaussien

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

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

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

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

$$I(Z) = \left[\bar{X}_n \pm S_n t_{1-\alpha/2}^{(n-1)} / \sqrt{n} \right], \quad S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

où $t_{\beta}^{(n-1)}$ est le quantile de niveau β de la loi de Student à $n-1$ degrés de liberté. Considérons le test :

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

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

$$G(Z; v) = \sqrt{n}(\bar{X}_n - v) / S_n,$$

La taille de $\phi_{\mu_0}(Z)$ est égale à α pour le test (I-3.9). 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

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

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. $\diamond$

Cet exemple est un cas particulier du principe de dualité entre intervalles de confiance et tests que nous présentons maintenant. Soit $(Z, \mathcal{Z}, \{\mathbb{P}_{\theta}, \theta \in \Theta\})$ une expérience statistique. Soit $g : \Theta \rightarrow \mathbb{R}^q$. Dans l'exemple I-3.22, nous avons simplement $q = 1$, $\theta = (\mu, \sigma^2)$ et $g(\theta) = \mu$.

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

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

Notons

$$\Theta_0^v = \{ \theta \in \Theta : g(\theta) = v \}. \quad (\text{I-3.11})$$

Pour tout $\theta \in \Theta_0^v$, nous avons

$$\mathbb{P}_{\theta}(\phi_v(Z) = 1) \leq \alpha. \quad (\text{I-3.12})$$

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

$$S(z) = \{v \in \mathbb{R}^q : \phi_v(z) = 0\} . \quad (\text{I-3.13})$$

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

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

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 $\theta \in \Theta$,

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

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

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

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

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

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

I-3.5 Méthode du pivot

Les pivots, ou fonctions pivotales $(z, \theta) \mapsto G(z, \theta)$, ont été introduits à la définition I-3.17. Étant donné un pivot G , on peut déterminer, pour tout $\alpha \in [0, 1]$, des réels t_α^- et t_α^+ tels que

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

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 intervalle de confiance. Cependant, l'exemple suivant décrit des situations classiques dans lesquelles cette résolution est en fait possible.

Exemple I-3.23 (Test sur la moyenne d'une loi $N(\mu, \sigma^2)$ avec σ^2 connue). Soit $(X_1, \dots, X_n)$ un n -échantillon du modèle $\{N(\mu, \sigma^2), \mu \in \mathbb{R}\}$ où la variance σ^2 connue. La fonction

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

est pivotale. Pour tout $\mu \in \mathbb{R}$, la loi de $G(\bar{X}_n, \mu)$ est gaussienne centrée réduite. Ainsi pour le test

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

on a sous H_0 :

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

Comme :

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

une région d'acceptation peut donc être définie, pour un test de taille α , par :

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

Ici la fonction puissance $\beta(\mu)$ peut être déterminée car la loi de $(\bar{X}_n - \mu_0)/(\sigma/\sqrt{n})$ reste gaussienne quand μ est différent de μ_0 . En effet :

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

où Φ est la fonction de répartition de la loi de Gauss centrée réduite. La fonction $\beta(\mu)$ à un minimum au point μ_0 (qui vaut précisément α). Par ailleurs on vérifie aisément que $\beta(\mu) \rightarrow 1$ quand $n \rightarrow \infty$ aussi bien lorsque $\mu > \mu_0$ que lorsque $\mu < \mu_0$. $\diamond$

Exemple I-3.24 (Test sur la moyenne d'une loi $N(\mu, \sigma^2)$ avec σ^2 inconnue). Soit $(X_1, \dots, X_n)$ un n -échantillon du modèle gaussien

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

On considère la fonction

$$G(X_1, \dots, X_n; \mu) = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \quad (\text{I-3.14})$$

où S_n^2 est la variance de l'échantillon. La fonction $G(X_1, \dots, X_n; \mu)$ est une fonction pivotale. Le théorème de Gosset (Théorème IV-5.24) montre que, pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$,

- (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 $N(\mu, \sigma^2/n)$.
- (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 $G(X_1, \dots, X_n; \mu)$ suit une loi de Student $t(n-1)$ (voir Section IV-5.6 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, nous avons donc

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

Comme

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

le test

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

est de taille α . La puissance du test est donnée par

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

La loi de $(\bar{X}_n - \mu_0)/(S_n/\sqrt{n})$ ne suit plus une loi de Student car μ_0 n'est plus la moyenne de $\bar{X}_n$. Nous avons

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

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

Exemple I-3.25 (Test sur la variance d'une loi $N(\mu, \sigma^2)$ avec μ inconnu). Soit $(X_1, \dots, X_n)$ un n -échantillon du modèle gaussien

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

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

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

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

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

ce qui conduit à considérer le test

$$\delta_\alpha(X_1, \dots, X_n; \sigma_0^2) = \mathbb{1} \left\{ \chi_{\alpha/2}^2(n-1) < \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{1-\alpha/2}^2(n-1) \right\}$$

◇

I-3.6 Utilisation d'inégalités de déviation

La méthode du pivot permet d'obtenir directement les quantiles requis pour déterminer les seuils critiques des test. Toutefois, il est en pratique assez rare d'avoir de tels pivots 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-1.6 consiste à utiliser une approche asymptotique.

Pour fixer les idées, considérons $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ 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 zone de rejet de la forme suivante,

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

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 $\mathbb{P}_\theta(T(Z) - g(\theta) > \varepsilon)$ de la forme : pour tout $\varepsilon > 0$,

$$\mathbb{P}_\theta(T(Z) - g(\theta) > \varepsilon) \leq \gamma_\theta(\varepsilon). \quad (\text{I-3.15})$$

On a donc

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

En supposant que $\gamma(\varepsilon) \xrightarrow{\varepsilon \rightarrow +\infty} 0$, 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 $\mathcal{R} = \{z \in Z : T(z) \geq g_0 + y_\alpha\}$,

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta (Z \in \mathcal{R}_\alpha) \leq \gamma(y_\alpha) \leq \alpha . \quad (\text{I-3.17})$$

Le test de région de rejet $\mathcal{R} = \{z \in Z : T(z) \geq g_0 + y_\alpha\}$ est de niveau α .

Supposons que $Z := (X_1, \dots, X_n)$ est un n -échantillon d'un modèle statistique

$$(\mathcal{X}, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\}) .$$

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-3.2 :

$$\text{pour tout } \varepsilon > 0, \quad \mathbb{P}_\theta \left(\left| n^{-1} \sum_{i=1}^n t(X_i) - g(\theta) \right| \geq \varepsilon \right) \leq \frac{\text{Var}_\theta(t(X_1))}{n\varepsilon^2} .$$

On déduit de cette inégalité qu'on peut utiliser dans la majoration (I-3.16) la fonction $\gamma(\varepsilon) = v^2(\Theta_0)/(n\varepsilon^2)$, où $v^2(\Theta_0) = \sup_{\theta \in \Theta_0} \text{Var}_\theta(t(X_1))$. 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

$$\mathcal{R}_\alpha = \left\{ z \in Z : T(z) \geq g_0 + \frac{v(\Theta_0)}{\sqrt{n\alpha}} \right\} = \left\{ T(Z) \geq g_0 + \frac{v(\Theta_0)}{\sqrt{n\alpha}} \right\} ,$$

est de niveau α .

Exemple I-3.26. Soit $Z = (X_1, \dots, X_n)$ un n -échantillon de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta = [0, 1]\})$$

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 $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Pour appliquer la construction précédente, il reste à évaluer

$$v^2(\Theta_0) = \sup_{\theta \leq 1/2} \text{Var}_\theta(X_1) = \sup_{\theta \leq 1/2} \theta(1-\theta) = \sup_{\theta \leq 1/2} \theta(1-\theta) = \frac{1}{4} .$$

On en déduit que le test de région de rejet

$$\mathcal{R} = \left\{ (x_1, \dots, x_n) \in \{0, 1\}^n : n^{-1} \sum_{i=1}^n x_i > \frac{1}{2} + \frac{1}{2\sqrt{n\alpha}} \right\} ,$$

est de niveau α . ◇

Lorsque les variables sont bornées, on peut remplacer l'inégalité de Tchebycheff par celle d'Hoeffding (Théorème IV-3.9). Plus précisément, supposons que l'observation $Z = (X_1, \dots, X_n)$ où $X_1, \dots, X_n$ sont des variables aléatoires i.i.d. à valeurs dans $(\mathcal{X}, \mathcal{Z})$. Supposons qu'il existe une fonction $t : \mathcal{X} \rightarrow \mathbb{R}$ mesurable telle que pour tout $\inf_{x \in \mathcal{X}} t(x) = a > -\infty$ et $b = \sup_{x \in \mathcal{X}} t(x) < \infty$ 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

$$\mathbb{P}_\theta \left(n^{-1} \sum_{i=1}^n t(X_i) - g(\theta) > \varepsilon \right) \leq e^{-2n\varepsilon^2/(b-a)^2} .$$

Notons que la borne $\gamma_\theta(x)$ dans (I-3.15) ne dépend pas de θ (elle dépend par contre de a et de b), on peut donc prendre directement $\gamma(\varepsilon) = e^{-2n\varepsilon^2/(b-a)^2}$ dans (I-3.16), 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

$$\mathcal{R} = \left\{ (x_1, \dots, x_n) \in \mathbb{X}^n : n^{-1} \sum_{i=1}^n t(x_i) > g_0 + (b-a) \sqrt{\frac{1}{2n} \ln \left(\frac{1}{\alpha} \right)} \right\} ,$$

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

$$\mathcal{R} = \left\{ z \in \{0, 1\}^n : \bar{x}_n > \frac{1}{2} + \sqrt{\frac{1}{2n} \ln \left(\frac{1}{\alpha} \right)} \right\} ,$$

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 ! $\diamond$

Chapitre I-4

Bases de la théorie de la décision

I-4.1 Règles de décision, pertes et risques

I-4.1.1 Perte et risque d'une règle de 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.

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 $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique paramétrique et $\ell : \Theta \times A \rightarrow \mathbb{R}_+$ une fonction de perte, où A est un ensemble d'actions. On suppose A muni d'une tribu $\mathcal{A}$ telle que $\ell(\theta, \cdot) : (A, \mathcal{A}) \rightarrow \mathbb{R}$ soit mesurable pour tout $\theta \in \Theta$.

Une règle de décision est une fonction mesurable $\delta : (Z, \mathcal{Z}) \rightarrow (A, \mathcal{A})$; le risque d'une règle de décision δ est défini comme l'espérance

$$R_\ell(\theta, \delta) = \mathbb{E}_\theta[\ell(\theta, \delta(Z))] . \quad (\text{I-4.1})$$

L'estimation ponctuelle peut être naturellement formulée comme un problème de décision. En effet, soit $g : \Theta \rightarrow Y$ une fonction à estimer, où $(Y, \mathcal{Y})$ est un espace mesurable. En prenant pour espace d'actions $A = Y$, un estimateur $T : (Z, \mathcal{Z}) \rightarrow (Y, \mathcal{Y})$ 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 $W : Y \times Y \rightarrow \mathbb{R}_+$ mesurable, on peut naturellement associer la fonction de perte, définie pour tout $a \in A = Y$ et $\theta \in \Theta$:

$$\ell(\theta, a) = W(g(\theta), a) .$$

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 $g : \Theta \rightarrow \mathbb{R}^p$, $\theta \mapsto g(\theta)$,

— la *perte quadratique* $\ell(\theta, a) = \|a - g(\theta)\|_2^2$ associée au carré de la norme euclidienne

$$W(u, v) = \|u - v\|_2^2 = \sum_{i=1}^p (u^{(i)} - v^{(i)})^2 .$$

— la *perte absolue* $\ell(\theta, a) = \|a - g(\theta)\|_1$ associée à la norme 1

$$W(u, v) = \|u - v\|_1 = \sum_{i=1}^p |u^{(i)} - v^{(i)}| .$$

◇

Notons que le test d'hypothèse peut également se formuler comme un problème de décision. En effet, si Θ est la réunion disjointe de Θ_0 et Θ_1 , on considère l'ensemble d'actions $A = \{0, 1\}$ et la fonction de perte

$$\ell(\theta, 0) = \ell_0 \mathbb{1}_{\Theta_1}(\theta) , \quad \ell(\theta, 1) = \ell_1 \mathbb{1}_{\Theta_0}(\theta) ,$$

pour $\ell_0, \ell_1 > 0$. Une règle de décision correspond alors à un test pur $\phi : Z \rightarrow \{0, 1\}$ (définition I-3.2), et le risque d'un tel test vaut

$$R_\ell(\theta, \phi) = L_0 \mathbb{P}_\theta(\phi(Z) = 0) \mathbb{1}_{\Theta_1}(\theta) + L_1 \mathbb{P}_\theta(\phi(Z) = 1) \mathbb{1}_{\Theta_0}(\theta) .$$

Remarque I-4.3 (Perte vectorielle). Dans certains cas, il arrive que l'on soit amené à contrôler simultanément plusieurs critères. Formellement, cela revient à considérer une fonction de perte $\ell : \Theta \times A \rightarrow \mathbb{R}_+^N$ *vectorielle* et non plus scalaire ; les composantes $\ell_1, \dots, \ell_N : \Theta \times A \rightarrow \mathbb{R}_+$ de ℓ sont alors des fonctions de perte (scalaires). Comme dans la définition I-4.1, il est possible de définir le *risque* vectoriel d'une règle de décision $\delta : Z \rightarrow A$ par :

$$R_\ell(\theta, \delta) = \mathbb{E}_\theta[\ell(\theta, \delta(Z))] = (R_{\ell_1}(\theta, \delta), \dots, R_{\ell_N}(\theta, \delta)) . \quad (\text{I-4.2})$$

Les pertes vectorielles interviennent naturellement lorsque l'on cherche à estimer simultanément plusieurs fonctions $g_1, \dots, g_N$ (par exemple, la moyenne et la variance de $Z \sim \mathbb{P}_\theta$), et que l'on souhaite contrôler le risque pour chacune de ces procédures séparément.

Sauf mention explicite du contraire, nous considérons dans ce cours des pertes scalaires. La principale raison en est que, pour développer une théorie satisfaisante de l'optimalité des règles de décision, il est nécessaire de combiner les différentes composantes de la perte vectorielle en une perte scalaire "agrégée". En effet, la qualité d'une règle de décision dépend de l'objectif considéré, c'est-à-dire de la composante du vecteur de perte considérée. Il existe ainsi un compromis entre les différentes composantes du risque ; afin de comparer des règles de décision, il faut spécifier des *préférences* sur les vecteurs de perte dans $\mathbb{R}_+^N$, c'est-à-dire une relation de *préordre* totale $\preceq$ sur $\mathbb{R}_+^N$. Sous des hypothèses raisonnables (comme par exemple la "continuité" des préférences), il existe une fonction $U : \mathbb{R}_+^N \rightarrow \mathbb{R}_+$, appelée *fonction d'utilité*, telle que pour tous $L, L' \in \mathbb{R}_+^N$, $L \preceq L'$ si et seulement si $U(L) \leq U(L')$; on est alors ramené à l'étude de la fonction de perte scalaire $\ell' = U \circ \ell$.

Nous exposerons également dans la remarque I-4.5 une réduction générique des pertes vectorielles aux pertes scalaires. ◇

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-4.4 (Admissibilité). Une règle de décision $\delta : Z \rightarrow A$ est dite *inadmissible* (pour la perte ℓ) s'il existe une règle δ^* telle que

$$\begin{array}{ll} R_\ell(\theta, \delta^*) \leq R_\ell(\theta, \delta) & \text{pour tout } \theta \in \Theta \\ R_\ell(\theta, \delta^*) < R_\ell(\theta, \delta) & \text{pour au moins un } \theta \in \Theta \end{array}$$

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.

Remarque I-4.5. Il est possible d'adapter les notions de *préférabilité* et d'*admissibilité* des règles de décision au cas de pertes vectorielles, en remplaçant l'inégalité (dans $\mathbb{R}$) par l'inégalité coordonnée par coordonnée (dans $\mathbb{R}^N$).

Il est toutefois possible de se ramener à une perte scalaire de la façon suivante. On considère le nouvel espace de paramètres $\tilde{\Theta} = \Theta \times \{1, \dots, N\}$, et la fonction de perte $\tilde{\ell}(a, (\theta, i)) := \ell_i(\theta, a)$. Alors, une règle de décision δ_1 est préférable à une autre règle δ_2 pour ℓ si et seulement si elle l'est pour $\tilde{\ell}$, et une règle δ est admissible pour ℓ si et seulement si elle l'est pour $\tilde{\ell}$. Ainsi, de nombreux résultats établis dans ce chapitre pour les pertes scalaires se transfèrent directement au cas de pertes vectorielles. $\diamond$

La propriété d'admissibilité est en quelque sorte un prérequis minimal sur une règle de décision : dire qu'une règle est inadmissible signifie qu'il existe une autre règle dont le risque est moindre *quelle que soit* la vraie valeur du paramètre θ . En revanche, il existe en général beaucoup de règles incomparables ainsi que des règles admissibles indésirables, comme le montre l'exemple suivant.

Exemple I-4.6 (Estimateurs admissibles indésirables). Considérons un n -échantillon du modèle de Bernoulli,

$$(\{0, 1\}^n, \mathcal{P}(\{0, 1\}^n), \{\text{Ber}_{\theta}^{\otimes n}, \theta \in \Theta = [0, 1]\}).$$

On cherche à estimer le paramètre θ , et l'on considère la fonction de perte quadratique $\ell(\theta, a) = (\theta - a)^2$ (ici $\mathbf{A} = \Theta$). Pour tout $\vartheta \in]0, 1[$, l'estimateur constant $\hat{\theta}_{\vartheta} \equiv \vartheta \in [0, 1]$ est admissible : en effet $R_{\ell}(\vartheta, \hat{\theta}_{\vartheta}) = 0$ et si T est un estimateur tel que $R_{\ell}(\vartheta, T) = 0$ alors $T = \hat{\theta}_{\vartheta}$. En revanche, cet estimateur n'est pas très intéressant, puisqu'il ne sait estimer qu'un seul paramètre ! $\diamond$

Ainsi, la propriété d'admissibilité n'est pas suffisante pour garantir de bons estimateurs. Cela amène à considérer des critères d'optimalité plus restrictifs. Nous allons en considérer deux d'usage courant en statistiques : le risque minimax et le risque intégré (ou bayésien).

I-4.1.3 Approche minimax

Partons de l'exemple I-4.6. La faiblesse des estimateurs constants est qu'ils ne se comportent bien que pour une valeur du paramètre. Comme on ne connaît pas cette valeur a priori, il est naturel de rechercher des estimateurs qui admettent un risque contrôlé uniformément sur Θ . Cela conduit à la définition suivante.

Définition I-4.7 (Risque minimax). Soit δ une règle de décision. Le risque dans le pire des cas de δ est par définition

$$\bar{R}_{\ell}(\delta) = \sup_{\theta \in \Theta} R_{\ell}(\theta, \delta). \quad (\text{I-4.3})$$

Le risque minimax pour la fonction de perte ℓ est l'infimum sur la classe des règles de décision du risque de pire des cas :

$$\bar{R}_{\ell} = \inf_{\delta} \bar{R}_{\ell}(\delta) = \inf_{\delta} \sup_{\theta \in \Theta} R_{\ell}(\theta, \delta). \quad (\text{I-4.4})$$

Enfin, une règle δ^* est dite optimale au sens minimax, ou tout simplement minimax, si son risque maximal est égal au risque minimax, i.e.

$$\bar{R}_{\ell}(\delta^*) = \bar{R}_{\ell} = \inf_{\delta} \bar{R}_{\ell}(\delta). \quad (\text{I-4.5})$$

L'approche minimax, qui sera étudiée plus en détail en ??, consiste ainsi à chercher des règles de décision dont le risque dans le pire des cas $\bar{R}_{\ell}(\delta)$ est minimal. Les procédures minimax ont l'avantage d'être "sûres", au sens où elles se prémunissent contre toutes les valeurs du paramètre θ , de manière uniforme. En revanche, il arrive que la recherche de garanties uniformes sur le risque conduise à des procédures trop conservatrices.

I-4.1.4 Approche bayésienne

Une approche alternative consiste à introduire une mesure de probabilité Π sur Θ permettant de pondérer les différentes valeurs possibles du paramètre θ , puis à considérer l'intégrale du risque selon la loi a priori Π (plutôt que de considérer son supremum). Plus précisément, supposons Θ muni d'une tribu $\mathcal{T}$ telle que :

- pour tout $B \in \mathcal{Z}$, la fonction $(\Theta, \mathcal{T}) \rightarrow \mathbb{R}_+$, $\theta \mapsto \mathbb{P}_\theta(B)$ est mesurable.
- la fonction de perte $\ell : (\Theta \times A, \mathcal{T} \otimes \mathcal{A}) \rightarrow \mathbb{R}_+$ est mesurable.

Par un argument classique (en passant des fonctions indicatrices de pavés aux indicatrices de parties mesurables arbitraires en prenant la tribu engendrée, puis aux fonctions étagées par combinaison linéaire, et enfin aux fonctions mesurables positives par passage à la limite), la première condition implique que, pour toute fonction mesurable positive

$$f : (\Theta \times Z, \mathcal{T} \otimes \mathcal{Z}) \rightarrow \mathbb{R}_+,$$

la fonction

$$\theta \mapsto \int_Z f(\theta, z) \mathbb{P}_\theta(dz)$$

est $\mathcal{T}$ -mesurable. En particulier, en appliquant cela à la fonction $(\theta, z) \mapsto \ell(\theta, \delta(z))$ (mesurable par la seconde hypothèse et par mesurabilité de δ), la fonction

$$\theta \mapsto R_\ell(\theta, \delta) = \int_Z \ell(\theta, \delta(z)) \mathbb{P}_\theta(dz)$$

est $\mathcal{T}$ -mesurable. Nous pouvons alors former la définition suivante :

Définition I-4.8 (Risque de Bayes). Soit Π une mesure de probabilité sur $(\Theta, \mathcal{T})$, appelée loi a priori. Pour toute règle δ , le risque de Bayes (ou risque intégré) de δ par rapport à Π est défini par

$$R_\ell(\Pi, \delta) = \int_\Theta R_\ell(\theta, \delta) \Pi(d\theta). \quad (\text{I-4.6})$$

Le risque de Bayes (ou bayésien) selon Π est défini comme l'infimum du risque intégré selon toutes les règles δ :

$$\underline{R}_\ell(\Pi) = \inf_{\delta} R_\ell(\Pi, \delta). \quad (\text{I-4.7})$$

Une règle δ_Π est dite bayésienne selon Π si elle atteint le risque de Bayes selon Π , i.e. si

$$R_\ell(\Pi, \delta_\Pi) = \underline{R}_\ell(\Pi) = \inf_{\delta} R_\ell(\Pi, \delta).$$

Le risque et les règles de Bayes seront étudiés en détail en section ???. De plus, quoique les points de vue minimax et bayésiens diffèrent sensiblement, nous verrons en ??? qu'ils entretiennent des liens utiles à l'étude de procédures minimax.

I-4.1.5 Optimalité sous contraintes

Nous concluons cette section en mentionnant une dernière approche de l'optimalité des estimateurs (ou des règles de décision), consistant à se restreindre à certaines "classes" d'estimateurs sujets à des contraintes. 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 suivante.

Une autre classe remarquable est constituée des estimateurs invariants sous l'action de certains groupes de transformations. Ces approches sont élégantes mais elles ne s'appliquent qu'à des classes très limitées de problèmes (on consultera par exemple [?] pour approfondir de telles constructions).

Exemple I-4.9 (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 $(X_1, \dots, X_n)$ un n -échantillon du modèle

$$\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \{\mathbb{P}_\theta = p_\theta \cdot \lambda_{\text{Leb}}, \theta \in \Theta\} \right).$$

Considérons le problème de l'estimation du paramètre de translation θ pour une fonction de perte $W(u, v) = w(u - v)$ où $w : \mathbb{R}^d \rightarrow \mathbb{R}_+$ est une fonction mesurable positive. Remarquons que, pour tout $c \in \mathbb{R}^d$ et $\theta \in \Theta$,

$$\mathbb{E}_\theta[h(X_j + c)] = \mathbb{E}_{\theta+c}[h(X_j)]$$

pour toute fonction h mesurable positive. Il est donc naturel de se restreindre à la classe d'estimateurs satisfaisant la propriété d'invariance par translation, i.e. pour tout $c \in \mathbb{R}^d$,

$$T(x_1 + c, \dots, x_n + c) = T(x_1, \dots, x_n) + c, \quad \text{pour tous } x_1, \dots, x_n \in \mathbb{R}^d. \quad (\text{I-4.8})$$

Les estimateurs qui satisfont (I-4.8) sont dits *équivariants*. Notons $T_1(x_1, \dots, x_{n-1}) = T(0, x_1, \dots, x_{n-1})$ pour tous $x_1, \dots, x_{n-1} \in \mathbb{R}^d$. Un estimateur T est équivariant si et seulement si, pour tous $x_1, \dots, x_n \in \mathbb{R}^d$,

$$T(x_1, \dots, x_n) = x_1 + T_1(x_2 - x_1, \dots, x_n - x_1). \quad (\text{I-4.9})$$

Si T est un estimateur équivariant, alors le risque de T pour la perte $W(u, v) = w(u - v)$ ne dépend pas de $\theta \in \Theta$. En effet

$$\begin{aligned} R_W(\theta, T) &= \mathbb{E}_\theta[w(T(X_1, \dots, X_n) - \theta)] \\ &= \mathbb{E}_\theta[w(T(X_1 - \theta, \dots, X_n - \theta))] \\ &= \mathbb{E}_0[w(T(X_1, \dots, X_n))] = R_w(0, T) =: R_w(T). \end{aligned}$$

Tout estimateur équivariant T_0 pour lequel $\inf_T R_w(T)$ est réalisé est donc optimal dans la classe des estimateurs équivariants. $\diamond$

I-4.2 Le cas du risque quadratique : estimation non biaisée

Dans toute cette section, nous considérons le problème de l'estimation d'une fonction vectorielle $g : \Theta \rightarrow \mathbb{R}^p$ 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-4.10 (Biais). Soient $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ une expérience statistique, $g : \Theta \rightarrow \mathbb{R}^p$ 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 :

$$b_g(\theta, T) = \mathbb{E}_\theta[T(Z)] - g(\theta). \quad (\text{I-4.10})$$

T est un estimateur sans biais de $g : \theta \mapsto g(\theta)$ si pour tout $\theta \in \Theta$, $b_g(\theta, T) = 0$, i.e. $\mathbb{E}_\theta[T(Z)] = g(\theta)$.

Exemple I-4.11 (Moyenne empirique). Soit $(X_1, \dots, X_n)$ un n -échantillon de $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{P}_\theta, \theta \in \Theta\})$. 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 $\mu : \theta \mapsto \mu(\theta) := \mathbb{E}_\theta[X_1]$. $\diamond$

Exemple I-4.12 (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 $\sigma^2(\theta) = \text{Var}_\theta(X_1)$. Une idée naturelle est de considérer l'estimateur

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (\text{I-4.11})$$

Il s'avère que cet estimateur n'est pas sans biais. En effet,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 &= \frac{1}{n} \sum_{i=1}^n [(X_i - \mu(\theta)) - (\bar{X}_n - \mu(\theta))]^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu(\theta))^2 - 2(\bar{X}_n - \mu(\theta)) \frac{1}{n} \sum_{i=1}^n (X_i - \mu(\theta)) + (\bar{X}_n - \mu(\theta))^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu(\theta))^2 - 2(\bar{X}_n - \mu(\theta))^2 + (\bar{X}_n - \mu(\theta))^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu(\theta))^2 - (\bar{X}_n - \mu(\theta))^2, \end{aligned}$$

d'où :

$$\mathbb{E}_\theta[\hat{\sigma}_n^2] = \frac{1}{n} \sum_{i=1}^n \text{Var}_\theta(X_i) - \text{Var}_\theta(\bar{X}_n) = \sigma^2(\theta) - \frac{\sigma^2(\theta)}{n} = \frac{n-1}{n} \sigma^2(\theta).$$

Le biais de l'estimateur de la variance empirique est donc égal à

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

En moyenne, la variance empirique sous-estime donc la variance de la loi étudiée. On déduit de ce qui précède que l'estimateur S_n^2 ci-dessous de $\sigma^2(\theta)$, appelé *variance empirique*, est sans biais :

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (\text{I-4.12})$$

◇

Exemple I-4.13 (Inexistence d'un estimateur sans biais). La propriété pour un estimateur d'être sans biais est restrictive ; dans certains cas, il n'existe même pas d'estimateur sans biais ! En effet, soit X une variable aléatoire de loi binomiale $\text{Bin}(n, \theta)$, avec $\theta \in]0, 1]$. Alors, il n'existe pas d'estimateur sans biais de $g : \theta \mapsto 1/\theta$: pour tout estimateur $T : \{0, \dots, n\} \rightarrow \mathbb{R}$,

$$\mathbb{E}_\theta[T(X)] = \sum_{k=0}^n \binom{n}{k} T(k) \theta^k (1-\theta)^{n-k} \quad (\text{I-4.13})$$

qui est une fonction bornée de $\theta \in]0, 1]$, et ne peut donc pas être égale à $g(\theta)$ (qui diverge en 0).

Plus précisément, une fonction $g : \Theta \rightarrow \mathbb{R}$ admet un estimateur sans biais si et seulement si g est un polynôme 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-4.13), pour toute fonction $T : \{0, \dots, n\} \rightarrow \mathbb{R}$, 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 $\{0, \dots, n\} \rightarrow \mathbb{R}$ 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*. Afin d'écrire cette décomposition sous sa forme générale vectorielle, il est commode de noter, pour $u, v \in \mathbb{R}^p$, $u \otimes v := uv^T \in \text{Mat}_p(\mathbb{R})$, et $u^{\otimes 2} := u \otimes u = uu^T$.

Proposition I-4.14 (Décomposition biais-variance vectorielle). Soient $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ une expérience statistique, $g : \Theta \rightarrow \mathbb{R}^p$ 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 $\theta \in \Theta$. Alors, pour tout $\theta \in \Theta$,

$$\mathbb{E}_\theta[(T(Z) - g(\theta))^{\otimes 2}] = \text{Var}_\theta(T(Z)) + b_g(\theta, T)^{\otimes 2} \quad (\text{I-4.14})$$

où $\text{Var}_\theta(T(Z)) \in \text{Mat}_p(\mathbb{R})$ désigne la matrice de covariance de $T(Z)$ sous $\mathbb{P}_\theta$.

Démonstration. On a $T(Z) - g(\theta) = T(Z) - \mathbb{E}_\theta[T(Z)] + b_g(\theta, T)$. Par bilinéarité de l'application $(u, v) \mapsto u \otimes v = uv^T$, il vient

$$\begin{aligned} (T(Z) - g(\theta))^{\otimes 2} &= (T(Z) - \mathbb{E}_\theta[T(Z)])^{\otimes 2} + (T(Z) - \mathbb{E}_\theta[T(Z)]) \otimes b_g(\theta, T) + \\ &\quad b_g(\theta, T) \otimes (T(Z) - \mathbb{E}_\theta[T(Z)]) + b_g(\theta, T)^{\otimes 2}. \end{aligned}$$

L'équation (I-4.14) s'en déduit en prenant l'espérance sous $\mathbb{P}_\theta$ et en notant que :

$$\begin{aligned} \mathbb{E}_\theta[(T(Z) - \mathbb{E}_\theta[T(Z)]) \otimes b_g(\theta, T)] &= (\mathbb{E}_\theta[T(Z)] - \mathbb{E}_\theta[T(Z)]) \otimes b_g(\theta, T) = 0 \\ \mathbb{E}_\theta[b_g(\theta, T) \otimes (T(Z) - \mathbb{E}_\theta[T(Z)])] &= 0 \\ \mathbb{E}_\theta[(T(Z) - \mathbb{E}_\theta[T(Z)])^{\otimes 2}] &= \text{Var}_\theta(T(Z)). \end{aligned} \quad \square$$

En prenant la trace dans l'équation (I-4.14) et en notant que, pour tout $a \in \mathbb{R}^p$, $\text{Tr}(a^{\otimes 2}) = \text{Tr}(a^T a) = \|a\|_2^2$, on obtient la décomposition suivante du risque quadratique :

Corollaire I-4.15 (Décomposition biais-variance). Reprenons les hypothèses de la proposition I-4.14. Pour tout $\theta \in \Theta$, on a

$$R(\theta, T) = \mathbb{E}_\theta[\|T(Z) - \mathbb{E}_\theta[T(Z)]\|_2^2] + \|b_g(\theta, T)\|_2^2.$$

En particulier, si T est un estimateur sans biais de $g : \Theta \rightarrow \mathbb{R}^p$ alors $R(\theta, T) = \text{Var}_\theta(T(Z))$.

Exemple I-4.16 (Meilleur estimateur linéaire sans biais de l'espérance). Soit $(X_1, \dots, X_n)$ un n -échantillon de

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{P}_\theta, \theta \in \Theta\}).$$

Supposons que, pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[X_1^2] < \infty$. On cherche à estimer l'espérance $\mu(\theta) = \mathbb{E}_\theta[X_1]$, et l'on note $\sigma^2(\theta) = \text{Var}_\theta(X_1)$.

Pour tout $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ tel que $\sum_{i=1}^n \alpha_i = 1$, la combinaison linéaire

$$\hat{\mu}_\alpha := \sum_{i=1}^n \alpha_i X_i$$

est un estimateur sans biais de $\mu : \theta \mapsto \mu(\theta)$. Ainsi, par la décomposition biais-variance (I-4.15), il vient

$$R(\theta, \hat{\mu}_\alpha) = \text{Var}_\theta(\hat{\mu}_\alpha) = \sigma^2(\theta) \sum_{i=1}^n \alpha_i^2 \quad (\text{I-4.15})$$

Ainsi, le risque de $\hat{\mu}_\alpha$ est minimisé pour tout $\theta \in \Theta$ par le $\alpha \in \mathbb{R}^n$ de somme 1 qui minimise $\sum_{i=1}^n \alpha_i^2$. Par l'inégalité de Cauchy-Schwarz (ou par stricte convexité de la fonction carré), on montre que cette quantité est minimisée (uniquement) pour $\alpha_i = 1/n$ pour tout $i = 1, \dots, n$. Ainsi, la moyenne empirique est le meilleur estimateur linéaire sans biais de l'espérance. 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 $(n^{-1}, \dots, n^{-1})$. $\diamond$

Exemple I-4.17 (Estimation de la variance dans le cas gaussien). Soit $(X_1, \dots, X_n)$ un n -échantillon du modèle

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

Comparons le risque quadratique des estimateurs S_n^2 et $\hat{\sigma}_n^2$ (cf. l'exemple I-4.12) de la variance σ^2 . Nous avons vu dans l'exemple I-4.12 que S_n^2 est un estimateur sans biais de σ^2 . Par le Théorème IV-5.24, la variance de l'estimateur S_n^2 est donnée par

$$\text{Var}_{\mu, \sigma^2}(S_n^2) = \frac{2\sigma^4}{n-1}.$$

Par ailleurs, en notant que $\hat{\sigma}_n^2 = \{(n-1)/n\} S_n^2$ et en utilisant la décomposition biais-variance, il vient pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$,

$$\begin{aligned} \mathbb{E}[\hat{\sigma}_n^2] &= \frac{n-1}{n} \sigma^2 \\ \text{Var}_{\mu, \sigma^2}(\hat{\sigma}_n^2) &= \frac{(n-1)^2}{n^2} \text{Var}_{\mu, \sigma^2}(S_n^2) = \frac{2(n-1)\sigma^4}{n^2} \\ \mathbb{E}_\theta \left[\left(\hat{\sigma}_n^2 - \sigma^2 \right)^2 \right] &= \left(\frac{n-1}{n} \sigma^2 - \sigma^2 \right)^2 + \frac{2(n-1)\sigma^4}{n^2} \\ &= \frac{1}{n^2} \sigma^4 + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}. \end{aligned}$$

La différence $R((\mu, \sigma^2), S_n^2) - R((\mu, \sigma^2), \hat{\sigma}_n^2)$ 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 . $\diamond$

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.

Exemple I-4.18 (Estimation du support d'une loi uniforme). Soit $(X_1, X_2, \dots, X_n)$ un n -échantillon du modèle

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), (\text{Unif}([0, \theta]))_{\theta \in \Theta = \mathbb{R}_+^*}).$$

On note $X_{n:n} = \max(X_1, \dots, X_n)$. Pour tout $\theta \in \Theta$ et $x \in [0, \theta]$, nous avons

$$\mathbb{P}_\theta(X_{n:n} \leq x) = \mathbb{P}_\theta(X_1 \leq x, \dots, X_n \leq x) = (x/\theta)^n.$$

La densité de la variable $X_{n:n}$ est donc donnée pour $x \in [0, \theta]$, par

$$p_\theta(x) = \frac{1}{\theta} n \left(\frac{x}{\theta} \right)^{n-1} = n \frac{x^{n-1}}{\theta^n} \quad (\text{I-4.16})$$

d'où :

$$\begin{aligned} \mathbb{E}_\theta[X_{n:n}] &= \int_0^\theta x \cdot n \frac{x^{n-1}}{\theta^n} dx = \frac{n}{n+1} \frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1} \theta, \\ \mathbb{E}_\theta[X_{n:n}^2] &= \int_0^\theta x^2 \cdot n \frac{x^{n-1}}{\theta^n} dx = \frac{n}{n+2} \frac{\theta^{n+2}}{\theta^n} = \frac{n}{n+2} \theta^2. \end{aligned}$$

Par conséquent, l'estimateur

$$T_n^{(1)}(X_1, \dots, X_n) = \frac{n+1}{n} X_{n:n} \quad (\text{I-4.17})$$

est un estimateur sans biais du paramètre θ . Cet estimateur est toutefois inadmissible pour le risque quadratique. En effet, considérons l'estimateur $T_n^{(2)}(X_1, \dots, X_n) = a_n X_{n:n}$ pour un scalaire a_n . Calculons le risque quadratique de cet estimateur

$$\begin{aligned}\mathbb{E}_\theta[(a_n X_{n:n} - \theta)^2] &= a_n^2 \mathbb{E}_\theta[X_{n:n}^2] - 2a_n \theta \mathbb{E}_\theta[X_{n:n}] + \theta^2 \\ &= \frac{na_n^2}{n+2} \theta^2 - \frac{2a_n n}{n+1} \theta^2 + \theta^2 = \theta^2 \left\{ \frac{na_n^2}{n+2} - \frac{2a_n n}{n+1} + 1 \right\}.\end{aligned}$$

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

$$R(\theta, T_n^{(2)}) = \mathbb{E}_\theta[(T_n^{(2)}(X_1, \dots, X_n) - \theta)^2] = \frac{\theta^2}{(n+1)^2}.$$

◇

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 $g : \theta \mapsto \mathbb{R}$. Par la décomposition biais-variance (Corollaire I-4.15), le risque d'un tel estimateur vaut $R(\theta, T) = \text{Var}_\theta(T(Z))$. Cela conduit naturellement à la définition suivante :

Définition I-4.19. Soient $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique, $g : \Theta \rightarrow \mathbb{R}$, $\theta \mapsto g(\theta)$ une fonction et T, T' des estimateurs sans biais de g .

- On dit que T est uniformément meilleur que T' si $\text{Var}_\theta(T(Z)) \leq \text{Var}_\theta(T'(Z))$ 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 $F : \Theta \times Z \rightarrow \mathbb{R}$, $(\theta, z) \mapsto F(\theta, z)$ une fonction. Si pour $z \in Z$, la fonction $\theta \mapsto F(\theta, z)$ est deux fois dérivable par rapport à θ , nous notons :

$$H_F(\theta, z) := \left[\frac{\partial^2 F}{\partial \theta^{(i)} \partial \theta^{(j)}}(\theta, z) \right]_{1 \leq i, j \leq d}. \quad (\text{I-4.18})$$

Si $A = (a_{ij})_{1 \leq i, j \leq d} \in \text{Mat}_d(\mathbb{R})$, nous notons $\|A\|$ sa *norme de Frobenius*, définie par :

$$\|A\|^2 := \text{Tr}(A^T A) = \sum_{1 \leq i, j \leq d} a_{ij}^2.$$

Définition I-4.20 (Modèle régulier). Considérons un modèle paramétrique $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$, 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 $(Z, \mathcal{Z})$, avec 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 $z \in Z$:

$$\ell(\theta, z) := \log p(\theta, z) = \log p_\theta(z). \quad (\text{I-4.19})$$

- (ii) Pour μ -presque tout $z \in Z$, 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 $\mathcal{V}_0$ de θ_0 et deux fonctions mesurables positives $g, h : (Z, \mathcal{Z}) \rightarrow \mathbb{R}$ telles que

$$\int_Z \{1 + h(z)\} g(z) \mu(dz) < \infty \quad (\text{I-4.20})$$

et, pour tout $\theta \in \mathcal{V}_0$ et μ -presque tout $z \in Z$,

$$\|\nabla \ell(\theta, z)\|^2 \leq h(z), \quad \|\mathbf{H}_\ell(\theta, z)\| \leq h(z), \quad p(\theta, z) \leq g(z) \quad (\text{I-4.21})$$

- (iv) Pour tout $\theta \in \Theta$, la matrice $\mathbb{E}_\theta[\mathbf{H}_\ell(\theta, Z)]$ (bien définie par l'hypothèse (iii)) est inversible.

Lemme I-4.21. Soit $(Z, \mathcal{Z}, \{p_\theta \cdot \mu, \theta \in \Theta\})$ un modèle statistique régulier. La matrice

$$\mathbb{I}(\theta) := \mathbb{E}_\theta [\nabla \ell(\theta, Z) \nabla^T \ell(\theta, Z)] ,$$

est bien définie et ne dépend pas du choix de la mesure de domination.

Démonstration. Notons que, pour tout $a \in \mathbb{R}^d$, $\|aa^T\| = \text{Tr}(aa^T aa^T)^{1/2} = \text{Tr}(a^T aa^T a)^{1/2} = \|a\|^2$. Ainsi, pour tout $\theta_0 \in \Theta$, on a

$$\mathbb{E}_{\theta_0} [\|\nabla \ell(\theta_0, Z) \nabla \ell(\theta_0, Z)^T\|] = \mathbb{E}_{\theta_0} [\|\nabla \ell(\theta_0, Z)\|^2] \leq \int_Z h(z) g(z) \mu(dz) < \infty$$

où g, h sont comme dans l'hypothèse (iii) de régularité. Ainsi, l'intégrale (I-4.22) 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 $\lambda = \mu + \nu$. Nous avons montré (Remarque I-1.13) qu'il existe une fonction mesurable positive h telle que

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

On en déduit que $\nabla(\log p_\theta(z)) = \nabla(\log q_\theta(z))$ $\lambda - \text{p.p.}$, ce qui conclut en substituant dans (I-4.22). $\square$

Définition I-4.22 (Information de Fisher). L'information de Fisher du modèle régulier $(Z, \mathcal{Z}, \{p_\theta \cdot \mu, \theta \in \Theta\})$ est la fonction $\mathbb{I} : \theta \mapsto \mathbb{I}(\theta)$ définie par

$$\mathbb{I}(\theta) := \mathbb{E}_\theta [\nabla \ell(\theta, Z) \nabla^T \ell(\theta, Z)] . \quad (\text{I-4.22})$$

Proposition I-4.23. Soit $(Z, \mathcal{Z}, \{p_\theta \cdot \mu, \theta \in \Theta\})$ Les propriétés suivantes sont vérifiées :

- (i) Pour tout $\theta \in \Theta$,

$$\mathbb{E}_\theta [\nabla \ell(\theta, Z)] = 0 . \quad (\text{I-4.23})$$

- (ii) L'information de Fisher admet l'expression alternative suivante :

$$\mathbb{I}(\theta) = -\mathbb{E}_\theta [\mathbf{H}_\ell(\theta, Z)] . \quad (\text{I-4.24})$$

- (iii) Soient deux modèles statistiques réguliers $(Z_1, \mathcal{Z}_1, \{\mathbb{P}_{1,\theta}, \theta \in \Theta\})$ et $(Z_2, \mathcal{Z}_2, \{\mathbb{P}_{2,\theta}, \theta \in \Theta\})$ 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 :

$$(Z_1 \times Z_2, \mathcal{Z}_1 \otimes \mathcal{Z}_2, \{\mathbb{P}_{1,\theta} \otimes \mathbb{P}_{2,\theta}, \theta \in \Theta\}) ,$$

il vient pour tout $\theta \in \Theta$:

$$\mathbb{I}_{12}(\theta) = \mathbb{I}_1(\theta) + \mathbb{I}_2(\theta) . \quad (\text{I-4.25})$$

En particulier, si $(X_1, \dots, X_n)$ est un n -échantillon du modèle statistique régulier

$$\mathcal{E} = (\mathbf{X}, \mathcal{X}, \{q_\theta, \theta \in \Theta\})$$

alors, pour tout $\theta \in \Theta$ et $n \geq 1$,

$$\mathbb{I}_{\mathcal{E}^n}(\theta) = n \mathbb{I}_{\mathcal{E}}(\theta),$$

où $\mathbb{I}_{\mathcal{E}}(\theta)$ et $\mathbb{I}_{\mathcal{E}^n}(\theta)$ sont les informations de Fisher associée aux modèles $\mathcal{E}$ et $\mathcal{E}^n$.

Remarque I-4.24. Soit $(Z, \mathcal{Z}, \{p_\theta \cdot \mu, \theta \in \Theta\})$ un modèle statistique régulier. Notons que, pour tout $z \in Z$, la matrice $\nabla \ell(\theta, z) \nabla \ell(\theta, z)^T$ est symétrique et semi-définie positive. Comme

$$\mathbb{I}(\theta) = \mathbb{E}_\theta[\nabla \ell(\theta, Z) \nabla \ell(\theta, Z)^T] = \int \nabla \ell(\theta, z) \nabla \ell(\theta, z)^T \mathbb{P}_\theta(dz)$$

est symétrique et semi-définie positive pour tout $\theta \in \Theta$. Par l'égalité I-4.24 et Définition I-4.20-(iv), $\mathbb{I}(\theta)$ est de plus inversible, donc définie positive. $\diamond$

Démonstration. Commençons par noter que, pour tout $\theta \in \Theta$ et $z \in Z$,

$$\nabla \ell(\theta, z) = \frac{1}{p(\theta, z)} \nabla p(\theta, z); \quad (\text{I-4.26})$$

en différenciant à nouveau, il vient :

$$\begin{aligned} \mathbf{H}_\ell(\theta, z) &= \frac{1}{p(\theta, z)} \mathbf{H}_p(\theta, z) - \frac{1}{p^2(\theta, z)} \nabla p(\theta, z) \nabla p(\theta, z)^T \\ &= \frac{1}{p(\theta, z)} \mathbf{H}_p(\theta, z) - \nabla \ell(\theta, z) \nabla \ell(\theta, z)^T. \end{aligned} \quad (\text{I-4.27})$$

Nous pouvons alors écrire, pour tout $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}_\theta[\nabla \ell(\theta, Z)] &= \int_Z \nabla \ell(\theta, z) p(\theta, z) \mu(dz) = \int_Z \nabla p(\theta, z) \mu(dz) \\ &\stackrel{(*)}{=} \nabla \int_Z p(\theta, z) \mu(dz) = \nabla 1 = 0, \end{aligned}$$

ce qui établit le premier point, sous réserve que l'interversion $(*)$ soit justifiée. De même, on a

$$\begin{aligned} \mathbb{E}_\theta[\mathbf{H}_\ell(\theta, Z)] &= \int_Z \mathbf{H}_\ell(\theta, z) p(\theta, z) \mu(dz) = \int_Z \mathbf{H}_p(\theta, z) \mu(dz) - \int_Z \nabla \ell(\theta, z) \nabla \ell(\theta, z)^T p(\theta, z) \mu(dz) \\ &\stackrel{(**)}{=} \nabla \int_Z \nabla p(\theta, z)^T \mu(dz) - \mathbb{I}(\theta) = \nabla 0 - \mathbb{I}(\theta) = -\mathbb{I}(\theta), \end{aligned}$$

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 $\theta_0 \in \Theta$, $\mathcal{V}_0$ un voisinage de θ_0 dans Θ et g, h des fonctions comme dans l'hypothèse (iii) de régularité. On a pour tout $\theta \in \mathcal{V}_0$,

$$\|\nabla p(\theta, z)\| = \|\nabla \ell(\theta, z)\| p(\theta, z) \leq (1 + \|\nabla \ell(\theta, z)\|^2) p(\theta, z) \leq (1 + h(z)) g(z),$$

avec $\int_Z (1 + h(z)) g(z) \mu(dz) < \infty$ par (I-4.20). La proposition IV-1.43 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 $\theta \in \mathcal{V}_0$ et tout $z \in Z$, l'égalité (I-4.27) implique

$$\begin{aligned} \|\mathbf{H}_p(\theta, z)\| &\leq \|\mathbf{H}_\ell(\theta, z)\| p(\theta, z) + \|\nabla \ell(\theta, z) \nabla \ell(\theta, z)^T\| p(\theta, z) \\ &\leq (\|\mathbf{H}_\ell(\theta, z)\| + \|\nabla \ell(\theta, z) \nabla \ell(\theta, z)^T\|) p(\theta, z) \\ &\leq 2h(z)g(z) \end{aligned}$$

qui est intégrable sous μ . Comme $H_p(\theta, z) = \nabla(\nabla p(\theta, z)^T)$, la proposition IV-1.43 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 $\ell_{12}(\theta, (z_1, z_2)) = \ell_1(\theta, z_1) + \ell_2(\theta, z_2)$. $\square$

Exemple I-4.25. Soit $(X_1, \dots, X_n)$ un n -échantillon de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta =]0, 1[\})$$

Par la Proposition I-4.23, 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 $X_1 \sim \text{Ber}(\theta)$. 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-4.23 :

$$\begin{aligned} \mathbb{I}_1(\theta) &= -\mathbb{E}_\theta[\ell''(\theta, X)] = -\theta \ell''(\theta, 1) - (1-\theta) \ell''(\theta, 0) \\ &= \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

L'information de Fisher du n -échantillon de Bernoulli est donc $\mathbb{I}(\theta) = n/(\theta(1-\theta))$. $\diamond$

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-4.26 (Estimateur régulier). Soit $(Z, \mathcal{Z}, \{p_\theta \cdot \mu, \theta \in \Theta\})$, où Θ est un ouvert de $\mathbb{R}^d$, un modèle régulier. Un estimateur $T : (Z, \mathcal{Z}) \rightarrow \mathbb{R}$ d'une fonction $g : \Theta \rightarrow \mathbb{R}$ 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 $\theta \in \Theta$

$$\nabla \mathbb{E}_\theta[T(Z)] = \int_Z T(z) \nabla p(\theta, z) \mu(dz). \quad (\text{I-4.28})$$

Remarque I-4.27. Notons que, par la proposition IV-1.43 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 $\mathcal{V}_0$ de θ_0 dans Θ et une fonction mesurable positive g telle que

$$\int_Z |T(z)| g(z) \mu(dz) < \infty$$

et il existe un ensemble $Z_0 \in \mathcal{Z}$ tel que $\mu(Z_0^c) = 0$ pour tout $z \in Z_0$ la fonction $\theta \mapsto p(\theta, z)$ est différentiable sur Θ et pour tout $\theta \in \mathcal{V}_0$ et $z \in Z_0$,

$$\|\nabla p(\theta, z)\| \leq g(z). \quad \diamond$$

Théorème I-4.28 (Cramér-Rao). Soit $(Z, \mathcal{Z}, \{p_\theta \cdot \mu, \theta \in \Theta\})$ où Θ est un ouvert de $\mathbb{R}^d$, un modèle régulier, $g : \Theta \rightarrow \mathbb{R}$ une fonction dérivable à valeurs réelles, et T un estimateur sans biais et régulier de $g : \theta \mapsto g(\theta)$. Pour tout $\theta \in \Theta$,

$$\text{Var}_\theta(T(Z)) \geq \nabla g(\theta)^T \mathbb{I}(\theta)^{-1} \nabla g(\theta). \quad (\text{I-4.29})$$

Démonstration. $T(Z)$ étant un estimateur sans biais et régulier, il vient :

$$\begin{aligned}\nabla g(\theta) &= \nabla \mathbb{E}_\theta T(Z) = \int T(z) \nabla p(\theta, z) \mu(dz) \\ &= \int T(z) \nabla \ell(\theta, z) p(\theta, z) \mu(dz) = \mathbb{E}_\theta [T(Z) \nabla \ell(\theta, Z)] .\end{aligned}$$

Or $\mathbb{E}_\theta [\nabla \ell(\theta, Z)] = 0$ (Proposition I-4.23), 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 $u \in \mathbb{R}^d$,

$$\begin{aligned}\{\nabla g(\theta)^T u\}^2 &= \mathbb{E}_\theta [\{T(Z) - g(\theta)\} \nabla \ell(\theta, Z)^T u]^2 \\ &\leq \mathbb{E}_\theta [\{T(Z) - g(\theta)\}^2] \mathbb{E}_\theta [\{\nabla \ell(\theta, Z)^T u\}^2]\end{aligned}$$

par l'inégalité de Cauchy-Schwarz. Or, pour tout $a \in \mathbb{R}^d$,

$$\{a^T u\}^2 = u^T a a^T u = u^T (a a^T) u ,$$

donc en prenant respectivement $a = \nabla g(\theta)$ et $a = \nabla \ell(\theta, Z)$, l'inégalité précédente devient :

$$\begin{aligned}u^T \nabla g(\theta) \nabla g(\theta)^T u &\leq \mathbb{E}_\theta [(T(Z) - g(\theta))^2] \mathbb{E}_\theta [u^T \nabla \ell(\theta, Z) \nabla \ell(\theta, Z)^T u] \\ &= \text{Var}_\theta [T(Z)] u^T \mathbb{I}(\theta) u\end{aligned}$$

pour tout $u \in \mathbb{R}^d$. En particulier, en prenant $u = \mathbb{I}(\theta)^{-1} \nabla g(\theta)$, l'inégalité précédente devient

$$(\nabla g(\theta)^T \mathbb{I}(\theta)^{-1} \nabla g(\theta))^2 \leq \text{Var}_\theta [T(Z)] (\nabla g(\theta)^T \mathbb{I}(\theta)^{-1} \nabla g(\theta)) ,$$

ce qui établit l'inégalité (I-4.29) $(\nabla g(\theta)^T \mathbb{I}(\theta)^{-1} \nabla g(\theta))$ est positif car $\mathbb{I}(\theta)$ est positive ; si ce terme est nul l'inégalité (I-4.29) est triviale, sinon on simplifie ce terme strictement positif). $\square$

Une conséquence importante du théorème I-4.28 est la suivante :

Corollaire I-4.29. Soit $(Z, \mathcal{Z}, \{p_\theta \cdot \mu, \theta \in \Theta\})$ 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 $\theta \in \Theta$,

$$\text{Var}_\theta (T(Z)) \geq \mathbb{I}(\theta)^{-1} \tag{I-4.30}$$

i.e. la matrice symétrique $\text{Var}_\theta (T(Z)) - \mathbb{I}(\theta)^{-1}$ 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-4.28 et en notant que $\nabla g(\theta) = \lambda$ et $\text{Var}_\theta (\lambda^T T(Z)) = \lambda^T \text{Var}_\theta (T(Z)) \lambda$, il vient :

$$\lambda^T \text{Var}_\theta (T(Z)) \lambda \geq \lambda^T \mathbb{I}(\theta)^{-1} \lambda .$$

Cette inégalité étant valide pour tout $\lambda \in \mathbb{R}^d$, il vient $\text{Var}_\theta (T(Z)) \geq \mathbb{I}(\theta)^{-1}$ comme voulu. $\square$

Du théorème I-4.28 de Cramér-Rao et de l'additivité de l'information de Fisher (Proposition I-4.23), on déduit immédiatement le corollaire suivant :

Corollaire I-4.30. Soient $(X, \mathcal{X})$ un espace mesurable et μ une mesure σ -finie sur $(X, \mathcal{X})$. Soient $(X_1, \dots, X_n)$ un n -échantillon du modèle régulier

$$(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$$

et $T_n(X_1, \dots, X_n)$ un estimateur sans biais et régulier de la fonction $g : \Theta \rightarrow \mathbb{R}$. En notant

$$\mathbb{I}_1(\theta) = \mathbb{E}_\theta [\nabla \ell(\theta, X_1) \nabla \ell(\theta, X_1)^T], \quad \ell(\theta, x) = \log q_\theta(x),$$

l'information de Fisher associée à une seule observation, on a pour tout $\theta \in \Theta$:

$$\text{Var}_\theta(T_n(X_1, \dots, X_n)) \geq \frac{1}{n} \nabla g(\theta)^T \mathbb{I}_1(\theta)^{-1} \nabla g(\theta). \quad (\text{I-4.31})$$

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-4.30 montre que, pour un modèle régulier, la variance (et donc le risque quadratique) d'un estimateur sans biais est au moins d'ordre n^{-1} , ce qui signifie que la différence $T(Z) - g(\theta)$ est au moins d'ordre $n^{-1/2}$.

Exemple I-4.31. La variance des estimateurs sans biais $\bar{X}_n$ de l'espérance $\mu(\theta) = \mathbb{E}_\theta[Z]$ (exemple I-4.11, 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 $\sigma^2(\theta) = \text{Var}_\theta[Z]$ (exemple I-4.17, 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-4.18), 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. En effet, la densité $p(\theta, x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^+$ de la loi uniforme sur $[0, \theta]$ est égale à $p(\theta, x) = \mathbb{1}(x \leq \theta)$; pour tout $x > 0$, la fonction $\theta \mapsto p(\theta, x)$ est discontinue en x . $\diamond$

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 $\nabla g(\theta)^T \mathbb{I}_1(\theta)^{-1} \nabla g(\theta)$ 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-4.32 (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-4.25 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

$$\text{Var}_\theta(\bar{X}_n) = \frac{\text{Var}_\theta(X_1)}{n} = \frac{\theta(1-\theta)}{n};$$

cet estimateur est par conséquent E.S.B.V.M.. $\diamond$

Exemple I-4.33 (Estimation de la moyenne d'un échantillon gaussien). Soit $(X_1, \dots, X_n)$ un n -échantillon d'un modèle gaussien

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\theta, \sigma_0^2), \theta \in \mathbb{R}\}),$$

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

$$\text{Var}_\theta(\bar{X}_n) = \frac{\sigma_0^2}{n}.$$

Par conséquent $\text{Var}_\theta(\bar{X}_n) = \mathbb{I}(\theta)^{-1}$, l'estimateur atteint la borne de Cramér-Rao. $\diamond$

Définition I-4.34 (Efficacité). Soit $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle régulier, et $g : \Theta \rightarrow \mathbb{R}$ 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

$$\text{Var}_\theta(T(Z)) = \nabla g(\theta)^T \mathbb{I}(\theta)^{-1} \nabla g(\theta) \quad \text{pour tout } \theta \in \Theta. \quad (\text{I-4.32})$$

Notons qu'un estimateur sans biais efficace est E.S.B.V.M. (au sens de la définition I-4.19), bien que la réciproque soit fautive en général.

Chapitre I-5

Optimalité en théorie des tests

L'objectif du chapitre est de présenter deux notions d'optimalité en théorie des tests. La première consiste, parmi les tests de niveau α , à rechercher ceux dont la puissance est la plus grande possible. Ce problème est résolu dans deux cas simples par le test du rapport de vraisemblance, mais est difficile dans la majorité des situations où il n'existe pas de tests optimaux en ce sens.

I-5.1 Tests uniformément plus puissants

On considère un modèle statistique $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$. 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 $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique et $\phi : Z \rightarrow [0, 1]$ 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 $Z \in Z$, 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 l'hypothèse alternative 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 zone de rejet et une région d'acceptation, un test randomisé fait une partition en 3 ensembles.

- L'ensemble $\mathcal{Z}_1 = \{z \in Z : \phi(z) = 1\}$ est la région dans laquelle le test rejette H_0 .
- L'ensemble $\mathcal{Z}_0 = \{z \in Z : \phi(z) = 0\}$ est la région dans laquelle le test accepte H_0 .
- L'ensemble $\mathcal{Z}_p = \{z \in Z : 0 < \phi(z) < 1\}$ 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 d'erreur de première espèce, d'erreur de seconde espèce, de taille et de niveau.

Définition I-5.2 (Taille et niveau d'un test). La taille d'un test de fonction critique ϕ (ou risque de première espèce) 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 $\mathcal{K}_\alpha(\Theta_0)$ 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 , il suffit en effet de considérer le test randomisé tel que $\phi(z) = \alpha$ pour toute valeur de $z \in Z$.

Exemple I-5.3. Soit $(X_1, \dots, X_n)$ un n -échantillon du modèle de Poisson

$$(\mathbb{N}, \mathcal{P}(\mathbb{N}), \{p_\theta \cdot \mu, \theta \in \Theta \subset \mathbb{R}_+\})$$

où μ est la mesure de comptage sur $\mathbb{N}$ (voir Section IV-1.3.4) et

$$p_\theta(x) = p(\theta, x) = e^{-\theta} \frac{\theta^x}{x!}, \quad \theta \in \Theta = \mathbb{R}_+^*, \quad x \in \mathbb{N}.$$

Supposons que $\Theta = \{\theta_0\} \cup \{\theta_1\}$ où $\theta_0 \neq \theta_1$. Nous considérons le test d'hypothèses simples

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1$$

La log-vraisemblance des observations est donnée par

$$\ell_n(\theta; X_1, \dots, X_n) = -n\theta + \log(\theta)S_n - \sum_{i=1}^n \log(X_i!)$$

où $S_n = \sum_{i=1}^n X_i$. L'estimateur du maximum de vraisemblance du paramètre θ est donné par

$$\hat{\theta}_n = n^{-1} S_n .$$

Rappelons que si les variables aléatoires $X_1, \dots, X_n$ sont indépendantes et distribuées suivant des lois de Poisson de paramètre $\theta \in \mathbb{R}_+^*$, alors la variable $S_n = \sum_{i=1}^n X_i$ est distribuée suivant une loi de Poisson de paramètre $n\theta$.

Nous considérons le test pur de seuil critique c , $\delta(X_1, \dots, X_n; c) = \mathbb{1}_{\{S_n > c\}}$. Nous obtenons un test de niveau $\alpha \in]0, 1[$ en choisissant le seuil critique c_α .

$$c_\alpha = \inf \{c \in \mathbb{N} : \mathbb{P}_{\theta_0}(S_n > c) \leq \alpha\} .$$

Comme la variable S_n est à valeurs entières, $\mathbb{P}_{\theta_0}(S_n > c_\alpha) < \alpha$ 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 ϕ_γ définie de la façon suivante :

$$\phi_\gamma(s) = \begin{cases} 1 & \text{si } s > c_\alpha \\ \gamma & \text{si } s = c_\alpha \\ 0 & \text{si } s \leq c_\alpha - 1 \end{cases}$$

La taille du test de fonction critique ϕ_γ est donnée par

$$\bar{\alpha}(\phi_\gamma) = \mathbb{E}_{\theta_0}[\phi_\gamma(S_n)] = \mathbb{P}_{\theta_0}(S_n > c_\alpha) + \gamma \mathbb{P}_{\theta_0}(S_n = c_\alpha)$$

Si nous choisissons

$$\begin{aligned} 0 \leq \gamma &= \frac{\alpha - \mathbb{P}_{\theta_0}(S_n > c_\alpha)}{\mathbb{P}_{\theta_0}(S_n = c_\alpha)} \\ &\leq \frac{\mathbb{P}_{\theta_0}(S_n > c_\alpha - 1) - \mathbb{P}_{\theta_0}(S_n > c_\alpha)}{\mathbb{P}_{\theta_0}(S_n = c_\alpha)} \leq 1 , \end{aligned}$$

nous obtenons un test dont la taille est exactement égale à α . Pour $\alpha = 0.05$, cela revient à choisir $\gamma \approx 0,5$. $\diamond$

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. $\phi \in \mathcal{K}_\alpha(\Theta_0)$ et si sa fonction puissance β_ϕ vérifie pour tout $\theta \in \Theta_1$,*

$$\beta_\phi(\theta) = \sup_{\psi \in \mathcal{K}_\alpha(\Theta_0)} \beta_\psi(\theta) .$$

I-5.1.2 Le Théorème de Neyman-Pearson

S'il n'existe pas, en général, de test uniformément plus puissant, nous allons démontrer dans cette partie qu'il est possible de construire de tels tests dans des cas particuliers importants en pratique.

Nous allons tout d'abord nous intéresser au cas d'un test d'hypothèse simple.

Théorème I-5.5 (Neyman-Pearson). *Soit ν une mesure σ -finie sur un espace mesurable $(Z, \mathcal{Z})$. Considérons le modèle*

$$(Z, \mathcal{Z}, \{p_\theta \cdot \nu, \theta \in \Theta = \{\theta_0, \theta_1\}\})$$

où $\theta_0 \neq \theta_1$.

Pour tout $\alpha \in]0, 1[$, il existe des constantes $c_\alpha > 0$ et $\gamma_\alpha \in [0, 1]$, telles que la fonction critique :

$$\phi^*(z) = \begin{cases} 1 & \text{si } p_{\theta_1}(z) > c_\alpha p_{\theta_0}(z) , \\ \gamma_\alpha & \text{si } p_{\theta_1}(z) = c_\alpha p_{\theta_0}(z) , \\ 0 & \text{si } p_{\theta_1}(z) < c_\alpha p_{\theta_0}(z) , \end{cases} \quad (\text{I-5.1})$$

vérifie

$$\mathbb{E}_{\theta_0}[\phi^*(Z)] = \int \phi^*(z) p_{\theta_0}(z) \nu(dz) = \alpha .$$

Le test de fonction critique ϕ^* est U.P.P.(α) et sa puissance est supérieure ou égale à α ,

$$\mathbb{E}_{\theta_1}[\phi^*(Z)] = \int \phi^*(z) p_{\theta_1}(z) \nu(dz) \geq \alpha .$$

De plus, si ϕ^{**} est U.P.P.(α) alors, pour ν -presque tout $z \in Z$,

$$\phi^{**}(z) = \begin{cases} 1 & \text{si } p_{\theta_1}(z) > c_\alpha p_{\theta_0}(z) , \\ 0 & \text{si } p_{\theta_1}(z) < c_\alpha p_{\theta_0}(z) . \end{cases} \quad (\text{I-5.2})$$

Démonstration. Pour $c \geq 0$ et $\gamma \in [0, 1]$, considérons le test randomisé

$$\phi_{c,\gamma}(z) = \begin{cases} 1 & \text{si } p_{\theta_1}(z) > c p_{\theta_0}(z) , \\ \gamma & \text{si } p_{\theta_1}(z) = c p_{\theta_0}(z) , \\ 0 & \text{si } p_{\theta_1}(z) < c p_{\theta_0}(z) , \end{cases} \quad (\text{I-5.3})$$

Nous allons tout d'abord montrer que nous pouvons choisir les constantes c et γ de telle sorte que :

$$\mathbb{E}_{\theta_0}[\phi_{c,\gamma}(Z)] = \mathbb{P}_{\theta_0}(r(Z) > c) + \gamma \mathbb{P}_{\theta_0}(r(Z) = c) = \alpha , \quad (\text{I-5.4})$$

admet toujours une solution. La fonction $\bar{F}_r : \mathbb{R} \rightarrow [0, 1]$,

$$c \mapsto \bar{F}_r(c) := \mathbb{P}_{\theta_0}(r(Z) > c) = 1 - \mathbb{P}_{\theta_0}(r(Z) \leq c)$$

est décroissante sur $\mathbb{R}$, continue à droite et admet des limites à gauche, i.e. en tout point $c_0 \in \mathbb{R}$:

$$\bar{F}_r(c_0) = \lim_{c \downarrow c_0} \mathbb{P}_{\theta_0}(r(Z) > c) = \mathbb{P}_{\theta_0}(r(Z) > c_0) \quad \text{et} \quad \bar{F}_r(c_0-) = \lim_{c \uparrow c_0} \mathbb{P}_{\theta_0}(r(Z) > c) = \mathbb{P}_{\theta_0}(r(Z) \geq c_0).$$

Notons que

$$\bar{F}_r(c_0-) - \bar{F}_r(c_0) = \mathbb{P}_{\theta_0}(r(Z) = c_0) .$$

De plus, $\bar{F}_r(0-) = 0$ et $\lim_{c \rightarrow \infty} \bar{F}_r(c) = 0$. Pour tout $\alpha \in]0, 1[$, il existe donc $c_\alpha \geq 0$ tel que

$$\bar{F}_r(c_\alpha) = \mathbb{P}_{\theta_0}(r(Z) > c_\alpha) \leq \alpha \leq \bar{F}_r(c_\alpha-) = \mathbb{P}_{\theta_0}(r(Z) \geq c_\alpha) .$$

Pour obtenir l'équation (I-5.4), nous posons

$$\gamma_\alpha = \begin{cases} 0 & \text{si } \mathbb{P}_{\theta_0}(r(Z) > c_\alpha) = \alpha \\ \frac{\alpha - \mathbb{P}_{\theta_0}(r(Z) > c_\alpha)}{\mathbb{P}_{\theta_0}(r(Z) = c_\alpha)} & \text{si } \mathbb{P}_{\theta_0}(r(Z) > c_\alpha) < \alpha . \end{cases}$$

Notons que si $\mathbb{P}_{\theta_0}(r(Z) > c_\alpha) < \alpha$ nous avons

$$\begin{aligned} \mathbb{P}_{\theta_0}(r(Z) = c_\alpha) &= \mathbb{P}_{\theta_0}(r(Z) \geq c_\alpha) - \mathbb{P}_{\theta_0}(r(Z) > c_\alpha) \\ &\geq \alpha - \mathbb{P}_{\theta_0}(r(Z) > c_\alpha) > 0 \end{aligned}$$

et donc que $\gamma_\alpha \in [0, 1]$.

Nous allons maintenant démontrer que ϕ^* est U.P.P.(α). Soit ϕ un test de niveau α , $\mathbb{E}_{\theta_0}[\phi(Z)] \leq \alpha$.

- Si $\phi^*(z) - \phi(z) > 0$, alors $\phi^*(z) > 0$ et donc $p_{\theta_1}(z) \geq c_\alpha p_{\theta_0}(z)$.
- Si $\phi^*(z) - \phi(z) < 0$, alors $\phi^*(z) < 1$ et donc $p_{\theta_1}(z) \leq c_\alpha p_{\theta_0}(z)$.

Dans tous les cas, pour tout $z \in Z$,

$$\{\phi^*(z) - \phi(z)\} \{p_{\theta_1}(z) - c_\alpha p_{\theta_0}(z)\} \geq 0.$$

(I-5.5)

Risque !de
première
espèce
Risque !de
deuxième
espèce
Puissance d'un
test

Par conséquent, on a

$$\int \{\phi^*(z) - \phi(z)\} \{p_{\theta_1}(z) - c_\alpha p_{\theta_0}(z)\} \nu(dz) \geq 0,$$

ce qui peut se réécrire

$$\begin{aligned} \int \{\phi^*(z) - \phi(z)\} p_{\theta_1}(z) \nu(dz) &= \mathbb{E}_{\theta_1}[\phi^*(Z)] - \mathbb{E}_{\theta_1}[\phi(Z)] \\ &\geq c_\alpha \int \{\phi^*(z) - \phi(z)\} p_{\theta_0}(z) \nu(dz) = c_\alpha (\mathbb{P}_{\theta_0}[\phi^*(Z)] - \mathbb{P}_{\theta_0}[\phi(Z)]) . \end{aligned}$$

Comme le test ϕ est de niveau α ,

$$\mathbb{P}_{\theta_0}[\phi^*(Z)] - \mathbb{P}_{\theta_0}[\phi(Z)] = \alpha - \mathbb{P}_{\theta_0}[\phi(Z)] \geq 0$$

ce qui prouve que ϕ^* est U.P.P. (α) .

Considérons le test randomisé de fonction critique constante $\psi(x) = \alpha$. Le risque de première espèce de ce test est égal à α et le risque de deuxième espèce est égal à $(1 - \alpha)$. Comme le test de fonction critique ϕ^* défini ci-dessus est U.P.P. (α) , nous avons : $\mathbb{E}_{\theta_1}[\phi^*(Z)] \geq \alpha$.

Soit ϕ^{**} la fonction critique d'un test U.P.P. (α) . Dans ce cas, nous avons

$$\begin{aligned} \mathbb{E}_{\theta_1}[\phi^*(Z)] - \mathbb{E}_{\theta_1}[\phi^{**}(Z)] &= 0 \\ c_\alpha \{\mathbb{P}_{\theta_0}[\phi^*(Z)] - \mathbb{P}_{\theta_0}[\phi^{**}(Z)]\} &= c_\alpha \{\alpha - \mathbb{P}_{\theta_0}[\phi^{**}(Z)]\} \geq 0, \end{aligned}$$

ce qui implique

$$\int \{\phi^*(z) - \phi^{**}(z)\} p_{\theta_1}(z) \nu(dz) \leq c_\alpha \int \{\phi^*(z) - \phi^{**}(z)\} p_{\theta_0}(z) \nu(dz)$$

et donc

$$\int \{\phi^*(z) - \phi^{**}(z)\} \{p_{\theta_1}(z) - c_\alpha p_{\theta_0}(z)\} \nu(dz) \leq 0.$$

Comme par ailleurs (I-5.5) est valide pour $\phi = \phi^{**}$, on obtient que

$$\{z \in Z : \{\phi^*(z) - \phi^{**}(z)\} \{p_{\theta_1}(z) - c_\alpha p_{\theta_0}(z)\} > 0\}$$

est ν -négligeable, ce qui conclut la preuve. $\square$

Remarque I-5.6. Remarquons que, si la loi du rapport de vraisemblance $r(Z)$ n'a pas d'atomes sous $\mathbb{P}_{\theta_0}$, c'est-à-dire si $\mathbb{P}_{\theta_0}(r(Z) = c) = 0$ 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. $\diamond$

Exemple I-5.7 (Deux variables gaussiennes scalaires). Considérons un modèle statistique gaussien

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \Theta = \{\theta_0\} \cup \theta_1\})$$

Notons p_θ la densité de la loi gaussienne de paramètre $\theta = (\mu, \sigma^2)$,

$$p_{\theta_i}(z) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-(z - \mu_i)^2 / 2\sigma_i^2)$$

La région critique du test est donnée par :

$$-\frac{1}{2\sigma_1^2}(z - \mu_1)^2 + \frac{1}{2\sigma_0^2}(z - \mu_0)^2 \geq \log(c_\alpha) + (1/2)\log(\sigma_1^2/\sigma_0^2).$$

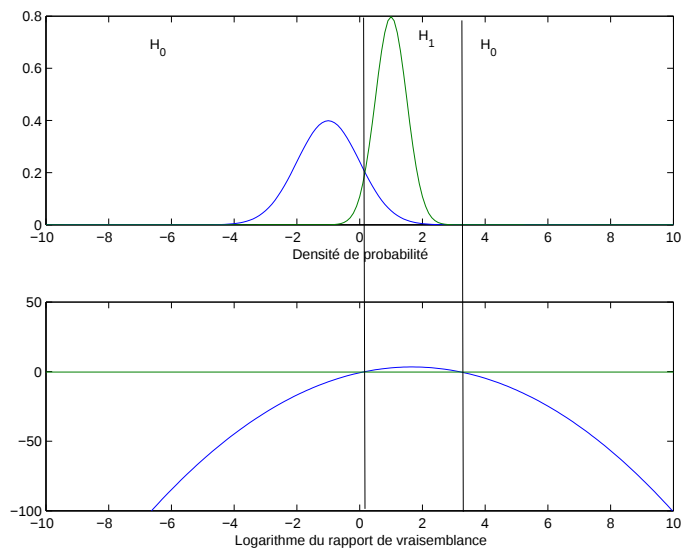


FIGURE I-5.1 – Figure du haut : densité de probabilité de deux variables aléatoires gaussiennes de moyenne et de variance $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 0.5)$. Figure du bas : rapport de vraisemblance $z \mapsto p_1(z)/p_0(z)$.

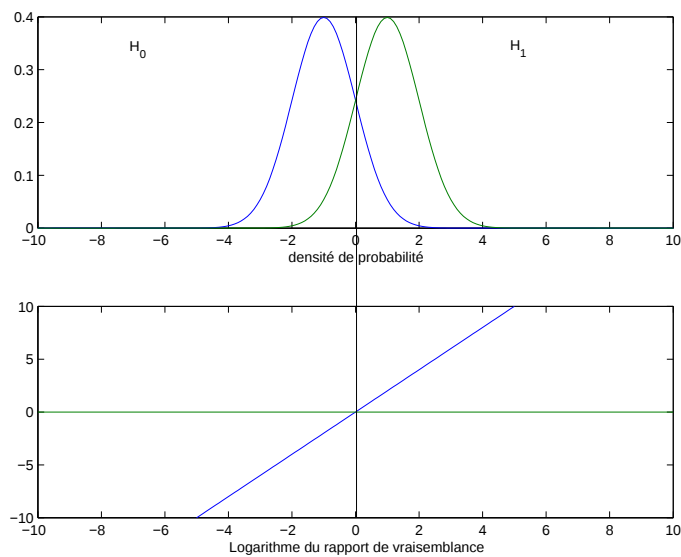


FIGURE I-5.2 – Panneau du haut : densité de probabilité de deux v.a. gaussiennes de moyenne et de variance $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 1)$. Panneau du bas : rapport de vraisemblance $z \mapsto p_1(z)/p_0(z)$.

Nous avons représenté dans la figure I-5.1 les régions critiques de ce test lorsque $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 0.5)$. Dans la figure I-5.2, nous avons visualisé les régions critiques du test lorsque $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 1)$, la variance est identique sous les deux alternatives. 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-5.8 (Test de la moyenne de variables aléatoires gaussiennes : variance connue). Soit $(X_1, \dots, X_n)$ un n -échantillon d'un modèle gaussien

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\mu, \sigma^2), \mu \in \{0, m\}\})$$

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 U.P.P. (α) . Nous formons le rapport de vraisemblance,

$$\begin{aligned} r(x_1, \dots, x_n) &= \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2 + \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right), \\ &= \exp \left(\frac{nm}{\sigma^2} \bar{x} - \frac{nm^2}{2\sigma^2} \right), \end{aligned}$$

où $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ et donc

$$r(x_1, \dots, x_n) = \tilde{r}(\bar{x}), \quad \tilde{r}(t) = \exp \left(\frac{nm}{\sigma^2} t - \frac{nm^2}{2\sigma^2} \right)$$

dépend uniquement de la statistique $\bar{x}$. On remarque que la fonction $t \rightarrow \tilde{r}(t)$ est une fonction strictement monotone de t , croissante si $m > 0$ et décroissante dans le cas contraire. Si $m \geq 0$, la condition $\tilde{r}(\bar{x}) > c_\alpha$ est équivalente à $\bar{x} > d_\alpha$. Pour déterminer le seuil d_α , nous devons résoudre l'équation :

$$\mathbb{P}_0(\bar{X} > d_\alpha) = \alpha. \quad (\text{I-5.6})$$

La loi image de $\mathbb{P}_0$ par la variable aléatoire $\sqrt{n}\bar{X}/\sigma$ est une loi gaussienne de moyenne nulle et de variance unité. L'équation (I-5.6) admet comme seule solution $d_\alpha = z_{1-\alpha}\sigma/\sqrt{n}$ où $z_{1-\alpha}$ est le quantile $1 - \alpha$ de la loi $N(0, 1)$. 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 :

$$\mathbb{P}_m(\bar{X} > z_{1-\alpha}\sigma/\sqrt{n}) = 1 - \Phi(z_{1-\alpha} - \sqrt{n}m/\sigma).$$

Nous avons visualisé dans la figure I-5.3 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 $X_1, \dots, X_n$ sont indépendantes de loi gaussienne de moyenne $m_1, \dots, m_n$ et de variance unité. Dans ce cas particulier, le rapport de vraisemblance vaut

$$r(x_1, \dots, x_n) = \exp \left(\frac{1}{\sigma^2} \sum_{i=1}^n m_i X_i - \frac{1}{2\sigma^2} \sum_{i=1}^n m_i^2 \right).$$

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 la loi image de $\mathbb{P}_0$ par la variable aléatoire

$$\frac{\sum_{i=1}^n m_i X_i}{\sigma \sqrt{\sum_{i=1}^n m_i^2}}$$

est une loi gaussienne centrée réduite, 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.

◇

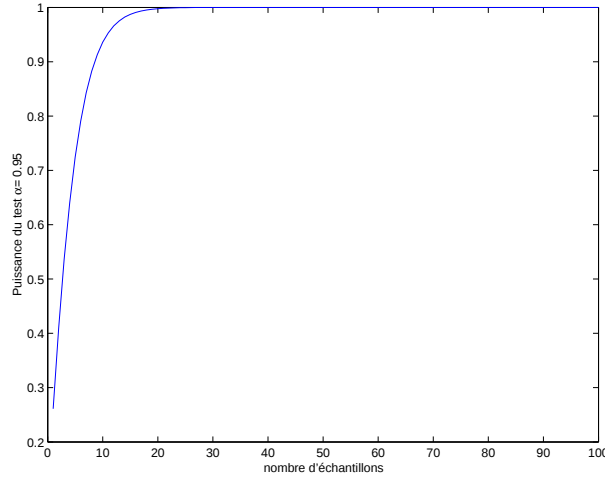


FIGURE I-5.3 – Puissance du test U.P.P.(α) de $H_0 = \{\mu = 0\}$ contre $H_1 = \{\mu = 1\}$ de niveau $\alpha = 0,05$ en fonction de la taille de l'échantillon.

Exemple I-5.9 (Variance d'une gaussienne : moyenne connue). Soit $(X_1, \dots, X_n)$ un n -échantillon de v.a. gaussiennes $N(0, \theta)$. 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 :

$$r(x_1, \dots, x_n) = \left(\frac{\theta_0}{\theta_1}\right)^{n/2} \exp\left(-\left(\frac{1}{2\theta_1} - \frac{1}{2\theta_0}\right) \sum_{i=1}^n x_i^2\right).$$

La condition $r(x_1, \dots, x_n) > c_\alpha$ 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 $\sum_{i=1}^n X_i^2 / \theta_0$ 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. $\diamond$

Exemple I-5.10 (Sondage). Soit $(X_1, \dots, X_n)$ un n -échantillon de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \{\theta_0\} \cup \{\theta_1\}\}) .$$

Considérons le test

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1$$

où $0 \leq \theta_0 < \theta_1 \leq 1$. 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

$$\phi^*(Z) := \begin{cases} 1 & \text{si } \tilde{r}(S(Z)) > c_\alpha \\ \gamma & \text{si } \tilde{r}(S(Z)) = c_\alpha \\ 0 & \text{si } \tilde{r}(S(Z)) < c_\alpha \end{cases}$$

soit U.P.P. (α). La fonction $s \mapsto \tilde{r}(s)$ est monotone en s , ce qui implique que le test précédent peut s'écrire

$$\phi^*(Z) := \begin{cases} 1 & \text{si } S(Z) > m_\alpha \\ \gamma & \text{si } S(Z) = m_\alpha \\ 0 & \text{si } S(Z) < m_\alpha \end{cases}.$$

Hypothèse !unilatérale
Test statis-
tique !\UPP\

Rapport de
vraisem-
blance !monotone

Pour que le test soit de niveau α , les constantes $m_\alpha \in \mathbb{N}$ et γ doivent vérifier l'équation

$$\alpha = \mathbb{P}_{\theta_0}[\phi^*(Z)] = \mathbb{P}_{\theta_0}(S(Z) > m_\alpha) + \gamma \mathbb{P}_{\theta_0}(S(Z) = m_\alpha).$$

Comme $S(Z)$ est distribuée suivant une loi binômiale de paramètre θ sous $\mathbb{P}_0$, 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. $\diamond$

I-5.2 Rapport de vraisemblance monotone

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'hypothèse de base est *unilatérale*. Plus précisément, considérons un modèle statistique $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ où $\Theta \subseteq \mathbb{R}$. Un test d'hypothèse est dit unilatéral si

$$\begin{aligned} &H_0 : \theta \leq \theta_0, \quad \text{contre} \quad H_1 : \theta > \theta_0 \\ \text{ou} \quad &H_0 : \theta \geq \theta_0, \quad \text{contre} \quad H_1 : \theta < \theta_0 \end{aligned}$$

où $\theta_0 \in \Theta$. De façon générale, le test le plus puissant de l'hypothèse H_0 contre l'alternative $\{\theta = \theta_1\}$, avec $\theta_1 > \theta_0$ dépend de la valeur de θ_1 , et on ne sait pas construire de test uniformément plus puissant de l'hypothèse

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

Nous allons voir toutefois qu'il existe des tests U.P.P. lorsque l'on impose une hypothèse supplémentaire sur la structure statistique du modèle.

H I-5.11 (Famille à rapport de vraisemblance monotone). Soient μ une mesure σ -finie sur $(Z, \mathcal{Z})$ et $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique où $\Theta \subseteq \mathbb{R}$. La famille est à *rapport de vraisemblance monotone* si

1. il existe $\psi : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ et $T : Z \rightarrow \mathbb{R}$ et $h : Z \rightarrow \mathbb{R}^+$ mesurables telles que

$$p(\theta, z) = p_\theta(z) = h(z) \psi(T(z); \theta), \quad (\text{I-5.7})$$

où $z \mapsto p(\theta, z)$ est la densité de $\mathbb{P}_\theta^Z$.

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 . $\diamond$

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 car nous pouvons toujours changer $T(z)$ en $-T(z)$ dans (I-5.7).

Exemple I-5.12 (Loi gaussienne $N(\theta, 1)$). Soit $Z = (X_1, \dots, X_n)$ un n -échantillon gaussien

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\theta, 1), \theta \in \mathbb{R}\}) .$$

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-5.11 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-5.13 (Loi gaussienne $N(0, \theta^{-2})$). Soit $(X_1, \dots, X_n)$ un n -échantillon de gaussien

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(0, \theta^{-2}), \theta \in \Theta = \mathbb{R}_+^*\}) .$$

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 $\mathbb{R}^n$ de vraisemblance et en posant $z = (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$p(\theta, z) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\theta^2}{2} \sum_{i=1}^n x_i^2 + n \log(\theta)\right) .$$

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-5.14 (Loi binômiale). Soient $(X_1, \dots, X_n)$ un n -échantillon de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta = [0, 1]\}) .$$

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-5.11 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 $(X_1, \dots, X_n)$ est un n -échantillon d'une famille exponentielle (Chapitre IV-6) associée à la paire (h, T) , où T est une statistique scalaire, c'est-à-dire que la densité de X_1 sous $\mathbb{P}_\theta$ s'écrit

$$p(\theta, x) = h(x) \exp(\phi(\theta)T(x) - \psi(\theta)) .$$

Un tel modèle est à rapports de vraisemblance monotones en $T(Z)$ si et seulement si la fonction $\theta \rightarrow \phi(\theta)$ est monotone. Par exemple, si la fonction $\theta \rightarrow \phi(\theta)$ est strictement croissante, alors le modèle est à rapports de vraisemblance croissants en $T(Z)$. Si la fonction $\theta \rightarrow \phi(\theta)$ est strictement décroissante, le modèle est à rapports de vraisemblance croissants en $-T(Z)$.

Remarquons que **H I-5.11** implique que, pour tout $\theta > \theta'$ et tout d , la condition $p(\theta, z)/p(\theta', z) \geq d$ s'écrit de manière équivalente $T(z) \geq c(\theta, \theta', d)$.

Le lemme suivant est utile pour étudier la propriété U.P.P. des tests sur les familles à rapports de vraisemblance monotones.

Lemme I-5.15. *Supposons **H I-5.11**. Soit $\phi : \mathbb{R} \rightarrow \mathbb{R}$ une fonction monotone croissante (au sens large) telle que, pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[|\phi \circ T(Z)|] < \infty$. Alors la fonction*

$$g : \theta \mapsto g(\theta) = \mathbb{E}_\theta[\phi \circ T(Z)]$$

est une fonction croissante (au sens large)

Démonstration. Soit $\theta' < \theta$. Sous **H I-5.11**, 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. Nous avons donc :

$$A := \{z \in Z : p(\theta', z) > p(\theta, z)\} = \{z \in Z : T(z) < d\}$$

où $d = \sup \{t \in \mathbb{R} : \psi(t; \theta)/\psi(t; \theta') < 1\}$. Par hypothèse, la fonction ϕ est monotone croissante ; par conséquent,

$$b := \inf_{z \in A^c} \phi \circ T(z) \geq a := \sup_{z \in A} \phi \circ T(z)$$

et

$$\begin{aligned} g(\theta) - g(\theta') &= \int \phi \circ T(z) \{p(\theta, z) - p(\theta', z)\} \mu(dz) \\ &\geq a \int_A \{p(\theta, z) - p(\theta', z)\} \mu(dz) + b \int_{A^c} \{p(\theta, z) - p(\theta', z)\} \mu(dz) \\ &= (b - a) \int_{A^c} \{p(\theta, z) - p(\theta', z)\} \mu(dz) \geq 0 . \end{aligned}$$

□

Théorème I-5.16. *Supposons **H I-5.11**. Soient $\theta_0 \in \Theta$ et $\alpha \in (0, 1)$. Alors :*

(i) Il existe un test U.P.P. (α) de

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

La fonction critique de ce test est donnée par :

$$\phi(z) = \begin{cases} 1 & \text{si } T(z) > c, \\ \gamma & \text{si } T(z) = c, \\ 0 & \text{si } T(z) < c, \end{cases} \quad (\text{I-5.8})$$

où les constantes c et γ sont solutions de l'équation :

$$\mathbb{E}_{\theta_0}[\phi(Z)] = \mathbb{P}_{\theta_0}(T(Z) > c) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c) = \alpha .$$

(ii) La puissance du test

$$\theta \mapsto \beta_\phi(\theta) := \mathbb{E}_\theta[\phi(Z)]$$

est une fonction strictement croissante sur l'ensemble $\{\theta \in \Theta : \beta_\phi(\theta) < 1\}$.

Démonstration. Considérons tout d'abord pour $\theta_1 > \theta_0$ le test d'hypothèses simple

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1, \quad (\text{I-5.9})$$

Posons

$$r_{\theta_0, \theta_1}(z) = \frac{p(\theta_1, z)}{p(\theta_0, z)}.$$

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

$$\mathbb{P}_{\theta_0}(r_{\theta_0, \theta_1}(Z) > d_{\theta_0, \theta_1}) + \gamma_{\theta_0, \theta_1} \mathbb{P}_{\theta_0}(r_{\theta_0, \theta_1}(Z) = d_{\theta_0, \theta_1}) = \alpha. \quad (\text{I-5.10})$$

et que le test de fonction critique $\phi_{\theta_0, \theta_1}$

$$\phi_{\theta_0, \theta_1}(z) = \begin{cases} 1 & \text{si } r_{\theta_0, \theta_1}(z) > d_{\theta_0, \theta_1}, \\ \gamma_{\theta_0, \theta_1} & \text{si } r_{\theta_0, \theta_1}(z) = d_{\theta_0, \theta_1}, \\ 0 & \text{si } r_{\theta_0, \theta_1}(z) < d_{\theta_0, \theta_1} \end{cases}$$

est U.P.P. (α) pour le test d'hypothèses simple (I-5.9).

Comme la fonction $t \mapsto \psi(t; \theta_1)/\psi(t; \theta_0)$ est strictement croissante en t , chercher des constantes (d, γ) telles que

$$\mathbb{P}_{\theta_0}(r_{\theta_0, \theta_1}(Z) > d) + \gamma \mathbb{P}_{\theta_0}(r_{\theta_0, \theta_1}(Z) = d) = \alpha \quad (\text{I-5.11})$$

équivalent à chercher des constantes (c, γ) telles que

$$\mathbb{P}_{\theta_0}(T(Z) > c) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c) = \alpha. \quad (\text{I-5.12})$$

La seconde équation ne dépend pas de θ_1 et par conséquent les solutions c, γ de (I-5.12) ne dépendent pas de θ_1 . Ceci implique que les solutions d, γ de (I-5.11) ne dépendent pas de θ_1 .

Dans la suite, nous posons $c_{\theta_0, \theta_1} = c_{\theta_0}$ et $\gamma_{\theta_0, \theta_1} = \gamma_{\theta_0}$. Considérons la fonction critique

$$\phi_{\theta_0}(z) = \begin{cases} 1 & \text{si } T(z) > c_{\theta_0}, \\ \gamma_{\theta_0} & \text{si } T(z) = c_{\theta_0}, \\ 0 & \text{si } T(z) < c_{\theta_0} \end{cases} \quad (\text{I-5.13})$$

Par construction, $\beta_{\phi_{\theta_0}}(\theta_0) = \mathbb{E}_{\theta_0}[\phi_{\theta_0}(Z)] = \alpha$ et, pour tout $\theta_1 > \theta_0$, le test de fonction critique ϕ_{θ_0} est U.P.P. (α) pour le test d'hypothèses simples :

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1.$$

Par conséquent, le test de fonction critique ϕ_{θ_0} est U.P.P. (α) pour le test d'hypothèse

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

La fonction critique ϕ_{θ_0} définie par (I-5.13) 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-5.15 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 $\theta \leq \theta_0$,

$$\beta_{\phi_{\theta_0}}(\theta) \leq \beta_{\phi_{\theta_0}}(\theta_0)$$

et donc que le test randomisé de fonction critique ϕ_{θ_0} est un test de niveau α de l'hypothèse

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

ce que l'on écrit $\phi_{\theta_0} \in \mathcal{K}_\alpha(\{\theta \leq \theta_0\})$. Comme

$$\mathcal{K}_\alpha(\{\theta \leq \theta_0\}) \subset \mathcal{K}_\alpha(\{\theta = \theta_0\})$$

et comme ϕ_{θ_0} est U.P.P. (α) dans la classe $\mathcal{K}_\alpha(\{\theta = \theta_0\})$ contre les alternatives de la forme $H_1 = \{\theta = \theta_1\}$ pour tout $\theta_1 > \theta_0$, ϕ_{θ_0} est aussi U.P.P. (α) dans la classe $\mathcal{K}_\alpha(\{\theta \leq \theta_0\})$ contre toutes les alternatives $H_1 = \{\theta = \theta_1\}$ pour tout $\theta_1 > \theta_0$. Par conséquent ϕ_{θ_0} est U.P.P. (α) dans $\mathcal{K}_\alpha(\{\theta \leq \theta_0\})$ contre $H_1 = \{\theta > \theta_0\}$. Ceci démontre le point (i) du théorème. Le point (ii) a été établi lorsque nous avons invoqué le lemme I-5.15. $\square$

Remarque I-5.17. Si nous souhaitons tester

$$H_0 : \theta \geq \theta_0, \quad \text{contre} \quad H_1 : \theta < \theta_0$$

le théorème I-5.16 reste vrai en changeant le sens des inégalités dans la définition de la fonction critique du test

$$\phi(z) = \begin{cases} 1 & \text{si } T(z) < c, \\ \gamma & \text{si } T(z) = c, \\ 0 & \text{si } T(z) > c, \end{cases} \quad (\text{I-5.14})$$

où les constantes c et γ sont solutions de l'équation :

$$\mathbb{E}_{\theta_0}[\phi(Z)] = \mathbb{P}_{\theta_0}(T(Z) < c) + \gamma \mathbb{P}_{\theta_0}(T(Z) = c) = \alpha. \quad \diamond$$

Exemple I-5.18 (Modèle binomial). Soit $Z = (X_1, \dots, X_n)$ un n -échantillon de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta = [0, 1]\}).$$

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

$$p(\theta, z) = (1 - \theta)^n \left(\frac{\theta}{1 - \theta} \right)^{\sum_{i=1}^n x_i}.$$

Ce modèle satisfait **H** I-5.11 (voir exemple I-5.14) 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

$$\phi_{\alpha, \theta_0}(z) = \begin{cases} 1 & \text{si } T(z) > c_{\alpha, \theta_0}, \\ \gamma_{\alpha, \theta_0} & \text{si } T(z) = c_{\alpha, \theta_0}, \\ 0 & \text{si } T(z) < c_{\alpha, \theta_0}, \end{cases} \quad (\text{I-5.15})$$

où les constantes c_{α, θ_0} et $\gamma_{\alpha, \theta_0}$ sont solutions de l'équation :

$$\mathbb{E}_{\theta_0}[\phi_{\alpha, \theta_0}(Z)] = \mathbb{P}_{\theta_0}(T(Z) > c_{\alpha, \theta_0}) + \gamma_{\alpha, \theta_0} \mathbb{P}_{\theta_0}(T(Z) = c_{\alpha, \theta_0}) = \alpha.$$

Le théorème I-5.16 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}

$$c_{\alpha, \theta_0} = \inf \{c \in \{0, \dots, n\} : \mathbb{P}_{\theta_0}(T(Z) > c) \leq \alpha\}.$$

puis calculer

$$\gamma_{\alpha, \theta_0} = \frac{\alpha - \mathbb{P}_{\theta_0}(T(Z) > c_{\alpha})}{\mathbb{P}_{\theta_0}(T(Z) = c_{\alpha})}. \quad \diamond$$

A titre d'exemple, si $\theta_0 = 1/2$, $n = 1000$, et $\alpha = 0.05$. Nous avons $c_{0.05, 1/2} = 526$ et $\gamma_{0.05, 1/2} = 0.4832$

Loi !du
\$\chi^2\$
Fonction
quantile

Exemple I-5.19 (Variance d'une loi gaussienne (suite)). Soit $(X_1, \dots, X_n)$ un n -échantillon gaussien

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(0, \theta), \theta \in \Theta = \mathbb{R}_+^*\}) .$$

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 $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i^2$. Le test U.P.P. (α) rejette H_0 lorsque $T(X_1, \dots, X_n) \leq d_\alpha$, où d_α est solution de l'équation :

$$\mathbb{P}_{\theta_0}(T \leq d_\alpha) = \alpha .$$

Comme $T(X_1, \dots, X_n)/\theta_0 = \sum_{i=1}^n X_i^2/\theta_0$ est un χ^2 centré à n degrés de liberté, la constante critique du test est $\theta_0 \chi_\alpha^2(n)$ où $\chi_\alpha^2(n)$ est le quantile d'ordre α de la distribution χ_n^2 . $\diamond$

Exemple I-5.20 (Loi de Poisson). Soit $Z = (X_1, \dots, X_n)$ un n -échantillon de la loi de Poisson

$$(\mathbb{N}, \mathcal{P}(\mathbb{N}), \{\text{Poi}(\theta), \theta \in \mathbb{R}_+^*\}) .$$

La vraisemblance de l'observation (par rapport à la mesure de comptage) est donnée, en posant $z = (x_1, \dots, x_n) \in \mathbb{N}^n$,

$$p_\theta(z) = \frac{1}{\prod_{i=1}^n x_i!} e^{-n\theta} \theta^{T(z)} \quad T(z) := \sum_{i=1}^n x_i .$$

C'est une famille à rapports de vraisemblance monotones. Nous effectuons le test unilatéral

$$H_0 : \theta \leq \theta_0, \quad \text{contre} \quad H_1 : \theta > \theta_0 . \quad (\text{I-5.16})$$

Notons que $T(Z) := \sum_{i=1}^n X_i$ suit une loi de Poisson de paramètre $n\theta$. Soit $\alpha \in]0, 1[$. Considérons la fonction

$$\phi_{\alpha, \theta_0}(z) = \begin{cases} 1 & T(z) > c_{\alpha, \theta_0} \\ \gamma_{\alpha, \theta_0} & T(z) = c_{\alpha, \theta_0} \\ 0 & T(z) < c_{\alpha, \theta_0} \end{cases} \quad (\text{I-5.17})$$

où c_{α, θ_0} et $\gamma_{\alpha, \theta_0}$ solutions de l'équation

$$\alpha = \sum_{j=c+1}^{\infty} \frac{e^{-n\theta_0} (n\theta_0)^j}{j!} + \gamma \frac{e^{-n\theta_0} (n\theta_0)^c}{c!} .$$

Le test de fonction critique (I-5.17) est U.P.P. (α) pour le test unilatéral (I-5.16). $\diamond$

Exemple I-5.21 (Loi Uniforme). Soit $(X_1, \dots, X_n)$ un n -échantillon d'une loi uniforme

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}^+), \{\text{Unif}(0, \theta), \theta \in \Theta = \mathbb{R}_+^*\}) .$$

Pour $\theta_0 \in \mathbb{R}_+^*$, considérons les deux hypothèses $H_0 : \theta \leq \theta_0$ et $H_1 : \theta > \theta_0$. 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 $\theta_1 < \theta_2$,

$$\frac{p_{\theta_2}(x_1, \dots, x_n)}{p_{\theta_1}(x_1, \dots, x_n)} := \frac{\theta_2^n \mathbb{1}_{[0, \theta_2]}(x_{n:n})}{\theta_1^n \mathbb{1}_{[0, \theta_1]}(x_{n:n})} ,$$

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. $x_{n:n} \leq \theta_2$. Donc cette famille vérifie **H** I-5.11. La distribution de $X_{n:n}$ sous $\mathbb{P}_{\theta_0}$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par $q_{\theta_0}(x) = n\theta_0^{-n} x^{n-1} \mathbb{1}_{[0, \theta_0]}(x)$, le test (pur) donné par (I-5.8) est U.P.P. de niveau α en choisissant c_α de telle sorte que

$$\alpha = \frac{n}{\theta_0^n} \int_{c_\alpha}^{\theta_0} x^{n-1} dx = 1 - \frac{c_\alpha^n}{\theta_0^n} .$$

Dans ce problème, le test U.P.P. n'est pas unique. On peut montrer que le test (randomisé) de fonction critique

$$\phi(X_1, \dots, X_n) = \begin{cases} 1 & \text{si } X_{n:n} > \theta_0 \\ \alpha & \text{si } X_{n:n} \leq \theta_0 , \end{cases}$$

est aussi U.P.P. de niveau α . $\diamond$

Il est intéressant de noter que la construction ci-dessus ne s'étend pas directement au cas des hypothèses bilatérales. Considérons $X_1, \dots, X_n$ un n -échantillon i.i.d. d'une famille exponentielle associée à (h, T) de densité (par rapport à une mesure de domination μ)

$$p(x; \theta) = h(x) \exp(\phi(\theta)T(x) - \psi(\theta)),$$

où $\theta \rightarrow \phi(\theta)$ est une fonction croissante de θ . Supposons que $\mathbb{P}_\theta(\sum_{i=1}^n T(X_i) = c) = 0$ 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 non randomisé et admettra les régions critiques $T(x_1, \dots, x_n) = \sum_{i=1}^n T(x_i) \geq c$ si $\theta_1 > \theta_0$ et $T(x_1, \dots, x_n) \leq c$ si $\theta_0 < \theta_1$. On voit que la structure des tests U.P.P. est différente suivant que l'on considère des alternatives $\theta_1 > \theta_0$ et $\theta_1 < \theta_0$. C'est pourquoi il n'existe pas de test U.P.P. dans ce cadre.

La notion d'optimalité U.P.P. est naturelle en adoptant le point de vue de Neyman. Toutefois, en dehors des tests que nous venons de décrire, il est très rare de pouvoir construire des tests U.P.P. (α) . Pour prendre un exemple très simple, considérons un modèle gaussien

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\theta, 1), \theta \in \mathbb{R}\}) .$$

On note Z l'observation canonique et

$$p_\theta(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z - \theta)^2\right)$$

Soit $\theta_1 > \theta_0$ et $\alpha \in]0, 1[$ et considérons le test

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta = \theta_1 \quad (\text{I-5.18})$$

Le théorème de Neyman-Pearson (Théorème I-5.5) montre qu'il existe une constante c_α telle que le test de fonction critique

$$\phi_\alpha^+(z) = \begin{cases} 1 & p_{\theta_1}(z) > c_\alpha p_{\theta_0}(z) \\ 0 & p_{\theta_1}(z) \leq c_\alpha p_{\theta_0}(z) \end{cases}$$

est U.P.P. (α) pour (I-5.18). On voit facilement que

$$\phi_{\theta_0, \alpha}^+(z) = \mathbb{1}_{\{z > z_{\theta_0, \alpha}^+\}},$$

où $z_{\theta_0, \alpha}^+ := \theta_0 + z_\alpha$ est le α -quantile d'une loi gaussienne centrée réduite. En fait, en reprenant la preuve de ce théorème dans ce cas simplifié, on voit que, la loi étant continue, pour toute fonction $\phi : \mathbb{R} \rightarrow [0, 1]$ mesurable,

$$\begin{aligned} \mathbb{E}_{\theta_1} \left[\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} - \phi(Z) \right] &= \mathbb{E}_{\theta_0} \left[\left(\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} - \phi(Z) \right) \frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} \right] \\ &= \mathbb{E}_{\theta_0} \left[\left(\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} - \phi(Z) \right) \left(\frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} - c_\alpha \right) \right] + c_\alpha \mathbb{E}_{\theta_0} \left[\left(\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} - \phi(Z) \right) \right] . \end{aligned}$$

Supposons que ϕ soit la fonction critique d'un test randomisé de niveau α . Dans ce cas, le second terme est positif ou nul. Supposons qu'il existe un ensemble Ω_+ de mesure de Lebesgue non nulle sur lequel $\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} > \phi(Z)$. Dans ce cas, $\phi_{\theta_0, \alpha}^+(z) = 1$ et donc $p_{\theta_1}(z) > c_\alpha p_{\theta_0}(z)$ ce qui implique

$$\mathbb{E}_{\theta_0} \left[\left(\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} - \phi(Z) \right) \left(\frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} - c_\alpha \right) \right] \geq \mathbb{E}_{\theta_0} \left[\left(\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} - \phi(Z) \right) \left(\frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} - c_\alpha \right) \mathbb{1}_{\Omega_+}(Z) \right] > 0 .$$

Il en va de même si $\mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}} < \phi(Z)$ sur un ensemble de mesure de Lebesgue non nulle. Dans ce cas en effet, nous avons nécessairement $\phi_\alpha^+(z) = 0$ et $p_{\theta_1}(z) \leq c_\alpha p_{\theta_0}(z)$. Ainsi, $\phi_{\theta_0, \alpha}^+(Z) = \mathbb{1}_{\{Z > z_{\theta_0, \alpha}^+\}}$ est l'unique test U.P.P. (α) à un ensemble Lebesgue-négligeable près.

Supposons maintenant que $\theta_1 < \theta_0$. On montre de même que $\phi_{\theta_0, \alpha}^-(Z) = \mathbb{1}_{\{Z < z_{\theta_0, \alpha}^-\}}$ où $z_{\theta_0, \alpha}^- := \theta_0 - z_\alpha$ est l'unique test U.P.P. (α) à ensemble de mesure nulle près de (I-5.18).

Soit $\phi : \mathbb{R} \rightarrow [0, 1]$ une fonction mesurable telle que $\mathbb{E}_{\theta_0}[\phi(Z)] \leq \alpha$, il existe nécessairement un ensemble de mesure non nulle sur lequel $\mathbb{1}_{\{z > z_{\theta_0, \alpha}^+\}} > \phi(z)$ ou $\mathbb{1}_{\{z < z_{\theta_0, \alpha}^-\}} > \phi(z)$. Supposons en effet que cela ne soit pas le cas. Nous aurions, sauf sur un ensemble Lebesgue-négligeable,

$$\mathbb{1}_{\{z > z_{\theta_0, \alpha}^+\}} \leq \phi(z) \quad \text{et} \quad \mathbb{1}_{\{z < z_{\theta_0, \alpha}^-\}} \leq \phi(z)$$

ce qui contredit $\mathbb{E}_{\theta_0}[\phi(Z)] \leq \alpha$. Le test ϕ est strictement moins puissant que $\phi_{\theta_0, \alpha}^+$ en $\theta_1 > 0$ ou que $\phi_{\theta_0, \alpha}^-$ en $\theta_1 < 0$. Il n'existe donc pas de tests U.P.P. (α) dans ce problème élémentaire.

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.

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 explicites (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. A titre d'exemple, soit $(X_1, \dots, X_n)$ un n -échantillon d'une famille paramétrique

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

Supposons que pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$,

$$\int x_1 \mathbb{Q}_{\mu, \sigma^2}(dx_1) = \mu \quad \text{et} \quad \int x_1^2 \mathbb{Q}_{\mu, \sigma^2}(dx_1) = \mu^2 + \sigma^2.$$

Nous effectuons le test :

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

où $\mu_0 \in \mathbb{R}$ est une valeur donnée. Nous avons déjà étudié ce test dans l'Exemple I-3.24 ; nous avons proposé à cette occasion la statistique de test donnée par

$$T_n(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$$

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 et la variance empiriques. Supposons tout que pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$, $\mathbb{Q}_{\mu, \sigma^2}$, est une loi gaussienne de moyenne μ et de variance σ^2 . Sous cette hypothèse, le théorème de Gosset (Théorème IV-5.24) montre que, sous $\mathbb{P}_{\mu, \sigma^2}^n = \mathbb{P}_{\mu, \sigma^2}^{\otimes n}$,

- (i) La moyenne empirique $\bar{X}_n$ suit une loi normale $N(\mu, \sigma^2/n)$.
- (ii) $(n-1)S_n^2/\sigma^2$ suit une loi du χ^2 centrée à $(n-1)$ degrés de liberté.
- (iii) La moyenne empirique et la variance empirique sont indépendantes.

Par conséquent, pour tout $\sigma^2 \in \mathbb{R}_+^*$, la distribution de T_n sous $\mathbb{P}_{\mu_0, \sigma^2}^n$ est une loi de Student à $n-1$ degrés de liberté, notée $t(n-1)$ (voir Section IV-5.6 pour la définition et les propriétés).

Évidemment, considérer un modèle gaussien $\{N(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}$ est très restrictif. Si le nombre d'observations est "suffisamment" grand, nous pouvons montrer que pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$,

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 critique $|T_n| > 1.96$ pour des observations indépendantes de distribution gaussienne ou exponentielle.

la loi de la statistique T_n sous $\mathbb{P}_{\mu, \sigma^2}^n$ 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 Chapitre IV-2. Plus précisément, nous pouvons établir que, pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$, T_n converge en loi vers une loi normale centrée réduite (voir Définition IV-2.24 et Exemple II-1.22), i.e. pour toute fonction h continue et bornée et $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$, nous avons

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\mu, \sigma^2}^n [h(T_n)] = \lim_{n \rightarrow \infty} \int \cdots \int h(T_n(x_1, \dots, x_n)) \mathbb{Q}_{\mu, \sigma^2}^{\otimes n} \left(\prod_{i=1}^n dx_i \right) = \int h(z) g(z) dz$$

où g est la densité d'une variable gaussienne centrée réduite. 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 $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\mu, \sigma^2}^n (T_n \in [-a, a]) = G(a) - G(-a) .$$

Pour $\alpha \in]0, 1[$, nous pouvons choisir a_α de telle sorte que $G(a_\alpha) - G(-a_\alpha) = 1 - \alpha$, et donc, pour tout $\sigma^2 \in \mathbb{R}_+^*$, nous avons

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\mu_0, \sigma^2} (|T_n| \geq a_\alpha) = \alpha .$$

Nous sommes ainsi capable de construire un test de niveau *asymptotiquement* égal à α , sous des hypothèses très générales, en utilisant la notion de convergence en loi.

La table II-1 donne la taille du test $\phi_n(X_1, \dots, X_n) = \mathbb{1}\{|T_n| > a_\alpha\}$ pour $\alpha = 0.05$ (dans ce cas $a_\alpha \approx 1.96$) lorsque la loi $\left\{ \mathbb{Q}_{\mu, \sigma^2}, (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^* \right\}$ 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 $n \geq 50$.

Comme dans le premier chapitre, nous prendrons quelques libertés avec la terminologie dans la suite de l'exposé pour éviter d'alourdir les énoncés des résultats. Nous dirons par exemple, conformément à un usage très répandu dans la littérature statistique

" $\{X_k, k \in \mathbb{N}^*\}$ est un échantillon du modèle $\mathcal{C}$ "

Cette expression synthétise plusieurs informations :

- On dispose, pour tout $n \geq 1$, d'un n -échantillon $(X, \mathcal{X}, \mathcal{C})^n$ d'un modèle statistique $(X, \mathcal{X}, \mathcal{C})$ (l'espace des observations est le plus souvent implicite tout comme la tribu des événements).
- Pour tout $i \in \mathbb{N}$, les statistiques $X_i : X^\infty \rightarrow X$, $X_i(\mathbf{x}) = x_i$ où $\mathbf{x} = \{x_i, i \in \mathbb{N}^*\}$ sont des statistiques indépendantes et de même loi, et le modèle induit par X_i est $(X, \mathcal{X}, \mathcal{C})$.

II-1.1 Consistance d'une suite d'estimateurs

Définition II-1.1 (Suite d'estimateurs consistant). *Considérons une suite d'expériences statistiques paramétriques,*

$$(Z_n, \mathcal{Z}_n, \{\mathbb{P}_{n,\theta} : \theta \in \Theta\}), \quad \text{où } \Theta \subset \mathbb{R}^d.$$

Soit $g : \Theta \rightarrow \mathbb{R}^\ell$ une fonction.

Une suite d'estimateurs $\{T_n, n \in \mathbb{N}^\}$ où pour tout $n \in \mathbb{N}^*$*

$$T_n : (Z_n, \mathcal{Z}_n) \rightarrow (\mathbb{R}^\ell, \mathcal{B}(\mathbb{R}^\ell))$$

est dite consistante pour g si pour tout $\theta \in \Theta$ et $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{n,\theta} (\|T_n - g(\theta)\| \geq \varepsilon) = 0$$

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. La loi des grands nombres joue, directement ou indirectement, un rôle important. En effet, la loi des grands nombres (voir Théorème IV-2.18) montre que si $\{X_k, k \in \mathbb{N}\}$ est un échantillon du modèle

$$(\mathbb{R}^\ell, \mathcal{B}(\mathbb{R}^\ell), \{\mathbb{Q}_\theta, \theta \in \Theta\})$$

et si pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[|X_1|] = \int |x_1| \mathbb{Q}_\theta(dx_1) < \infty$, alors la moyenne empirique définie pour tout $n \in \mathbb{N}^*$ par

$$T_n(X_1, \dots, X_n) = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i, \quad (\text{II-1.1})$$

est une suite consistante d'estimateurs de $\mu : \theta \mapsto \mu(\theta) := \mathbb{E}_\theta[X_1] = \int x_1 \mathbb{Q}_\theta(dx_1)$. Plus généralement, si $f : \mathbb{R}^\ell \rightarrow \mathbb{R}^q$ est une fonction mesurable telle que $\mathbb{E}_\theta[|f(X_1)|] = \int |f(x_1)| \mathbb{Q}_\theta(dx_1) < \infty$, alors la suite d'estimateurs $\{T_n, n \in \mathbb{N}^*\}$ définie par

$$T_n(X_1, \dots, X_n) = n^{-1} \sum_{i=1}^n f(X_i),$$

est une suite consistante pour la fonction

$$\mu_f : \theta \mapsto \mu_f(\theta) = \mathbb{E}_\theta[f(X_1)] = \int f(x_1) \mathbb{Q}_\theta(dx_1).$$

Exemple II-1.2 (Paramètre d'une loi exponentielle). Soient $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle exponentiel

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

Rappelons que pour $\theta \in \mathbb{R}_+^*$, la loi exponentielle $\text{Expo}(\theta)$ admet une densité par rapport à la mesure de Lebesgue donnée par

$$q(\theta, x) = \theta e^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x).$$

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 $\mathbb{E}_\theta[X_1] = \theta^{-1}$ et $\text{Var}_\theta(X_1) = \theta^{-2}$. Le corollaire IV-2.17 montre que

$$\bar{X}_n \xrightarrow{\mathbb{P}_\theta - \text{prob}} \mathbb{E}_\theta[X_1] = \theta^{-1}.$$

Considérons la fonction $g : \theta \mapsto \theta^{-1}$. Cette fonction est continue sur $\mathbb{R}_+^*$. Le théorème IV-2.6 montre que $g(\bar{X}_n) = \bar{X}_n^{-1} \xrightarrow{\mathbb{P}_\theta\text{-prob}} \theta$ et par conséquent que $\{\bar{X}_n^{-1}, n \in \mathbb{N}^*\}$ est une suite consistante pour θ . $\diamond$

Exemple II-1.3. Soit $\{X_n, n \in \mathbb{N}\}$ un échantillon du modèle uniforme

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\text{Unif}([0, \theta]), \theta > 0\}).$$

Pour tout $n \in \mathbb{N}^*$, nous considérons l'estimateur défini par

$$\hat{\theta}_n = \max(X_1, \dots, X_n) = X_{n:n}.$$

Comme, par définition $\mathbb{P}_\theta(\hat{\theta}_n \leq \theta) = 1$, pour tout $0 < \varepsilon < \theta$, nous avons

$$\begin{aligned} \mathbb{P}_\theta(|\hat{\theta}_n - \theta| > \varepsilon) &= \mathbb{P}_\theta(\hat{\theta}_n < \theta - \varepsilon) \\ &= \mathbb{P}_\theta(X_1 < \theta - \varepsilon, \dots, X_n < \theta - \varepsilon) = \prod_{i=1}^n \mathbb{P}_\theta(X_i < \theta - \varepsilon) = (1 - \varepsilon/\theta)^n. \end{aligned}$$

Par conséquent, pour tout $0 < \varepsilon < \theta$, $\lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{\theta}_n - \theta| > \varepsilon) = 0$ et $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est une suite consistante pour θ . $\diamond$

II-1.2 Vitesse de convergence et 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 . Idéalement, nous voudrions disposer de la distribution de $T_n - g(\theta)$ sous $\mathbb{P}_\theta$, pour pouvoir calculer la probabilité $\mathbb{P}_\theta(\|T_n - g(\theta)\| \geq \delta)$ par exemple. Cet objectif est en général trop ambitieux. En revanche, dans de nombreuses situations, la suite de variables $n^{1/2}\{T_n - g(\theta)\}$ converge vers une distribution non-dégénérée, le plus souvent gaussienne, c'est pourquoi nous introduisons la définition suivante.

Définition II-1.4 (Normalité asymptotique). Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle

$$(\mathcal{X}, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\}).$$

et $g : \Theta \rightarrow \mathbb{R}^\ell$ une fonction. La suite d'estimateurs $\{T_n, n \in \mathbb{N}\}$ est asymptotiquement normale pour $g : \theta \rightarrow g(\theta)$ si, pour tout $\theta \in \Theta$,

$$\sqrt{n}\{T_n - g(\theta)\} \xrightarrow{\mathbb{P}_\theta} N(0, \Gamma(\theta)),$$

où pour tout $\theta \in \Theta$, $\Gamma(\theta)$ est une matrice symétrique et positive.

Exemple II-1.5. Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon de

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{Q}_\theta, \theta \in \Theta\}).$$

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

$$\mathbb{E}_\theta[\|X_1\|^2] = \int \|x_1\|^2 \mathbb{Q}_\theta(dx_1) < \infty,$$

et que la matrice

$$\Sigma(\theta) = \text{Cov}_\theta(X_1) = \int \{x_1 - \mu(\theta)\}\{x_1 - \mu(\theta)\}^T \mathbb{Q}_\theta(dx_1) \quad \text{où} \quad \mu(\theta) = \int x_1 \mathbb{Q}_\theta(dx_1).$$

Pour tout $n \in \mathbb{N}^*$, posons

$$T_n(X_1, \dots, X_n) = n^{-1} \sum_{i=1}^n X_i.$$

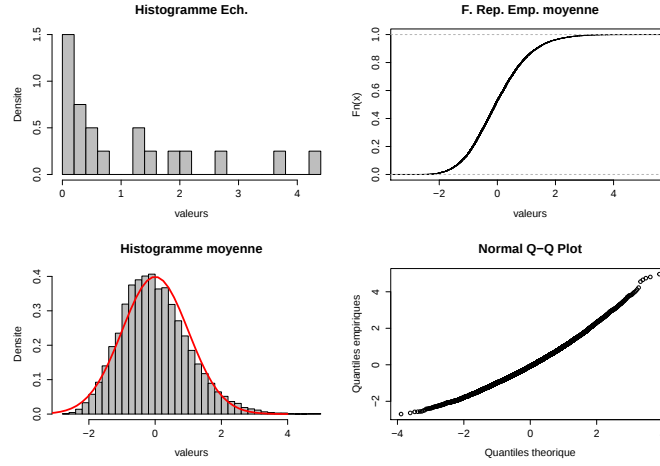


FIGURE II-1.1 – Loi exponentielle de paramètre 1. Figure en haut à gauche : histogramme de $n = 20$ échantillons. Figure en bas à gauche : histogramme de la moyenne pour 1000 réalisations indépendantes de $n = 20$ échantillons. Figure en haut à droite : fonction de répartition empirique de la moyenne pour 1000 réalisations indépendantes de $n = 20$ échantillons. Figure en bas à droite : diagramme quantile-quantile de la moyenne et d'une variable gaussienne centrée réduite

La normalité asymptotique de la suite $\{T_n, n \in \mathbb{N}\}$ découle du Théorème de la Limite Centrale ou T.L.C. (voir Théorème IV-2.39) : est inversible. Nous avons en effet, pour tout $\theta \in \Theta$,

$$\sqrt{n}\{\bar{X}_n - \mu(\theta)\} \xrightarrow{\mathbb{P}_\theta} N(0, \Sigma(\theta)) .$$

Par conséquent, la suite d'estimateurs $\{T_n, n \in \mathbb{N}\}$ est asymptotiquement normale pour moyenne $\mu : \theta \rightarrow \mu(\theta) = \mathbb{E}_\theta[X_1]$. 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-5.6). $\diamond$

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

$$T_n(X_1, \dots, X_n) = \hat{\theta}_n = \bar{X}_n^{-1} \quad \text{où} \quad \bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i .$$

En utilisant (IV-5.5) avec $p = 0$ et $r = 1, 2$, on montre que

$$\mathbb{E}_\theta[X_1] = \frac{1}{\theta} \quad \text{Var}_\theta(X_1) = \frac{1}{\theta^2} .$$

Le théorème de la limite centrale (Théorème IV-2.39) montre que, pour tout $\theta \in \Theta$,

$$n^{-1/2} \sum_{i=1}^n (X_i - 1/\theta) = n^{1/2}(\bar{X}_n - 1/\theta) \xrightarrow{\mathbb{P}_\theta} N(0, 1/\theta^2) .$$

Soit $h : \mathbb{R}_+^* \rightarrow \mathbb{R}$ une fonction dérivable. La δ -méthode (voir Théorème IV-2.47 et Exemple IV-2.48) montre que, pour tout $\theta \in \Theta$,

$$n^{1/2} \{h(\bar{X}_n) - h(1/\theta)\} \xrightarrow{\mathbb{P}_\theta} N(0, [h'(1/\theta)]^2 \text{Var}_\theta(X_1)) .$$

Considérons la fonction $h(u) = 1/u$. Nous avons donc, pour tout $u \in \mathbb{R}_+^*$, $h'(u) = -1/u^2$ et par conséquent $h'(1/\theta) = \theta^2$ pour tout $\theta \in \Theta$. Nous en déduisons

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, \theta^2) .$$

$\diamond$

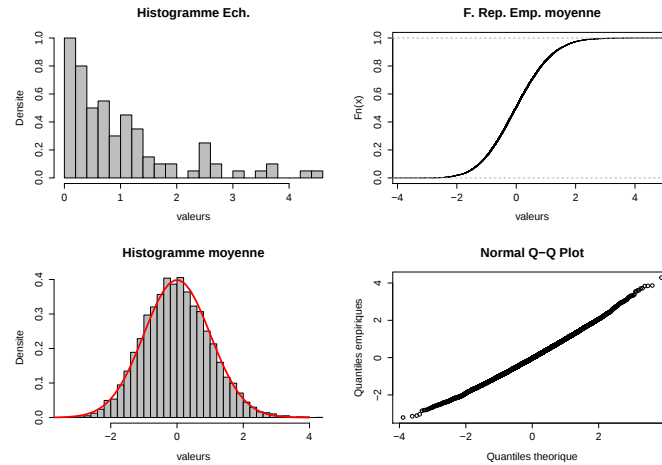


FIGURE II-1.2 – Loi exponentielle de paramètre 1. Figure en haut à gauche : histogramme de $n = 100$ échantillons. Figure en bas à gauche : histogramme de la moyenne pour 1000 réalisations indépendantes de $n = 100$ échantillons. Figure en haut à droite : fonction de répartition empirique de la moyenne pour 1000 réalisations indépendantes de $n = 100$ échantillons. Figure en bas à droite : diagramme quantile-quantile de la moyenne et d'une variable gaussienne centrée réduite

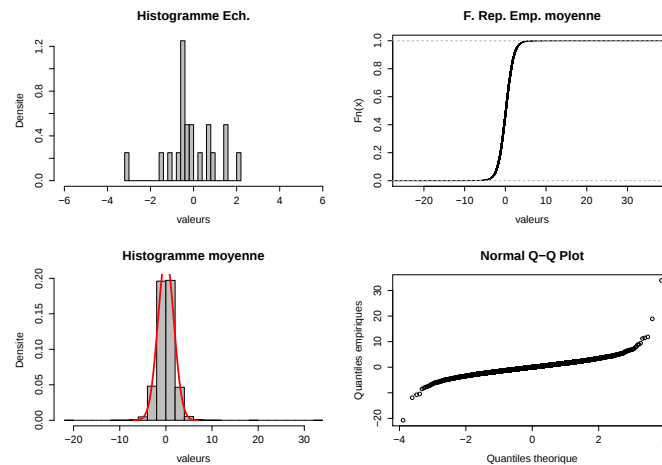


FIGURE II-1.3 – Loi de Student à 3 degrés de liberté. Figure en haut à gauche : histogramme de $n = 20$ échantillons. Figure en bas à gauche : histogramme de la moyenne pour 1000 réalisations indépendantes de $n = 20$ échantillons. Figure en haut à droite : fonction de répartition empirique de la moyenne pour 1000 réalisations indépendantes de $n = 20$ échantillons. Figure en bas à droite : diagramme quantile-quantile de la moyenne et d'une variable gaussienne centrée réduite

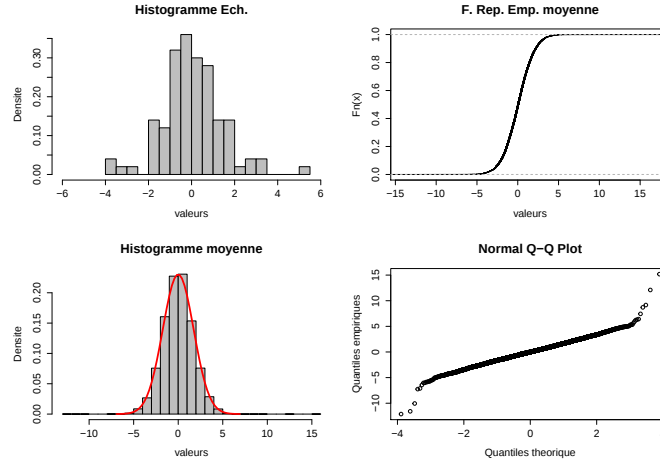


FIGURE II-1.4 – Loi de Student à 3 degrés de liberté. Figure en haut à gauche : histogramme de $n = 100$ échantillons. Figure en bas à gauche : histogramme de la moyenne pour 1000 réalisations indépendantes de $n = 100$ échantillons. Figure en haut à droite : fonction de répartition empirique de la moyenne pour 1000 réalisations indépendantes de $n = 100$ échantillons. Figure en bas à droite : diagramme quantile-quantile de la moyenne et d'une variable gaussienne centrée réduite

Exemple II-1.7 (Distribution limite pour le coefficient de régression). Dans ce chapitre introductif, nous nous intéressons exclusivement aux échantillons (observations 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 (voir ??). Soit h une densité par rapport à la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Supposons que

$$\int x^2 h(x) \lambda_{\text{Leb}}(dx) = 1 \quad \int x h(x) \lambda_{\text{Leb}}(dx) = 0 .$$

Considérons pour tout $n \in \mathbb{N}$, le modèle statistique

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{ p_{n,\theta} \cdot \lambda_{\text{Leb}}^{\otimes n}, \theta = (\beta_0, \beta_1, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^* \right\} \right),$$

où, pour tout $\theta = (\beta_0, \dots, \beta_1, \sigma^2) \in \Theta$, la densité $p_{n,\theta}$ sur $\mathbb{R}^n$ est donnée par

$$p_{n,\theta}(y_1, \dots, y_n) = \sigma^{-n} \prod_{i=1}^n h\left(\sigma^{-1} \{y_i - \beta_0 - \beta_1 x_i\}\right),$$

Notons $(Y_1, \dots, Y_n)$ les observations. Les observations sont indépendantes mais ne sont pas identiquement distribuées.

L'estimateur des moindres carrés des paramètres (β_0, β_1) pour les n observations $(Y_1, \dots, Y_n)$ est obtenu en minimisant la somme des carrés des erreurs

$$(\hat{\beta}_{n,0}, \hat{\beta}_{n,1}) = \arg \min_{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 .$$

Ce problème de régression admet une solution explicite. Nous allons nous intéresser particulièrement à l'estimateur $\hat{\beta}_{n,1}$, qui est donné par

$$\hat{\beta}_{n,1} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Nous allons montrer que pour que la suite $\{\beta_n, 1, n \in \mathbb{N}\}$ est consistante et que pour une normalisation appropriée, sa loi limite est gaussienne. Pour $\theta = (\beta_0, \beta_1, \sigma^2) \in \Theta$, définissons

$$\varepsilon_i(\theta) = \sigma^{-1} (Y_i - \beta_0 - \beta_1 x_i) .$$

Pour tout $n \in \mathbb{N}$ et $\theta \in \Theta$, le vecteur aléatoire $(\varepsilon_1(\theta), \dots, \varepsilon_n(\theta))$ est i.i.d. sous $\mathbb{P}_{n,\theta}$ et la loi de $\varepsilon_i(\theta)$ sous $\mathbb{P}_\theta$ est $h \cdot \lambda_{\text{Leb}}$. Par conséquent, sous $\mathbb{P}_{n,\theta}$ nous pouvons écrire

$$\sigma^{-1} \{ \hat{\beta}_{n,1} - \beta_1 \} = \frac{\sum_{i=1}^n \varepsilon_i(\theta) c_{n,i}}{\sum_{i=1}^n c_{n,i}^2}$$

où nous avons posé $c_{n,i} = (x_i - \bar{x}_n)$. Par conséquent, en utilisant la condition de Hajek-Sidak (Exemple IV-2.43), nous obtenons

$$\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \frac{\hat{\beta}_{n,1} - \beta_1}{\sigma} = \frac{\sum_{i=1}^n \varepsilon_i c_{n,i}}{\sigma \sqrt{\sum_{i=1}^n c_{n,i}^2}} \xrightarrow{\mathbb{P}} N(0, 1),$$

à condition que

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} \frac{(x_i - \bar{x}_n)^2}{\sum_{j=1}^n (x_j - \bar{x}_n)^2} = 0.$$

Cette condition est satisfaite par de très nombreux plans d'échantillonnage, par exemple lorsque $x_i = i/n$ pour tout $i \in \{1, \dots, n\}$. $\diamond$

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 $\{X_n, n \in \mathbb{N}\}$ un échantillon de

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{ \text{Unif}([0, \theta]), \theta \in \Theta = \mathbb{R}_+^* \}).$$

Pour tout $n \in \mathbb{N}^*$, considérons l'estimateur $\hat{\theta}_n = \max(X_1, X_2, \dots, X_n)$. Nous avons montré établi que la suite $\{\hat{\theta}_n, n \in \mathbb{N}^*\}$ est une suite consistant d'estimateurs de θ . Pour tout $x \geq 0$ et $\theta \in \Theta$, nous avons de plus

$$\begin{aligned} \mathbb{P}_\theta(n(\theta - \hat{\theta}_n) \geq x) &= \mathbb{P}_\theta(\hat{\theta}_n \leq \theta - x/n) = \mathbb{P}_\theta(X_1 \leq \theta - x/n, \dots, X_n \leq \theta - x/n) \\ &= \prod_{i=1}^n \mathbb{P}_\theta(X_i \leq \theta - x/n) = (1 - x/(n\theta))^n \end{aligned}$$

ce qui montre que

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(n(\theta - \hat{\theta}_n) \leq x) = 1 - e^{-x/\theta} = F_\theta(x).$$

La fonction $x \mapsto F_\theta(x)$ est 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. $\diamond$

II-1.3 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 $\{X_k, k \in \mathbb{N}\}$ un échantillon de l'expérience

$$(X, \mathcal{X}, \{ \mathbb{Q}_\theta, \theta \in \Theta \subset \mathbb{R}^d \}).$$

La méthode des moments consiste à estimer le paramètre θ en résolvant le système d'équations de d équations à d inconnues :

$$\frac{1}{n} \sum_{i=1}^n T_j(X_i) = \mathbb{E}_\theta[T_j(X_1)], \quad j = 1, \dots, d,$$

où $\{T_j\}_{j=1}^d$ sont des statistiques qui vérifient $\mathbb{E}_\theta[|T_j(X_1)|] < \infty$ pour tout $j \in \{1, \dots, d\}$. Supposons que ce système d'équations admette pour tout $n \in \mathbb{N}^*$ une solution unique notée $\hat{\theta}_n$. Alors $T_n(X_1, \dots, X_n) = \hat{\theta}_n$ est appelé *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

$$\text{pour tout } \theta_0 \in \Theta, \quad \hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \theta_0$$

et sont asymptotiquement normaux

$$\text{pour tout } \theta_0 \in \Theta, \quad \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \Sigma(\theta_0)) .$$

Les propriétés des estimateurs de moments découlent directement de la loi des grands nombres (Théorème IV-2.21), du T.L.C. (Théorème IV-2.39) et de la méthode-delta (Théorème IV-2.47). Pour simplifier la présentation des résultats, notons $\mathbf{T} = (T_1, \dots, T_d)$ et $\mathbf{e} : \Theta \mapsto \mathbb{R}^d$ la fonction à valeurs vectorielles

$$\mathbf{e}(\theta) = \mathbb{E}_{\theta}[\mathbf{T}(X)] .$$

Avec ces notations, l'estimateur des moments $\hat{\theta}_n$ est la solution du système d'équations

$$\hat{\mathbb{P}}_n \mathbf{T} := \frac{1}{n} \sum_{i=1}^n \mathbf{T}(X_i) = \mathbf{e}(\theta) ,$$

Une condition nécessaire pour que l'estimateur des moments soit défini est que l'estimateur empirique des moments $\hat{\mathbb{P}}_n \mathbf{T}$ 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

$$\hat{\theta}_n = \mathbf{e}^{-1}(\hat{\mathbb{P}}_n \mathbf{T}) \quad (\text{II-1.2})$$

et

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}\{\mathbf{e}^{-1}(\hat{\mathbb{P}}_n \mathbf{T}) - \theta_0\} .$$

Si l'estimateur des moments empiriques est asymptotiquement normal

$$\sqrt{n}\{\hat{\mathbb{P}}_n \mathbf{T} - \mathbf{e}(\theta_0)\} \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \Gamma(\theta_0))$$

et si la fonction $\mathbf{e}^{-1}$ est différentiable au point $\mathbf{e}(\theta_0)$, alors l'estimateur des moments est asymptotiquement normal par application de la méthode-delta (Théorème IV-2.47). En utilisant ce résultat nous pouvons maintenant donner un énoncé général pour l'existence, la convergence presque-sûre et la normalité asymptotique d'un estimateur des moments.

Théorème II-1.9. Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon de

$$(X, \mathcal{X}, \{\mathbb{Q}_{\theta}, \theta \in \Theta\})$$

où Θ est un ouvert de $\mathbb{R}^d$. Supposons que $\mathbf{e} : \Theta \rightarrow \mathbb{R}^d$ est une application continûment différentiable, injective et que, pour tout $\theta \in \Theta$, la matrice jacobienne $\mathbf{J}_{\mathbf{e}}(\theta)$ est inversible. Supposons de plus que, pour tout $\theta \in \Theta$, $\mathbb{E}_{\theta}[\|\mathbf{T}(X)\|^2] < \infty$.

Pour tout $\theta_0 \in \Theta$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\hat{\mathbb{P}}_n \mathbf{T} \in \mathbf{e}(\Theta)) = 1$$

et

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, [\mathbf{J}_{\mathbf{e}}(\theta_0)]^{-1} \text{Cov}_{\theta_0}(\mathbf{T}(X_1)) [\mathbf{J}_{\mathbf{e}}(\theta_0)]^{-T}) . \quad (\text{II-1.3})$$

Démonstration. Soit $\theta_0 \in \Theta$. Par le théorème d'inversion globale (??), $\mathbf{e}(\Theta)$ est un voisinage de $\mathbf{e}(\theta_0)$. Par la loi des grands nombres,

$$\hat{\mathbb{P}}_n \mathbf{T} := n^{-1} \sum_{i=1}^n \mathbf{T}(X_i) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \mathbb{E}_{\theta_0}[\mathbf{T}(X_1)] \quad (\text{II-1.4})$$

ce qui implique que

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} (\hat{\mathbb{P}}_n \mathbf{T} \in \mathbf{e}(\Theta)) = 1.$$

Si $\hat{\mathbb{P}}_n \mathbf{T} \in \mathbf{e}(\Theta)$ alors $\hat{\theta}_n$ est défini de façon unique et est donné par $\hat{\theta}_n = \mathbf{e}^{-1}(\hat{\mathbb{P}}_n \mathbf{T})$. 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-2.6 et du théorème IV-2.47. $\square$

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 :

$$e(\theta) = \mathbb{E}_{\theta} [T(X_1)] = \int_0^{+\infty} x \theta \exp(-\theta x) dx = \frac{1}{\theta}$$

$$\tilde{e}(\theta) = \mathbb{E}_{\theta} [\tilde{T}(X_1)] = \int_0^{+\infty} x^2 \theta \exp(-\theta x) dx = \frac{2}{\theta^2}.$$

Dans les deux cas considérés, ces équations ont des solutions uniques, qui définissent donc les deux estimateurs de moments suivants :

$$\hat{\theta}_{n,1} = \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^{-1}, \quad \text{et} \quad \hat{\theta}_{n,2} = \sqrt{2} \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right)^{-1/2}. \quad (\text{II-1.5})$$

Pour comparer ces estimateurs, nous appliquons le théorème II-1.9. Des calculs élémentaires montrent que

$$\begin{aligned} \text{Var}_{\theta}(X) &= \frac{1}{\theta^2} & \text{Var}_{\theta}(X^2) &= \frac{20}{\theta^4} \\ e'(\theta) &= -\frac{1}{\theta^2} & \tilde{e}'(\theta) &= -\frac{4}{\theta^3} \end{aligned}$$

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-3.6) sont données par :

$$v(\theta) = e'(\theta)^{-2} \text{Var}_{\theta}(X) = \theta^2$$

et

$$\tilde{v}(\theta) = \tilde{e}'(\theta)^{-2} \text{Var}_{\theta}(X^2) = \frac{20}{\theta^4} \frac{\theta^6}{16} = \frac{5}{4} \theta^2$$

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-3.6). $\diamond$

Exemple II-1.11 (Loi de Cauchy). Reprenons l'exemple I-2.3. Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon d'un modèle de Cauchy

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

On note $Y_i = \text{signe}(X_i) = T(X_i)$. On a

$$\mathbb{E}_{\theta} [T(X_1)] = \int \text{signe}(x) q(\theta, x) dx = 1 - 2F(-\theta),$$

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

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

L'estimateur des moments associé à T est donné par la valeur θ solution de

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

soit

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

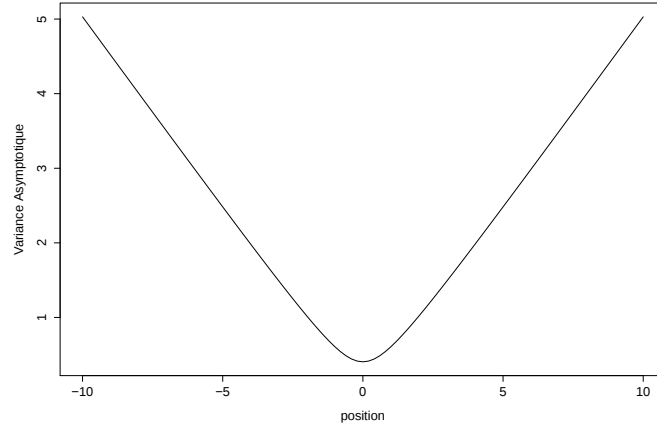


FIGURE II-1.5 – Variance asymptotique de l'estimateur des moments du paramètre de position d'une loi de Cauchy d'échelle 1

Par la loi des grands nombres, nous avons, pour tout $\theta_0 \in \Theta$,

$$\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 1 - 2F(-\theta_0) = \frac{2}{\pi} \arctan(\theta_0) ,$$

Le théorème IV-2.6 implique que $\hat{\theta}_n \xrightarrow{\theta_0 - \text{p.s.}} \theta_0$ pour tout $\theta_0 \in \Theta$. Comme

$$\begin{aligned} \text{Var}_{\theta_0}(Y_1) &= \mathbb{E}_{\theta_0}[Y_1^2] - (\mathbb{E}_{\theta_0}[Y_1])^2 = 1 - \{1 - 2F(-\theta_0)\}^2 \\ &= 4F(-\theta_0)\{1 - F(-\theta_0)\} = \{1 + (2/\pi) \arctan(\theta_0)\} \{1 - (2/\pi) \arctan(\theta_0)\} , \end{aligned}$$

le T.L.C. (Théorème IV-2.39) montre que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - \mathbb{E}_{\theta_0}[Y_i]\} \xRightarrow{\mathbb{P}_{\theta_0}} N(0, \{1 + (2/\pi) \arctan(\theta_0)\} \{1 - (2/\pi) \arctan(\theta_0)\}) .$$

Comme nous avons

$$e'(\theta) = \frac{\pi}{2} \frac{1}{1 + \theta^2} ,$$

le théorème II-1.9 montre que, pour tout $\theta_0 \in \mathbb{R}$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xRightarrow{\mathbb{P}_{\theta_0}} N\left(0, (2/\pi)^2 (1 + \theta_0^2)^2 \{1 + (2/\pi) \arctan(\theta_0)\} \{1 - (2/\pi) \arctan(\theta_0)\}\right) .$$

◇

Nous avons représenté la variance asymptotique dans la fig. II-1.5.

Exemple II-1.12 (Modèle de translation et d'échelle). Soit h est une densité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda_{\text{Leb}})$ vérifiant $\int xh(x)dx = 0$, $\int x^2 h(x)dx = m_2 > 0$ et $\int x^4 h(x)dx = m_4 < \infty$. Pour $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$, considérons la densité de probabilité

$$q(\theta, x) = \frac{1}{\sigma} h\left(\frac{x - \mu}{\sigma}\right) , \quad \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^* , \quad (\text{II-1.6})$$

Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon de

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{q_{\theta}, \theta \in \Theta = \mathbb{R} \times \mathbb{R}_+^*\})$$

Nous considérons l'estimateur donné pour tout $n \in \mathbb{N}^*$ par $\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n^2)$ où

$$\begin{cases} \hat{\mu}_n &= \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\sigma}_n^2 &= \frac{1}{nm_2} \sum_{i=1}^n (X_i - \hat{\mu}_n)^2 . \end{cases} \quad (\text{II-1.7})$$

Nous allons utiliser le théorème II-1.9 pour déterminer la loi asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta_0)$ sous $\mathbb{P}_{\theta_0}$. Notons tout d'abord qu'en appliquant le T.L.C. (Théorème IV-2.39) nous avons, pour tout $\theta_0 = (\mu_0, \sigma_0^2) \in \Theta$,

$$\sqrt{n} \left(\begin{bmatrix} n^{-1} \sum_{i=1}^n X_i \\ n^{-1} \sum_{i=1}^n X_i^2 \end{bmatrix} - \begin{bmatrix} \mu_0 \\ \mu_0^2 + m_2 \sigma_0^2 \end{bmatrix} \right) \xrightarrow{\mathbb{P}_{\theta_0}} N(\mathbf{0}_{2 \times 1}, \Sigma_{\theta_0})$$

$$\Sigma_{\theta_0} = \begin{bmatrix} \alpha_2(\theta_0) - \alpha_1^2(\theta_0) & \alpha_3(\theta_0) - \alpha_1(\theta_0)\alpha_2(\theta_0) \\ \alpha_3(\theta_0) - \alpha_1(\theta_0)\alpha_2(\theta_0) & \alpha_4(\theta_0) - \alpha_2^2(\theta_0) \end{bmatrix}$$

où nous avons posé, pour $r = 1, \dots, 4$,

$$\alpha_r(\theta_0) = \mathbb{E}_{\theta_0}[X_1^r] = \sum_{k=0}^r \binom{r}{k} \mu_0^k \sigma_0^{r-k} m_{r-k}, \quad m_j = \int x^j h(x) dx.$$

◇

Une application directe du théorème II-1.9 montre que, pour tout $\theta_0 \in \Theta$, nous avons

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(\mathbf{0}_{2 \times 1}, [\mathbf{J}_{\mathbf{e}}(\theta_0)]^{-1} \Sigma_{\theta_0} [\mathbf{J}_{\mathbf{e}}(\theta_0)]^{-T}).$$

II-1.4 Régions de confiance asymptotiques

Définition II-1.13 (Régions de confiance asymptotiques). Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon de

$$(\mathcal{X}, \mathcal{X}, \{\mathbb{Q}_{\theta}, \theta \in \Theta\})$$

Soit $g : \Theta \rightarrow \mathbb{R}^p$ et $\alpha \in]0, 1[$. Une suite d'ensembles aléatoires

$$\{\mathcal{C}_n(X_1, \dots, X_n), n \in \mathbb{N}\} \subset \mathcal{B}(\mathbb{R}^p)$$

définit une suite de régions de confiance au niveau asymptotique $1 - \alpha$, $\alpha \in]0, 1[$ pour la fonction $g : \theta \mapsto g(\theta)$ si

$$\inf_{\theta \in \Theta} \liminf_{n \rightarrow \infty} \mathbb{P}_{\theta}(g(\theta) \in \mathcal{C}_n(X_1, \dots, X_n)) \geq 1 - \alpha.$$

Remarque II-1.14. 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. ◇

Pour construire des intervalles de confiance asymptotiques, nous introduisons les fonctions asymptotiquement pivotales.

Définition II-1.15 (Fonctions asymptotiquement pivotales ou pivots asymptotiques). Une suite de fonctions mesurables $\{G_n, n \in \mathbb{N}\}$ où pour tout $n \in \mathbb{N}$, $G_n : \mathcal{X}^n \times \Theta \rightarrow \mathbb{R}^p$ est asymptotiquement pivotale si et seulement si il existe une probabilité μ sur $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ telle que, pour tout $\theta, \vartheta \in \Theta$ et $A \in \mathcal{B}(\mathbb{R}^p)$, on a

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta}(G_n(X_1, \dots, X_n, \theta) \in A) = \lim_{n \rightarrow \infty} \mathbb{P}_{\vartheta}(G_n(X_1, \dots, X_n, \vartheta) \in A) = \mu(A)$$

Considérons par exemple le cas $p = 1$. On choisit p_{α} et q_{α} de telle sorte que $\mu(q_{\alpha}) - \mu(p_{\alpha}) = 1 - \alpha$. Les ensembles

$$\mathcal{J}_n(\alpha, X_1, \dots, X_n) = \{\theta \in \Theta : p_{\alpha} \leq G(X_1, \dots, X_n; \theta) \leq q_{\alpha}\}$$

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

$$\mathbb{P}_\theta(\theta \in \mathcal{J}_n(\alpha, X_1, \dots, X_n)) = \mathbb{P}_\theta(p_\alpha \leq G(X_1, \dots, X_n; \theta) \leq q_\alpha) \rightarrow 1 - \alpha.$$

Intervalle de
Wilson
Intervalle de
Wald
Loi binomiale

C'est une suite de régions de confiance de niveau asymptotique $1 - \alpha$.

Exemple II-1.16 (Intervalle de confiance pour le paramètre d'une loi de Bernoulli). Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in]0, 1[)\}.$$

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 $\theta \in (0, 1)$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_\theta} \mathcal{N}(0, \theta(1 - \theta)).$$

et donc la suite de fonctions définies par :

$$G_n(X_1, \dots, X_n; \theta) = \sqrt{n} \frac{(\bar{X}_n - \theta)}{\sqrt{\theta(1 - \theta)}}$$

est une suite de fonctions asymptotiquement pivotales. 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 est constituée d'intervalle, dont les limites sont données par :

$$\frac{\bar{X}_n + \frac{z_{1-\alpha/2}^2}{2n} \pm z_{1-\alpha/2}^2 \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + z_{1-\alpha/2}^2/n} \quad (\text{II-1.8})$$

Ce dernier intervalle de confiance est appelé *intervalle de Wilson*. En utilisant le lemme de Slutsky (Lemme IV-2.33), il est facile de voir que la suite de fonctions définie par

$$G_n(X_1, \dots, X_n; \theta) = \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}$$

est aussi une fonction asymptotiquement pivotale. Les ensembles de confiance de probabilité de couverture asymptotiques $1 - \alpha$ associés à cette fonction pivotale sont des intervalles dont les bornes inférieures et supérieures sont cette fois-ci données par :

$$\bar{X}_n \pm z_{1-\alpha/2} \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}{\sqrt{n}}. \quad (\text{II-1.9})$$

◇

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) (voir Figure II-1.6-Figure II-1.7).

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.

II-1.5 Transformations de stabilisation de la variance

Une des utilisations principales de la δ -méthode est la construction de transformations de stabilisation de la variance (T.S.V.), 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 $\theta \in \mathbb{R}$. Soit $T_n = T(X_1, \dots, X_n)$ un estimateur du

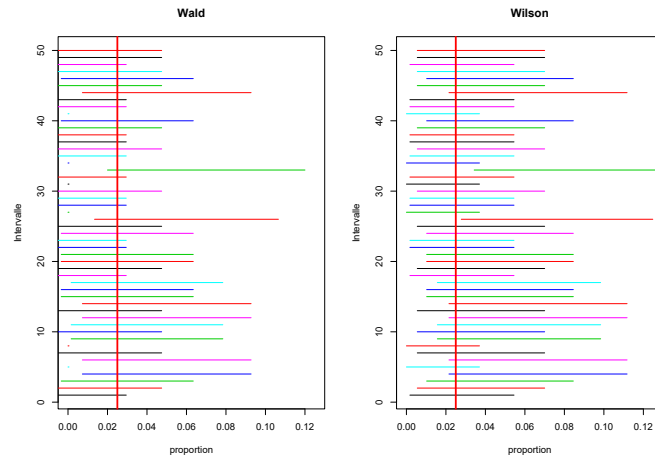


FIGURE II-1.6 – 50 intervalles de confiance de Wald et de Wilson de niveau de couverture de $1 - \alpha = 0.95$ pour une proportion $\theta = 0.025$ et $n = 100$

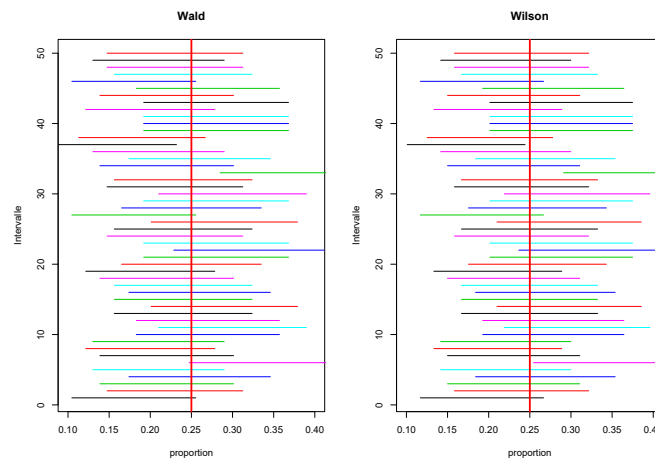


FIGURE II-1.7 – 50 intervalles de confiance de Wald et de Wilson de niveau de couverture de $1 - \alpha = 0.95$ pour une proportion $\theta = 0.25$ et $n = 100$

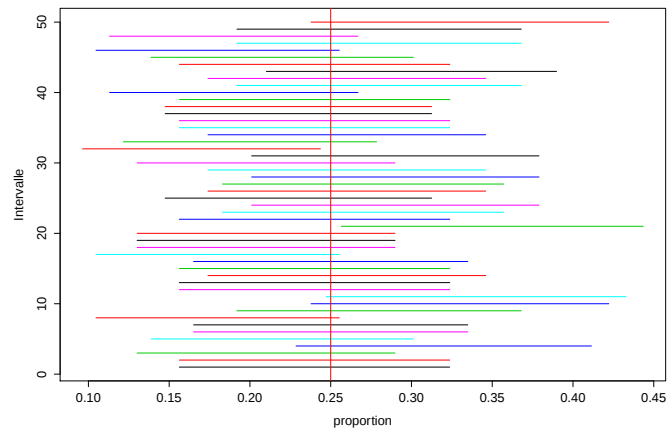


FIGURE II-1.8 – 50 intervalles de confiances de Wald de niveau de couverture de $1 - \alpha = 0.95$ pour une proportion $\theta = 0.25$ et $n = 100$

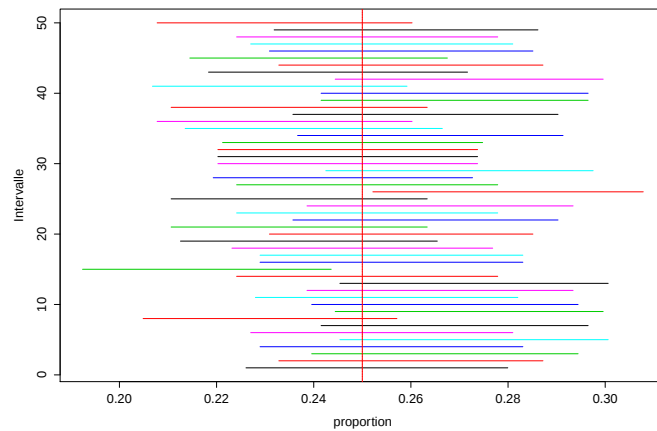


FIGURE II-1.9 – 50 intervalles de confiances de Wald de niveau de couverture de $1 - \alpha = 0.95$ pour une proportion $\theta = 0.25$ et $n = 1000$

paramètre θ . Supposons que, pour tout $\theta \in \Theta$, $\sqrt{n}(T_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, \sigma^2(\theta))$ et que $\theta \rightarrow \sigma(\theta)$ est continue et strictement positive. Pour $\alpha \in]0, 1[$ et $\theta \in \Theta$, nous avons donc

$$\mathbb{P}_\theta \left(T_n - z_{1-\frac{\alpha}{2}} \frac{\sigma(\theta)}{\sqrt{n}} \leq \theta \leq T_n + z_{1-\frac{\alpha}{2}} \frac{\sigma(\theta)}{\sqrt{n}} \right) \rightarrow_{n \rightarrow \infty} 1 - \alpha,$$

où $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ désigne le quantile d'ordre $1 - \alpha/2$ d'une gaussienne centrée réduite. Comme $T_n \xrightarrow{\mathbb{P}_{\theta-\text{prob}}} \theta$ et que la fonction σ est continue, le théorème de continuité (Théorème IV-2.30) montre que $\sigma(T_n) \xrightarrow{\mathbb{P}_{\theta-\text{prob}}} \sigma(\theta)$ pour tout $\theta \in \Theta$ et le lemme de Slutsky (Lemme IV-2.33) montre que

$$\sqrt{n}(T_n - \theta)/\sigma(T_n) \xrightarrow{\mathbb{P}_\theta} N(0, 1).$$

Par conséquent,

$$\mathbb{P}_\theta \left(T_n - z_{1-\frac{\alpha}{2}} \frac{\sigma(T_n)}{\sqrt{n}} \leq \theta \leq T_n + z_{1-\frac{\alpha}{2}} \frac{\sigma(T_n)}{\sqrt{n}} \right) \rightarrow_{n \rightarrow \infty} 1 - \alpha,$$

ce qui montre que

$$\left[T_n - z_{1-\frac{\alpha}{2}} \frac{\sigma(T_n)}{\sqrt{n}}, T_n + z_{1-\frac{\alpha}{2}} \frac{\sigma(T_n)}{\sqrt{n}} \right], \quad (\text{II-1.10})$$

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,

$$\sqrt{n}\{g(T_n) - g(\theta)\} \xrightarrow{\mathbb{P}_\theta} N(0, [g'(\theta)]^2 \sigma^2(\theta)).$$

Par conséquent, si nous choisissons la fonction g de telle sorte que, pour tout $\theta \in \Theta$,

$$[g'(\theta)]^2 \sigma^2(\theta) = k^2 \quad (\text{II-1.11})$$

où $k > 0$ est une constante, alors $\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{\mathbb{P}_\theta} N(0, k^2)$. Ceci nous permet alors de construire l'intervalle de confiance suivant pour $g(\theta)$, de niveau asymptotique $1 - \alpha$,

$$\left[g(T_n) - z_{1-\frac{\alpha}{2}} \frac{k}{\sqrt{n}}, g(T_n) + z_{1-\frac{\alpha}{2}} \frac{k}{\sqrt{n}} \right]$$

En supposant que la fonction g est bijective, nous en déduisons un intervalle de confiance pour θ :

$$\left[g^{-1} \left(g(T_n) - z_{1-\frac{\alpha}{2}} \frac{k}{\sqrt{n}} \right), g^{-1} \left(g(T_n) + z_{1-\frac{\alpha}{2}} \frac{k}{\sqrt{n}} \right) \right].$$

La raison pour laquelle cet intervalle de confiance est souvent préféré à (II-1.10) est qu'il n'est pas nécessaire d'utiliser un estimateur de substitution de la variance asymptotique (Définition II-3.6). Il découle de (II-1.11) que

$$g(\theta) = k \int_{\theta_0}^{\theta} \frac{1}{\sigma(\theta)} d\theta. \quad (\text{II-1.12})$$

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-1.17 (Loi Bernoulli). Soit $\{X_n, n \in \mathbb{N}\}$ un échantillon de Bernoulli

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\text{Ber}(\theta), \theta \in \Theta =]0, 1[\}) .$$

Pour tout $\theta \in \Theta$, le théorème de la limite centrale montre que

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, \theta(1 - \theta)).$$

Posons $\sigma(\theta) = \sqrt{\theta(1-\theta)}$ et $k = 1/2$. En utilisant (II-1.12), une transformation de stabilisation de la variance est donnée par

$$g(\theta) = \int \frac{1/2}{\sqrt{\theta(1-\theta)}} d\theta = \arcsin(\sqrt{\theta}) .$$

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

$$\sqrt{n} \left(\arcsin \left(\sqrt{\bar{X}_n/n} \right) - \arcsin(\sqrt{\theta}) \right) \xrightarrow{\mathbb{P}_\theta} N(0, 1/4) .$$

Donc, pour tout $\alpha \in]0, 1[$, si $\bar{X}_n \notin \{0, 1\}$,

$$\mathcal{J}_{n,\alpha} = \left[\sin^2(\arcsin(\sqrt{\bar{X}_n/n}) - z_{1-\alpha/2}/2\sqrt{n}), \sin^2(\arcsin(\sqrt{\bar{X}_n/n}) + z_{1-\alpha/2}/2\sqrt{n}) \right] ,$$

définit un intervalle de confiance de probabilité de couverture asymptotique $1 - \alpha$. $\diamond$

Exemple II-1.18 (Transformation de Fisher). Soit $\{(X_i, Y_i), i \in \mathbb{N}\}$ un échantillon gaussien

$$\left(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \left\{ N_2(\theta), \theta = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) \in \Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^* \times \mathbb{R}_+^* \times]-1, 1[\right\} \right)$$

où ρ est le coefficient de corrélation (voir Exemple IV-2.50). Nous avons montré que

$$\sqrt{n}(\hat{\rho}_n - \rho) \xrightarrow{\mathbb{P}_\theta} N(0, (1 - \rho^2)^2),$$

où $\hat{\rho}_n$ est le coefficient de corrélation empirique

$$\hat{\rho}_n = \frac{n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_{n,X} S_{n,Y}}$$

$\diamond$

Posons

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

g est une transformation de stabilisation de la variance pour le coefficient de corrélation. Cette transformation particulière est souvent appelée *Fisher z*. Nous avons

$$\sqrt{n}(\operatorname{arctanh}(\hat{\rho}_n) - \operatorname{arctanh}(\rho)) \xrightarrow{\mathbb{P}_\theta} N(0, 1)$$

L'intervalle de confiance de probabilité de couverture asymptotique $1 - \alpha$ pour le coefficient de corrélation ρ associé à la transformation Fisher- z est

$$\left[\tanh \left(\operatorname{arctanh}(\hat{\rho}_n) - \frac{z_{1-\alpha/2}}{\sqrt{n}} \right), \tanh \left(\operatorname{arctanh}(\hat{\rho}_n) + \frac{z_{1-\alpha/2}}{\sqrt{n}} \right) \right] .$$

Exemple II-1.19. 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 $\{X_n, n \in \mathbb{N}\}$ un échantillon du modèle uniforme

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{ \operatorname{Unif}([0, \theta]), \theta \in \Theta \in \mathbb{R}_+^* \}) .$$

Considérons la suite d'estimateurs définie pour tout $n \in \mathbb{N}$ par $\hat{\theta}_n = \max(X_1, X_2, \dots, X_n)$. Nous avons montré dans l'exemple II-1.19 que, pour tout $\theta \in \Theta$, $n(\theta - \hat{\theta}_n)$ converge en loi vers une loi exponentielle de paramètre θ ,

$$n(\theta - \hat{\theta}_n) \xrightarrow{\mathbb{P}_\theta} \operatorname{Exp}(\theta) .$$

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

$$g(\theta) = \int \frac{1}{\theta} d\theta = \log(\theta) .$$

Notons que, pour tout $\theta \in \Theta$, si Y est distribuée suivant la loi $\text{Exp}(\theta)$, alors Y/θ est distribuée suivant une loi exponentielle de paramètre 1. En utilisant le théorème IV-2.47 avec $r_n = n$ et $g(\theta) = \log(\theta)$, nous obtenons, pour tout $\theta \in \Theta$,

$$n(\log(\theta) - \log(\hat{\theta}_n)) \xrightarrow{\mathbb{P}_\theta} \text{Exp}(1) .$$

Comme pour $\alpha \in]0, 1[$, $\int_{-\log(1-\alpha)}^{\infty} \exp(-u) du = 1 - \alpha$, la relation précédente implique que, pour tout $\theta \in \Theta$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta (n(\log(\theta) - \log(\hat{\theta}_n)) \leq -\log(1 - \alpha)) = 1 - \alpha .$$

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 . $\diamond$

II-1.6 Tests asymptotiques

Nous avons construit dans la section I-3.5 des tests en utilisant des pivots. Comme pour les intervalles de confiance, il est possible aussi d'utiliser des pivots asymptotiques. Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{Q}_\theta, \theta \in \Theta \subset \mathbb{R}\}) .$$

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

$$\int x^2 \mathbb{Q}_\theta(dx) < \infty, \quad \text{et} \quad \theta = \int x \mathbb{Q}_\theta(dx) .$$

Posons

$$\sigma^2(\theta) = \int (x_1 - \mu(\theta))^2 \mathbb{Q}_\theta(dx) .$$

Le théorème de la limite centrale (théorème IV-2.39) assure alors que, pour tout $\theta \in \Theta$,

$$\sqrt{n} \frac{\bar{X}_n - \theta}{\sigma(\theta)} \xrightarrow{\mathbb{P}_\theta} \text{N}(0, 1) .$$

Si de plus l'application $\theta \mapsto \sigma(\theta)$ est continue, le lemme de Slutsky (lemme IV-2.33) garantit que

$$G_n(X_1, \dots, X_n, \theta) := \sqrt{n} \frac{\bar{X}_n - \theta}{\sigma(\bar{X}_n)} \xrightarrow{\mathbb{P}_\theta} \text{N}(0, 1) .$$

Considérons le test des hypothèses

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

et la statistique de test

$$\phi(X_1, \dots, X_n) = \mathbb{1}_{\{\bar{X}_n > \sigma(\bar{X}_n) z_{1-\alpha}/\sqrt{n}\}} .$$

Cette statistique de test vérifie, pour tout $\theta \leq 0$,

$$\begin{aligned} \mathbb{P}_\theta (\phi(X_1, \dots, X_n) = 1) &= \mathbb{P}_\theta \left(\bar{X}_n > \sigma(\bar{X}_n) \frac{z_{1-\alpha}}{\sqrt{n}} \right) \\ &\leq \mathbb{P}_\theta (G_n(X_1, \dots, X_n, \theta) > z_{1-\alpha}) \xrightarrow[n \rightarrow \infty]{} \alpha . \end{aligned}$$

Le test est de niveau asymptotique α .

Définition II-1.20 (Taille et niveau asymptotique). Soit $\{X_k, k \in \mathbb{N}^*\}$ un échantillon du modèle statistique

$$(\mathcal{X}, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$$

Soit $\{\phi_n, n \in \mathbb{N}^*\}$ une suite de tests (pour chaque $n \in \mathbb{N}$, $\phi_n : \mathcal{X}^n \rightarrow [0, 1]$ est un test randomisé). Considérons le problème

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

La taille asymptotique de la suite de tests $\{\phi_n, n \in \mathbb{N}\}$ est définie par

$$\sup_{\theta \in \Theta_0} \limsup_{n \rightarrow \infty} \beta_{\phi_n}(\theta) .$$

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

$$\text{pour tout } \theta \in \Theta_0, \quad \limsup_{n \rightarrow \infty} \beta_{\phi_n}(\theta) \leq \alpha .$$

L'erreur de deuxième espèce est contrôlée de la façon suivante.

Définition II-1.21 (Consistance). Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle

$$(\mathcal{X}, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$$

Soit $\{\phi_n, n \in \mathbb{N}\}$ une suite de test des hypothèses

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

La suite $\{\phi_n, n \in \mathbb{N}\}$ est consistante ou convergente si

$$\text{pour tout } \theta \in \Theta_1, \quad \liminf_{n \rightarrow \infty} \beta_{\phi_n}(\theta) = 1 .$$

Exemple II-1.22 (Distribution asymptotique de la statistique de Student). Soit $\{X_k, k \in \mathbb{N}^*\}$ une suite de variables aléatoires i.i.d. de loi $\{\mathbb{P}_{\mu, \sigma^2}, (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}$ telles que, $\mathbb{E}_{\mu, \sigma^2}[X_1] = \mu$ et $\mathbb{E}_{\mu, \sigma^2}[X_1^2] = \sigma^2$, pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$. Soit $\mu_0 \in \mathbb{R}$. Pour tester les hypothèses

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

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 la variance empirique. Nous avons

$$T_n \xrightarrow{\mathbb{P}_{\mu_0, \sigma^2}} N(0, 1) .$$

Donc, $\phi(X_1, \dots, X_n) = \mathbb{1}_{\{|T_n| > z_{1-\alpha/2}\}}$ est un test consistant de H_0 contre H_1 . ◇

Exemple II-1.23 (Test dans un modèle de Bernoulli). Reprenons l'exemple de Bernoulli I-1.1. L'estimateur $\bar{X}_n$ est asymptotiquement normal par le T.L.C. (Théorème IV-2.39) et $\bar{X}_n(1 - \bar{X}_n)$ est un estimateur consistant de $\text{Var}_p(X_1)$ par les théorèmes IV-2.6 et IV-2.18, donc le test de région de rejet

$$\mathcal{R}_n = \left\{ z = (x_1, \dots, x_n) \in \{0, 1\}^n : \bar{x}_n > \frac{1}{2} + z_{1-\alpha} \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \right\} ,$$

est asymptotiquement de taille égale à α , il est consistant d'après la loi faible des grands nombres (Théorème IV-2.18). 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 $N \sim N(0, 1)$, on a pour tout $x > 0$,

$$\mathbb{P}(N > x) \leq \frac{1}{2}e^{-x^2/2}.$$

Donc $z_{1-\alpha}^2 \leq 2\ln(\frac{1}{2\alpha}) \leq 2\ln(1/\alpha)$ 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. $\diamond$

II-1.6.1 Tests de Wald

Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle

$$(\mathcal{X}, \mathcal{F}, \{\mathbb{Q}_\theta, \theta \in \Theta\}).$$

Soit $T : \mathcal{X} \rightarrow \mathbb{R}$ une fonction mesurable. Supposons que pour tout $\theta \in \Theta$,

$$\int \{T(x)\}^2 \mathbb{Q}_\theta(dx) < \infty.$$

Soit $\mu : \Theta \rightarrow \mathbb{R}$ la fonction définie par

$$\mu : \theta \mapsto \mu(\theta) = \int T(x) \mathbb{Q}_\theta(dx).$$

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

$$G_n(X_1, \dots, X_n, \mu(\theta)) = \sqrt{n} \frac{n^{-1} \sum_{i=1}^n T(X_i) - \mu(\theta)}{\sqrt{n^{-1} \sum_{i=1}^n T^2(X_i) - (n^{-1} \sum_{i=1}^n T(X_i))^2}}. \quad (\text{II-1.13})$$

Par le T.L.C. (Théorème IV-2.39), la loi faible des grands nombres (Théorème IV-2.18), le théorème de l'application continue pour la convergence en probabilité (Théorème IV-2.6) et le lemme de Slutsky (Lemme IV-2.33),

$$G_n(X_1, \dots, X_n, \mu(\theta)) \xrightarrow{\mathbb{P}_\theta} N(0, 1). \quad (\text{II-1.14})$$

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

$$\mathcal{R}_n = \left\{ (x_1, \dots, x_n) \in \mathcal{X}^n : \left| n^{-1} \sum_{i=1}^n T(x_i) - f_0 \right| > t_{\text{crit}} \right\},$$

où, de manière équivalente, sous la forme

$$\mathcal{R}_n = \{ (x_1, \dots, x_n) \in \mathcal{X}^n : |G_n(x_1, \dots, x_n, f_0)| > t_{\text{crit}} \},$$

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

$$\mathbb{P}(|N(0, 1)| > t_{\text{crit}}) = 1 - \alpha,$$

c'est à dire $t_{\text{crit}} = z_{1-\alpha/2}$ le $(1 - \alpha/2)$ -quantile de la loi $N(0, 1)$. On a montré le résultat suivant.

Proposition II-1.24. *Le test de Wald de*

$$H_0 : \mu(\theta) = f_0, \quad \text{contre} \quad H_1 : \mu(\theta) \neq f_0 ,$$

de région de rejet

$$\mathcal{R}_n = \{x_1, \dots, x_n \in \mathbf{X}^n : |G_n(x_1, \dots, x_n, f_0)| > z_{1-\alpha/2}\} ,$$

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

$$\mathcal{R}_n = \{x_1, \dots, x_n \in \mathbf{X}^n : G_n(x_1, \dots, x_n, f_0) > z_{1-\alpha}\} , \quad (\text{II-1.15})$$

où $z_{1-\alpha}$ est le $(1 - \alpha)$ -quantile de la loi $N(0, 1)$. Soit θ tel que $\mu(\theta) \leq f_0$, on a alors, d'après (II-1.14),

$$\begin{aligned} \mathbb{P}_\theta (G_n(X_1, \dots, X_n, f_0) > z_{1-\alpha}) &\leq \mathbb{P}_\theta (G_n(X_1, \dots, X_n, \mu(\theta)) > z_{1-\alpha}) \\ &\xrightarrow[n \rightarrow \infty]{} \mathbb{P}(N(0, 1) > z_{1-\alpha}) = 1 - \alpha . \end{aligned}$$

On en déduit le résultat suivant :

Proposition II-1.25. *Le test de Wald, de*

$$H_0 : \mu(\theta) \leq f_0, \quad \text{contre} \quad H_1 : \mu(\theta) > f_0 ,$$

de région de rejet $\mathcal{R}_n$ définie en (II-1.15) est de niveau asymptotique α .

Chapitre II-2

Théorie asymptotique des (M,Z)-estimateurs

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.

Dans toute la suite de ce chapitre, nous supposons disposer d'un échantillon $\{X_k, k \in \mathbb{N}\}$ du modèle

$$\left(\mathcal{X}, \mathcal{X}, \left\{ \mathbb{Q}_\theta, \theta \in \Theta \subset \mathbb{R}^d \right\} \right).$$

Nous avons introduit les M -estimateurs dans la section I-2.4. Nous prenons ici une définition un peu plus générale, qui permet de couvrir aussi les situations où l'estimateur $\hat{\theta}_n$ est une solution approchée du problème d'optimisation. 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.

Soit $m : \Theta \times \mathcal{X} \rightarrow \mathbb{R}$, $(\theta, x) \mapsto m(\theta, x)$ une application mesurable et soit

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

On appelle M -estimateur associé à la fonction m tout estimateur $\hat{\theta}_n$ vérifiant la propriété suivante : il existe une suite de variables $\{\rho_n, n \in \mathbb{N}\}$ satisfaisant, pour tout $\theta_0 \in \Theta$, Les Z -estimateurs sont introduits dans la section I-2.2. Comme pour les M -estimateurs, nous prendrons ici une définition plus générale dans laquelle $\hat{\theta}_n$ est une solution approchée du système d'équations

$$\Psi_n(\theta) = n^{-1} \sum_{i=1}^n \psi(\theta, X_i) = \mathbf{0}_{d \times 1}.$$

II-2.1 Consistance des Z - et des M -estimateurs

Dans cette section, nous donnons des critères simples sur la famille paramétrique $\{\mathbb{P}_\theta, \theta \in \Theta\}$ 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

Il est clair que la consistance de ces estimateurs est liée au comportement asymptotique de la suite de processus aléatoires

$$\theta \mapsto M_n(\theta) .$$

Si, pour tout $\theta, \theta_0 \in \Theta$, $\mathbb{E}_{\theta_0}[|m(\theta, X_1)|] < \infty$, la loi des grands nombres (Théorème IV-2.18) montre, que pour tout $\theta \in \Theta$,

$$M_n(\theta) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} M_{\theta_0}(\theta) ,$$

où la fonction $\theta \mapsto M_{\theta_0}(\theta)$ est donnée par

$$M_{\theta_0}(\theta) = \mathbb{E}_{\theta_0}[m(\theta, X_1)] . \quad (\text{II-2.1})$$

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 $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \theta_0$, 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 $n \rightarrow \infty$) 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. Nous devons donc établir la “convergence uniforme” de cette suite de fonctions. Nous devons aussi supposer que le maximum de la fonction limite $\theta \mapsto M_{\theta_0}(\theta)$ 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 $M_n : \Theta \rightarrow \mathbb{R}$, $\theta \mapsto M_n(\theta)$ 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 $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} \left(\sup_{\theta \in \Theta} |M_n(\theta) - M_{\theta_0}(\theta)| \geq \varepsilon \right) = 0 .$$

(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

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\rho_n \geq \varepsilon) = 0 , \quad \text{pour tout } \varepsilon > 0 \quad (\text{II-2.2})$$

$$\liminf_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(M_n(\hat{\theta}_n) \geq M_n(\theta_0) - \rho_n) = 1 . \quad (\text{II-2.3})$$

◇

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 $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\|\hat{\theta}_n - \theta_0\| \geq \varepsilon) = 0 .$$

Démonstration. Soit $\theta_0 \in \Theta$. Comme θ_0 est le maximum de la fonction $\theta \rightarrow M_{\theta_0}(\theta)$, nous avons

$$\begin{aligned} 0 \leq M_{\theta_0}(\theta_0) - M_{\theta_0}(\hat{\theta}_n) &= M_{\theta_0}(\theta_0) - M_n(\theta_0) + M_n(\theta_0) - M_n(\hat{\theta}_n) + M_n(\hat{\theta}_n) - M_{\theta_0}(\hat{\theta}_n), \\ &\leq 2 \sup_{\theta \in \Theta} |M_n(\theta) - M_{\theta_0}(\theta)| + \rho_n + \{M_n(\theta_0) - M_n(\hat{\theta}_n) - \rho_n\} \mathbb{1}_{\{M_n(\theta_0) - \rho_n > M_n(\hat{\theta}_n)\}}. \end{aligned}$$

Par conséquent, pour tout $\eta > 0$, nous avons

$$\mathbb{P}_{\theta_0}(M_{\theta_0}(\theta_0) - M_{\theta_0}(\hat{\theta}_n) \geq \eta) = 0.$$

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

$$\{|\hat{\theta}_n - \theta_0| \geq \varepsilon\} \subset \{M_{\theta_0}(\hat{\theta}_n) \leq M_{\theta_0}(\theta_0) - \eta\}. \quad (\text{II-2.4})$$

Par conséquent, nous avons

$$\begin{aligned} \mathbb{P}_{\theta_0}(|\hat{\theta}_n - \theta_0| \geq \varepsilon) &\leq \mathbb{P}_{\theta_0}(M_{\theta_0}(\hat{\theta}_n) < M_{\theta_0}(\theta_0) - \eta) \\ &= \mathbb{P}_{\theta_0}(M_{\theta_0}(\theta_0) - M_{\theta_0}(\hat{\theta}_n) > \eta) \xrightarrow{n \rightarrow \infty} 0. \end{aligned} \quad \square$$

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 $\Theta \setminus B(\theta_0, \varepsilon)$ 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. $\square$

Exemple II-2.4 (Estimateur des moindres carrés). Soit q une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ qui satisfait $\sigma^2 := \int x^2 q(x) dx < \infty$ et $\int x q(x) dx = 0$. Soit Θ un sous ensemble compact de $\mathbb{R}$ (par exemple un segment). Pour tout $\theta \in \Theta$, définissons

$$q_\theta(x) = q(\theta, x) = q(x - \theta).$$

Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{q_\theta \cdot \lambda_{\text{Leb}}, \theta \in \Theta\}).$$

Considérons un M-estimateur de θ basé sur la fonction $m(\theta, x) = -(x - \theta)^2$. Pour tout $\theta \in \mathbb{R}$, la fonction

$$\begin{aligned} M_{\theta_0}(\theta) &= -\mathbb{E}_{\theta_0}[(X_1 - \theta)^2] = -\int (x - \theta)^2 q(x - \theta_0) dx = -\int (x + \theta_0 - \theta)^2 q(x) dx \\ &= -\sigma^2 - (\theta_0 - \theta)^2. \end{aligned}$$

Le maximum de la fonction $\theta \mapsto M_{\theta_0}(\theta)$ est atteint en θ_0 et H II-2.1-(ii) est satisfaite en appliquant le Lemme II-2.3. La fonction

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

admet un maximum unique, donné ici par $\hat{\theta}_n = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Par conséquent, $M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta)$ 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

$$\begin{aligned} M_n(\theta) - M_{\theta_0}(\theta) &= -\frac{1}{n} \sum_{i=1}^n X_i^2 + 2\theta \frac{1}{n} \sum_{i=1}^n X_i - \theta^2 + \sigma^2 + (\theta - \theta_0)^2 \\ &= -\left\{ \frac{1}{n} \sum_{i=1}^n X_i^2 - \sigma^2 - \theta_0^2 \right\} - 2\theta \left\{ \frac{1}{n} \sum_{i=1}^n X_i - \theta_0 \right\} \end{aligned}$$

Par conséquent, nous avons

$$\sup_{\theta \in \Theta} |M_n(\theta) - M_{\theta_0}(\theta)| \leq \left| \frac{1}{n} \sum_{i=1}^n X_i^2 - \sigma^2 - \theta_0^2 \right| + 2\text{diam}(\Theta) \left| \frac{1}{n} \sum_{i=1}^n X_i - \theta_0 \right|,$$

où $\text{diam}(\Theta)$ est le diamètre de Θ . Par la loi des grands nombres, pour tout $\theta_0 \in \Theta$, nous avons

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n X_i^2 &\xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \sigma^2 + \theta_0^2, \\ \frac{1}{n} \sum_{i=1}^n X_i &\xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \theta_0, \end{aligned}$$

ce qui montre que, pour tout $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} \left(\sup_{\theta \in \Theta} |M_n(\theta) - M_{\theta_0}(\theta)| \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \geq \varepsilon \right) = 0.$$

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. $\diamond$

Au prix de quelques hypothèses supplémentaires, nous pouvons d'ailleurs établir la consistance de la suite $\{\hat{\theta}_n, n \in \mathbb{N}\}$ sans supposer que l'espace des paramètres est compact (ce qui nous permettra de prendre $\Theta = \mathbb{R}$ dans l'exemple précédent).

II-2.1.2 Convergence des Z-estimateurs

Supposons que, pour tout $\theta, \theta_0 \in \Theta$, $\mathbb{E}_{\theta_0}[\|\psi(\theta, X_i)\|] < \infty$. D'après la loi des grands nombres (Théorème IV-2.18), pour tout $\theta, \theta_0 \in \Theta$, $\Psi_n(\theta) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \Psi_{\theta_0}(\theta)$. 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 .

H II-2.5 (Hypothèses pour la consistance d'un Z-estimateur).

(i) Pour tout $\theta_0 \in \Theta$ et $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} \left(\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi_{\theta_0}(\theta)\| \geq \varepsilon \right) = 0.$$

(ii) Pour tout $\theta_0 \in \Theta$, $\Psi_{\theta_0}(\theta_0) = 0$, et pour tout $\varepsilon > 0$,

$$\inf_{\|\theta - \theta_0\| \geq \varepsilon} \|\Psi_{\theta_0}(\theta)\| > 0.$$

(iii) La suite d'estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$ vérifie : pour tout $\theta_0 \in \Theta$ et tout $\varepsilon > 0$,

$$\mathbb{P}_{\theta_0} (\|\Psi_n(\hat{\theta}_n)\| \geq \varepsilon) = 0.$$

$\diamond$

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 $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} (\|\hat{\theta}_n - \theta_0\| \geq \varepsilon) = 0.$$

Démonstration. Le résultat découle du théorème II-2.2 avec $M_n(\theta) = -\|\Psi_n(\theta)\|$ et $M_{\theta_0}(\theta) = -\|\Psi_{\theta_0}(\theta)\|$. Pour vérifier la condition **H** II-2.1-(iii), on écrit

$$M_n(\hat{\theta}_n) - M_n(\theta_0) = \|\Psi_n(\theta_0)\| - \|\Psi_n(\hat{\theta}_n)\| = \|\Psi_n(\theta_0) - \Psi_{\theta_0}(\theta_0)\| - \|\Psi_n(\hat{\theta}_n)\| \geq -\|\Psi_n(\hat{\theta}_n)\| ,$$

et on utilise le fait que, par **H** II-2.5-(iii), le terme de droite converge en probabilité vers zéro. $\square$

II-2.1.3 Loi faible uniforme

Dans cette partie, nous donnons quelques éléments permettant de comprendre le type de résultats nécessaires pour obtenir la convergence uniforme de $\theta \mapsto M_n(\theta)$ ou $\theta \mapsto \Psi_n(\theta)$ dans les cas particuliers où ces fonctions sont de la forme $n^{-1} \sum_{i=1}^n \phi(\theta, X_i)$. Il existe bien entendu des techniques beaucoup plus sophistiquées.

Lemme II-2.7. *Considérons le modèle statistique $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta \subset K\})$ où K est un ensemble compact de $\mathbb{R}^d$. Soit $\phi : K \times X \rightarrow \mathbb{R}$, $(\theta, x) \mapsto \phi(\theta, x)$ une fonction. Supposons que*

- (i) *pour tout $x \in X$, la fonction $\theta \mapsto \phi(\theta, x)$ est continue sur K ,*
- (ii) *pour tout $\theta_0 \in \Theta$,*

$$\int \sup_{\theta \in K} |\phi(\theta, x)| \mathbb{Q}_{\theta_0}(dx) < \infty .$$

Alors, pour tout $\theta_0 \in \Theta$, la fonction

$$\theta \mapsto W_{\theta_0}(\theta) = \int \phi(\theta, x) \mathbb{Q}_{\theta_0}(dx)$$

est continue et

$$\lim_{\delta \rightarrow 0^+} \int_X w_\delta(x) \mathbb{Q}_{\theta_0}(dx) = 0 ,$$

où pour $x \in X$ et $\delta > 0$, $w_\delta(x)$ est le module de continuité de $\theta \mapsto \phi(\theta, x)$, défini par :

$$w_\delta(x) := \sup_{\{\theta, \vartheta \in K : \|\theta - \vartheta\| \leq \delta\}} |\phi(\theta, x) - \phi(\vartheta, x)| . \quad (\text{II-2.5})$$

Démonstration. On pose, pour tout $x \in X$, $\varphi(x) = \sup_{\theta \in K} |\phi(\theta, x)|$ (rappelons que toute fonction continue sur un ensemble compact atteint son maximum sur ce compact). Soit $\theta \in K$ et $\{\theta_n, n \in \mathbb{N}\}$ une suite d'éléments de K convergeant vers θ . Comme, pour tout $x \in X$, $\theta \rightarrow \phi(\theta, x)$ est continue, nous avons

$$\lim_{n \rightarrow \infty} \phi(\theta_n, x) = \phi(\theta, x) .$$

D'autre part, pour tout $n \in \mathbb{N}$, $|\phi(\theta_n, x)| \leq \varphi(x)$ et $\int \varphi(x) \mathbb{Q}_{\theta_0}(dx) < \infty$. Le théorème de convergence dominée montre que $W_{\theta_0}(\theta_n) \rightarrow W_{\theta_0}(\theta)$.

Toute fonction continue sur un compact est uniformément continue, nous avons donc, pour tout $x \in X$,

$$\lim_{\delta \rightarrow 0^+} w_\delta(x) = 0 .$$

D'autre part, nous avons, pour tout $\delta > 0$, $w_\delta(x) \leq 2\varphi(x)$. Pour tout suite $\{\delta_n, n \in \mathbb{N}\}$ telle que $\lim_{n \rightarrow \infty} \delta_n = 0$, le théorème de convergence dominée montre que

$$\lim_{n \rightarrow \infty} \int w_{\delta_n}(x) \mathbb{Q}_{\theta_0}(dx) = 0 . \quad \square$$

Théorème II-2.8. *Considérons le modèle statistique $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta \subset K\})$ où K est un ensemble compact de $\mathbb{R}^d$. Soit $\phi : K \times X \rightarrow \mathbb{R}$, $(\theta, x) \mapsto \phi(\theta, x)$ une fonction. Supposons que*

(i) pour tout $x \in X$, la fonction $\theta \mapsto \phi(\theta, x)$ est continue sur K .

(ii) pour tout $\theta_0 \in K$, $\int \sup_{\theta \in K} |\phi(\theta, x)| \mathbb{Q}_{\theta_0}(dx) < \infty$.

Alors, pour tout $\theta_0 \in K$ et $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} \left(\sup_{\theta \in K} |W_n(\theta) - W_{\theta_0}(\theta)| \geq \varepsilon \right) = 0$$

où, pour tout $\theta, \theta_0 \in \Theta \subset K$ et $n \in \mathbb{N}$,

$$W_n(\theta) := \frac{1}{n} \sum_{i=1}^n \phi(\theta, X_i) \quad \text{et} \quad W_{\theta_0}(\theta) := \mathbb{E}_{\theta_0}[\phi(\theta, X)] .$$

Démonstration. Soient $\varepsilon > 0$ et $\theta_0 \in K$. Le lemme II-2.7 montre qu'on peut choisir $\delta > 0$ tel que $\int w_\delta(x) \mathbb{Q}_{\theta_0}(dx) \leq \varepsilon$, où $w_\delta(x)$ est défini en (II-2.5). L'ensemble K étant compact, on peut extraire du recouvrement $\bigcup_{\theta \in K} B(\theta, \delta)$ un sous-recouvrement fini : il existe $m \in \mathbb{N}$ et $\{\theta^{(i)}\}_{i=1}^m$ tels que $K \subset \bigcup_{i=1}^m B(\theta^{(i)}, \delta)$. Notons que, pour tous $\theta, \vartheta \in \Theta$ tels que $\|\theta - \vartheta\| \leq \delta$, nous avons

$$|W_{\theta_0}(\theta) - W_{\theta_0}(\vartheta)| \leq \int |\phi(\theta, x) - \phi(\vartheta, x)| \mathbb{Q}_{\theta_0}(dx) \leq \int w_\delta(x) \mathbb{Q}_{\theta_0}(dx) .$$

ce qui implique

$$\sup_{\{\theta, \vartheta \in K : \|\theta - \vartheta\| \leq \delta\}} |W_{\theta_0}(\theta) - W_{\theta_0}(\vartheta)| \leq \varepsilon . \quad (\text{II-2.6})$$

Pour tout $\theta \in B(\theta^{(i)}, \delta)$, nous avons

$$|W_n(\theta) - W_{\theta_0}(\theta)| \leq |W_n(\theta) - W_n(\theta^{(i)})| + |W_n(\theta^{(i)}) - W_{\theta_0}(\theta^{(i)})| + |W_{\theta_0}(\theta^{(i)}) - W_{\theta_0}(\theta)|$$

ce qui implique

$$\begin{aligned} \sup_{\theta \in K} |W_n(\theta) - W_{\theta_0}(\theta)| &\leq \max_{1 \leq i \leq m} \sup_{\theta \in B(\theta^{(i)}, \delta)} \left\{ |W_n(\theta) - W_n(\theta^{(i)})| \right\} \\ &\quad + \max_{1 \leq i \leq m} |W_n(\theta^{(i)}) - W_{\theta_0}(\theta^{(i)})| + \max_{1 \leq i \leq m} \sup_{\theta \in B(\theta^{(i)}, \delta)} |W_{\theta_0}(\theta^{(i)}) - W_{\theta_0}(\theta)| . \end{aligned}$$

L'inégalité (II-2.6) montre que

$$\max_{1 \leq i \leq m} \sup_{\theta \in B(\theta^{(i)}, \delta)} |W_{\theta_0}(\theta^{(i)}) - W_{\theta_0}(\theta)| \leq \varepsilon .$$

La loi des grands nombres implique

$$\max_{1 \leq i \leq m} |W_n(\theta^{(i)}) - W_{\theta_0}(\theta^{(i)})| \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0 .$$

Nous concluons en remarquant que

$$\begin{aligned} \sup_{\theta \in B(\theta^{(i)}, \delta)} \left\{ |W_n(\theta) - W_n(\theta^{(i)})| \right\} &\leq \frac{1}{n} \sum_{j=1}^n \sup_{\theta \in B(\theta^{(i)}, \delta)} |\phi(\theta, X_j) - \phi(\theta^{(i)}, X_j)| \\ &\leq \frac{1}{n} \sum_{j=1}^n w_\delta(X_j) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \int w_\delta(x) \mathbb{Q}_{\theta_0}(dx) \leq \varepsilon . \quad \square \end{aligned}$$

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)). Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle statistique $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$ où Θ est un ensemble compact de $\mathbb{R}^d$. Supposons que

(i) Pour tout $x \in X$, la fonction $\theta \mapsto m(\theta, x)$ est continue.

(ii) Pour tout $\theta_0 \in \Theta$,

$$\int \sup_{\theta \in \Theta} |m(\theta, x)| \mathbb{Q}_{\theta_0}(dx) < \infty .$$

(iii) Pour tout $\theta \neq \theta_0 \in \Theta$, $M_{\theta_0}(\theta) < M_{\theta_0}(\theta_0)$ où $M_{\theta_0}(\theta) := \int m(\theta, x) \mathbb{Q}_{\theta_0}(dx)$.

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

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\rho_n \geq \varepsilon) = 0, \quad \text{pour tout } \varepsilon > 0 \quad (\text{II-2.7})$$

$$\liminf_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(M_n(\hat{\theta}_n) \geq M_n(\theta_0) - \rho_n) = 1. \quad (\text{II-2.8})$$

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 $\theta_0 \in \Theta$,

$$\sup_{\theta \in \Theta} |M_n(\theta) - M_{\theta_0}(\theta)| \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0,$$

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. $\square$

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)). Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle statistique $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta = \mathbb{R}^d\})$. Supposons que

(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 $\theta_0 \in \mathbb{R}^d$,

$$\int \sup_{\theta \in K} |m(\theta, x)| \mathbb{Q}_{\theta_0}(dx) < \infty .$$

(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

$$\int \sup_{\|\theta\| \geq a} |m(\theta, x)| \mathbb{Q}_{\theta_0}(dx) < \infty .$$

(v) Pour tout $\theta_0 \in \Theta$ et tout $x \in X$

$$\lim_{b \rightarrow \infty} \sup_{\|\theta\| > b} m(\theta, x) = -\infty .$$

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

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\rho_n \geq \varepsilon) = 0, \quad \text{pour tout } \varepsilon > 0 \quad (\text{II-2.9})$$

$$\liminf_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(M_n(\hat{\theta}_n) \geq M_n(\theta_0) - \rho_n) = 1. \quad (\text{II-2.10})$$

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

$$\lim_{b \rightarrow \infty} \int_X \left\{ \sup_{\|\theta\| > b} m(\theta, x) \right\} \mathbb{Q}_{\theta_0}(dx) = -\infty.$$

Soient $\theta_0 \in \mathbb{R}^d$ et b tel que

$$\int \left\{ \sup_{\|\theta\| > b} m(\theta, x) \right\} \mathbb{Q}_{\theta_0}(dx) < M_{\theta_0}(\theta_0). \quad (\text{II-2.11})$$

Remarquons que, par construction $b > \|\theta_0\|$. La loi forte des grands nombres montre que

$$\begin{aligned} \sup_{\|\theta\| > b} M_n(\theta) - M_n(\theta_0) &\leq \frac{1}{n} \sum_{i=1}^n \sup_{\|\theta\| > b} m(\theta, X_i) - m(\theta_0, X_i) \\ &\xrightarrow{\mathbb{P}_{\theta_0}\text{-prob}} \int \left\{ \sup_{\|\theta\| > b} m(\theta, x) \right\} \mathbb{Q}_{\theta_0}(dx) - M_{\theta_0}(\theta_0) < 0, \end{aligned}$$

et donc

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} \left(\sup_{\|\theta\| > b} M_n(\theta) - M_n(\theta_0) + \rho_n \geq 0 \right) = 0.$$

On pose $K := \overline{B(0, b)}$. Par construction, l'ensemble K est compact. On a

$$\begin{aligned} \{\hat{\theta}_n \notin K\} &\subset \{M_n(\hat{\theta}_n) \leq M_n(\theta_0) - \rho_n\} \cup \{\hat{\theta}_n \notin K, M_n(\hat{\theta}_n) \geq M_n(\theta_0) - \rho_n\} \\ &\subset \{M_n(\hat{\theta}_n) \leq M_n(\theta_0) - \rho_n\} \cup \left\{ \sup_{\|\theta\| \geq b} M_n(\theta) \geq M_n(\theta_0) - \rho_n \right\}. \end{aligned}$$

Donc

$$\limsup_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\hat{\theta}_n \notin K) = 0.$$

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 $\int |x| F(dx) < \infty$, $F(x) = 1 - F(-x)$ pour tout $x \in \mathbb{R}$ et $F(\eta) > 1/2$ pour tout $\eta > 0$. Pour $\theta \in \Theta = \mathbb{R}$ et $x \in X$, notons

$$F_\theta(x) = F(x - \theta),$$

et notons par $\mathbb{Q}_\theta$ la loi sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ de fonction de répartition F_θ . Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle statistique

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{Q}_\theta, \theta \in \Theta\}) .$$

Nous considérons le M -estimateur associé à la fonction $m(\theta, x) = -|x - \theta|$, i.e., pour tout $\theta, \theta_0 \in \Theta$,

$$M_n(\theta) = -\frac{1}{n} \sum_{i=1}^n |X_i - \theta| \quad \text{et} \quad M_{\theta_0}(\theta) = -\int |x - \theta| F_{\theta_0}(dx) .$$

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 $|m(\theta, x)| \leq |\theta| + |x|$ et, pour tout $\theta_0 \in \mathbb{R}$ et tout sous-ensemble compact $K \subset \mathbb{R}$,

$$\int \left\{ \sup_{\theta \in K} |m(\theta, x)| \right\} F_{\theta_0}(dx) \leq 2 \text{diam}(K) + \int |x| F(dx) < \infty ,$$

ce qui établit la condition théorème II-2.10-(ii). Pour tout $a \geq 0$ et $x \in \mathbb{X}$ on a

$$m(a, x) \leq \sup_{\|\theta\| \geq a} m(\theta, x) \leq 0$$

et donc

$$\left| \sup_{\|\theta\| \geq a} m(\theta, x) \right| \leq |m(a, x)| .$$

Comme, pour tout $\theta_0 \in \mathbb{R}$, $\int |m(a, x)| F_{\theta_0}(dx) < \infty$, la condition théorème II-2.10-(iv) est satisfaite. Comme $|\theta| - |x| \leq |\theta - x|$, nous avons $-|\theta - x| \leq |x| - |\theta|$ ce qui implique

$$\sup_{\|\theta\| \leq b} m(\theta, x) \leq |x| - b .$$

Par conséquent, pour tout $x \in \mathbb{R}$, $\lim_{b \rightarrow \infty} \sup_{\|\theta\| \leq b} m(\theta, x) = -\infty$ et la condition théorème II-2.10-(v) est satisfaite.

Considérons maintenant **H** II-2.19-(ii). Ici, pour tout $\theta, \theta_0 \in \Theta$,

$$\begin{aligned} M_{\theta_0}(\theta) &= -\int |x - \theta| F_{\theta_0}(dx) = -\int |x + (\theta_0 - \theta)| F(dx) \\ &= -\int_0^\infty \{|x + (\theta_0 - \theta)| + |x - (\theta_0 - \theta)|\} F(dx) . \end{aligned}$$

Un calcul élémentaire montre que, pour tout $\theta, \theta_0 \in \mathbb{R}$,

$$\int_0^\infty \{|x + (\theta - \theta_0)| + |x - (\theta - \theta_0)|\} F(dx) = \int_{-\infty}^\infty |x| F(dx) + 2 \int_0^{|\theta - \theta_0|} (|\theta - \theta_0| - x) F(dx) .$$

Comme, pour tout $\eta > 0$, $F(\eta) > 0$, $\int_0^\eta (\eta - x) F(dx) > 0$, on a donc, pour tout $\theta \neq \theta_0$, $M_{\theta_0}(\theta) < M_{\theta_0}(\theta_0)$. $\diamond$

II-2.2 Loi limite des Z- et M-estimateurs

Nous précisons les résultats de la section précédente, en cherchant une vitesse de convergence $\alpha_n \rightarrow \infty$ de sorte que l'erreur normalisée

$$\alpha_n(\hat{\theta}_n - \theta)$$

converge vers une 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 Loi limite des Z-estimateurs

Nous démontrons les résultats dans le cas $\Theta \subset \mathbb{R}$ pour simplifier. Étant données, d'une part une fonction $\psi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ définissant un Z-estimateur, et d'autre part un modèle statistique $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$, 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° de Θ ,

- (i) il existe un voisinage $\mathcal{V}(\theta_0)$ tel que pour tout $x \in X$, la fonction $\theta \mapsto \psi(\theta, x)$ est continûment différentiable sur $\mathcal{V}(\theta_0)$. Il existe une fonction mesurable g , telle que, pour tout $x \in X$,

$$\sup_{\theta \in \mathcal{V}(\theta_0)} |\psi(\theta, x)| \leq g(x), \quad \text{et} \quad \mathbb{E}_{\theta_0}[g(X)] < +\infty.$$

où

$$\dot{\psi}(\theta, x) := \frac{\partial \psi}{\partial \theta}(\theta, x).$$

De plus, $\int \dot{\psi}(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) \neq 0$.

- (ii) $\int \psi^2(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) < +\infty$ et $\int \psi(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) = 0$. ◇

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 $\theta_0 \in \Theta^\circ$. Ceci exclut les paramètres de la frontière de Θ dans le cas où Θ n'est pas un ouvert. ◇

Remarque II-2.14. La condition II-2.12-(i) est vérifiée dans de nombreux exemples, mais n'est pas vérifiée par exemple si $\psi(\theta, x) = \text{signe}(x - \theta)$. La normalité asymptotique de la médiane peut être établie, mais en utilisant d'autres méthodes comme nous l'avons vu en (IV-4.24). ◇

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). Supposons **H II-2.12**. Soit $\{\hat{\theta}_n, n \in \mathbb{N}\}$ une suite d'estimateurs consistante telle que, pour tout $\theta_0 \in \Theta^\circ$,

$$n^{1/2} \|\Psi_n(\hat{\theta}_n)\| \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0. \quad (\text{II-2.12})$$

Pour tout $\theta_0 \in \Theta^\circ$, nous avons

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, v_\psi(\theta_0))$$

où pour tout $\theta \in \Theta$,

$$v_\psi(\theta) := \frac{\mathbb{E}_\theta[\psi^2(\theta, X_1)]}{\left(\mathbb{E}_\theta[\dot{\psi}(\theta, X_1)]\right)^2} = \frac{\int \psi^2(\theta, x) \mathbb{Q}_\theta(dx)}{\left(\int \dot{\psi}(\theta, x) \mathbb{Q}_\theta(dx)\right)^2}.$$

Démonstration. Soit $x \in X$ un point tel que $\theta \rightarrow \psi(\theta, x)$ soit continûment dérivable sur $\mathcal{V}(\theta_0)$. Pour tout t tel que $\theta_0 + t \in \mathcal{V}(\theta_0)$, nous avons

$$\psi(\theta_0 + t, x) = \psi(\theta_0, x) + \dot{\psi}(\theta_0, x)t + r(t, x) \quad (\text{II-2.13})$$

où

$$r(t, x) := \int_0^1 \{\dot{\psi}(\theta_0 + wt, x) - \dot{\psi}(\theta_0, x)\} dw. \quad (\text{II-2.14})$$

Pour $\delta > 0$, définissons

$$R_\delta(x) = \sup_{|t| \leq \delta} |r(t, x)|.$$

Nous avons $R_\delta(x) \leq 2g(x)$ et $\lim_{\delta \rightarrow 0} R_\delta(x) = 0$. Le théorème de convergence dominée implique que $\lim_{\delta \rightarrow 0} \int R_\delta(x) \mathbb{Q}_{\theta_0}(dx) = 0$. En utilisant (II-2.13), nous avons, $\mathbb{P}_{\theta_0}$ -presque sûrement

$$\Psi_n(\theta_0 + t) = \frac{1}{n} \sum_{i=1}^n \psi(\theta_0, X_i) + t \int \dot{\psi}(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) + t R_n(t), \quad (\text{II-2.15})$$

où nous avons posé $R_n(t) = R_n^{(1)} + R_n^{(2)}(t)$ avec

$$R_n^{(1)} := \frac{1}{n} \sum_{i=1}^n \left\{ \psi(\theta_0, X_i) - \int \psi(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) \right\}$$

$$R_n^{(2)}(t) := \frac{1}{n} \sum_{i=1}^n r(t, X_i) .$$

Sous **H** II-2.12-(i), la loi faible des grands nombres (Théorème IV-2.18) montre que

$$R_n^{(1)} \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0 . \quad (\text{II-2.16})$$

Pour toute suite $\{\delta_n, n \in \mathbb{N}\}$ telle que $\lim_{n \rightarrow \infty} \delta_n = 0$, nous avons

$$\sup_{|t| \leq \delta_n} |R_n^{(2)}(t)| \leq \frac{1}{n} \sum_{i=1}^n R_{\delta_n}(X_i) \quad \text{et} \quad \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta_0}[R_{\delta_n}(X_i)] = \mathbb{E}_{\theta_0}[R_{\delta_n}(X_1)] \rightarrow_{n \rightarrow \infty} 0 .$$

Par l'inégalité de Markov (lemme IV-3.1), nous avons donc

$$\sup_{|t| \leq \delta_n} |R_n^{(2)}(t)| \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0 . \quad (\text{II-2.17})$$

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 $\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(|\Delta_n| \geq \delta_n) = 0$. Comme, pour tout $\varepsilon > 0$,

$$\mathbb{P}_{\theta_0}(|R_n^{(2)}(\Delta_n)| \geq \varepsilon) \leq \mathbb{P}_{\theta_0}(\sup_{|t| \leq \delta_n} |R_n^{(2)}(t)| \geq \varepsilon) + \mathbb{P}_{\theta_0}(|\Delta_n| \geq \delta_n) ,$$

on a $R_n^{(2)}(\Delta_n) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0$. En combinant ce résultat avec (II-2.17), nous obtenons

$$R_n(\Delta_n) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0 , \quad (\text{II-2.18})$$

et donc, en utilisant (II-2.15),

$$\sqrt{n} \Psi_n(\hat{\theta}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(\theta_0, X_i) + \sqrt{n}(\hat{\theta}_n - \theta_0) \{ \mathbb{E}_{\theta_0}[\psi(\theta_0, X_1)] + R_n(\Delta_n) \} , \quad (\text{II-2.19})$$

Le T.L.C. (Théorème IV-2.39) montre que, sous **H** II-2.12-(ii),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(\theta_0, X_i) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \mathbb{E}_{\theta_0}[\psi^2(\theta_0, X_1)]) .$$

On conclut en appliquant le lemme IV-2.33. □

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-dégénérée¹ avec la normalisation $\sqrt{n}$ implique que si l’on choisit une autre normalisation $\alpha_n \rightarrow \infty$, alors l’erreur normalisée

$$\alpha_n(\hat{\theta}_n - \theta)$$

tend vers 0 en probabilité si $\alpha_n/\sqrt{n} \rightarrow 0$ et “explose”² si $\alpha_n/\sqrt{n} \rightarrow \infty$. ◇

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, ψ est une fonction $\mathbb{R}^d \times \mathbf{X} \rightarrow \mathbb{R}^d$. Considérons les hypothèses suivantes.

1. C’est-à-dire une loi gaussienne de matrice de variance-covariance $V(\theta)$ non singulière.

2. Dans le sens suivant : pour tout $M > 0$, $\liminf_{n \rightarrow \infty} \mathbb{P}_{\theta} [|\alpha_n(\hat{\theta}_n - \theta)| \geq M] > 0$.

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 $x \in X$, la fonction $\theta \mapsto \psi(\theta, x)$ est continûment différentiable sur un voisinage $\mathcal{V}(\theta_0)$ et il existe une fonction mesurable g telle que, pour tout $x \in X$,

$$\sup_{\theta \in \mathcal{V}(\theta_0)} \sum_{i=1}^d \left\| \frac{\partial \psi}{\partial \theta^{(i)}}(\theta, x) \right\| \leq g(x) \quad \text{et} \quad \int g(x) \mathbb{Q}_{\theta_0}(dx) < \infty.$$

De plus, la matrice $J_\psi(\theta_0)$ est inversible, où

$$J_\psi(\theta_0) := \left[\int \left[\frac{\partial \psi}{\partial \theta^{(1)}}(\theta_0, x) \right] \mathbb{Q}_{\theta_0}(dx) \dots \int \left[\frac{\partial \psi}{\partial \theta^{(d)}}(\theta_0, x) \right] \mathbb{Q}_{\theta_0}(dx) \right]. \quad (\text{II-2.20})$$

- (ii) $\int \|\psi(\theta_0, x)\|^2 \mathbb{Q}_{\theta_0}(dx) < \infty$ et $\int \psi(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) = 0$. ◇

Théorème II-2.18 (Loi limite des Z-estimateurs : cas vectoriel). Supposons **H II-2.17**. Soit $\{\hat{\theta}_n, n \in \mathbb{N}\}$ une suite d'estimateurs consistante telle que, pour tout $\theta_0 \in \Theta$

$$\sqrt{n} \Psi_n(\hat{\theta}_n) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0. \quad (\text{II-2.21})$$

Alors, pour tout $\theta_0 \in \Theta^o$, nous avons

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, [J_\psi(\theta_0)]^{-1} G_\psi(\theta_0) [J_\psi(\theta_0)]^{-T}), \quad (\text{II-2.22})$$

où $J_\psi(\theta_0)$ est définie en (II-2.20) et

$$G_\psi(\theta_0) := \int \psi(\theta_0, x) \psi^T(\theta_0, x) \mathbb{Q}_{\theta_0}(dx). \quad (\text{II-2.23})$$

Démonstration. La preuve est tout à fait similaire au cas scalaire. □

II-2.2.2 Loi limite 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 fonction $\theta \mapsto m(\theta, x)$ est différentiable sur Θ^o pour tout $x \in X$ 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

$$\nabla m(\theta, x) := \left[\frac{\partial m}{\partial \theta^{(1)}}(\theta, x), \dots, \frac{\partial m}{\partial \theta^{(d)}}(\theta, x) \right]^T. \quad (\text{II-2.24})$$

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 $x \in \mathbf{X}$, la fonction $\theta \mapsto m(\theta, x)$ est deux fois continûment différentiable sur un voisinage $\mathcal{V}(\theta_0)$. Il existe une fonction mesurable g telle que, pour tout $x \in \mathbf{X}$,

$$\sup_{\theta \in \mathcal{V}(\theta_0)} \|\mathbf{H}_m(\theta, x)\| \leq g(x) \quad \text{et} \quad \int g(x) \mathbb{Q}_{\theta_0}(dx) < \infty,$$

où $\mathbf{H}_m(\theta, x)$ désigne la matrice

$$\mathbf{H}_m(\theta, x) := \left[\frac{\partial^2 m}{\partial \theta^{(i)} \partial \theta^{(j)}}(\theta, x) \right]_{1 \leq i, j \leq d}.$$

De plus, la matrice $\int \mathbf{H}_m(\theta_0, x) \mathbb{Q}_{\theta_0}(dx)$ est inversible.

- (ii) $\int \|\nabla m(\theta_0, x)\|^2 \mathbb{Q}_{\theta_0}(dx) < \infty$ et $\int \nabla m(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) = 0$.

◇

Théorème II-2.20 (Loi limite des M-estimateurs : cas vectoriel). Supposons **H II-2.19**. Soit $\{\hat{\theta}_n, n \in \mathbb{N}\}$ une suite consistante d'estimateurs telle que, pour tout $\theta_0 \in \Theta^o$,

$$\sqrt{n} \nabla M_n(\hat{\theta}_n) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0. \quad (\text{II-2.25})$$

Alors, pour tout $\theta_0 \in \Theta^o$, nous avons

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} \mathbf{N}(0, U_m(\theta_0) G_m(\theta_0) U_m^T(\theta_0)) \quad (\text{II-2.26})$$

où

$$G_m(\theta_0) := \int \left[\nabla m(\theta_0, x) (\nabla m(\theta_0, x))^T \right] \mathbb{Q}_{\theta_0}(dx), \quad (\text{II-2.27})$$

$$U_m(\theta_0) := \left(\int \mathbf{H}_m(\theta_0, x) \mathbb{Q}_{\theta_0}(dx) \right)^{-1}. \quad (\text{II-2.28})$$

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

II-3.1 Consistance de l'estimateur du M.V.

La méthode d'estimation au sens du maximum de vraisemblance 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 $(X, \mathcal{X})$. Considérons $\{X_j, j \in \mathbb{N}\}$ un échantillon du modèle

$$(X, \mathcal{X}, \{q_\theta \cdot \mu : \theta \in \Theta\})$$

L'estimateur du maximum de vraisemblance $\hat{\theta}_n^{MV}$, s'il existe, peut s'interpréter comme le M -estimateur associé à la fonction

$$\ell(\theta, x) = \log q(\theta, x) \quad \text{où} \quad q(\theta, x) = q_\theta(x).$$

Le théorème I-2.24 montre que, pour tout $\theta_0 \in \Theta$, la valeur $\theta = \theta_0$ maximise la fonction

$$\theta \rightarrow M_{\theta_0}(\theta) := \int \log q(\theta, x) \mathbb{Q}_{\theta_0}(dx) = -\text{KL}(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta) + \int \log q(\theta_0, x) \mathbb{Q}_{\theta_0}(dx). \quad (\text{II-3.1})$$

Si, pour μ -presque tout $x \in X$, la fonction $\theta \rightarrow \log q(\theta, x)$ est différentiable, alors le maximum de vraisemblance peut être aussi vu comme un Z -estimateur associé à la fonction

$$\psi(\theta, x) = \nabla \log q(\theta, x) = \left[\frac{\partial \log q}{\partial \theta^{(1)}}(\theta, x), \dots, \frac{\partial \log q}{\partial \theta^{(d)}}(\theta, x) \right] \quad \theta \in \Theta, x \in \mathbb{R}$$

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)). *Supposons que Θ est un compact de $\mathbb{R}^d$ et que,*

- (i) *pour tout $x \in X$, la fonction $\theta \mapsto \log q(\theta, x)$ est continue.*
- (ii) *pour tout $\theta_0 \in \Theta$,*

$$\int \sup_{\theta \in \Theta} |\log q(\theta, x)| \mathbb{Q}_{\theta_0}(dx) < \infty.$$

- (iii) *pour tout $\theta \neq \theta_0 \in \Theta$, $\mathbb{Q}_\theta \neq \mathbb{Q}_{\theta_0}$.*

Alors, pour tout $\theta_0 \in \Theta$, la fonction

$$\theta \mapsto \int \log q(\theta, x) \mathbb{Q}_{\theta_0}(dx)$$

a un maximum unique au point $\theta = \theta_0$. De plus, si pour tout $\theta_0 \in \Theta$, l'estimateur $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ vérifie

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\ell_n(\hat{\theta}_n^{\text{MV}}) \geq \ell_n(\theta_0)) = 1, \quad \text{où} \quad \ell_n(\theta) := n^{-1} \sum_{i=1}^n \log q(\theta, X_i),$$

alors $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ est consistant.

Démonstration. Il suffit d'appliquer le théorème II-2.9 avec

$$m(\theta, x) := \log q(\theta, x).$$

Il faut donc vérifier les conditions théorème II-2.9-(i)-(ii) découlent des conditions (i) et (ii). Par le Théorème I-2.24 et (II-3.1), sous la condition (iii), $M_{\theta_0}(\theta_0) > M_{\theta_0}(\theta)$. $\square$

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)). Supposons que

- (i) pour μ -presque tout x , la fonction $\theta \mapsto \log q(\theta, x)$ est continue sur $\mathbb{R}^d$.
- (ii) pour tout $\theta_0 \in \mathbb{R}^d$ et tout compact K de $\mathbb{R}^d$,

$$\int \sup_{\theta \in K} |\log q(\theta, x)| \mathbb{Q}_{\theta_0}(dx) < \infty.$$

- (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

$$\int \left| \sup_{\|\theta\| \geq a} \log q(\theta, x) \right| \mathbb{Q}_{\theta_0}(dx) < \infty.$$

- (v) Pour μ -presque tout x ,

$$\lim_{b \rightarrow \infty} \sup_{\|\theta\| \geq b} \log q(\theta, x) = -\infty.$$

Si pour tout $\theta_0 \in \mathbb{R}^d$, l'estimateur $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ vérifie

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\ell_n(\hat{\theta}_n^{\text{MV}}) \geq \ell_n(\theta_0)) = 1, \quad \text{où} \quad \ell_n(\theta) := n^{-1} \sum_{i=1}^n \log q(\theta, X_i),$$

alors $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ est consistant.

Démonstration. C'est une application immédiate du théorème II-2.10. $\square$

II-3.2 Loi limite de l'estimateur du Maximum de Vraisemblance

Fisher,
information
de

Nous nous intéresserons dans la suite aux modèles $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ réguliers, au sens de la définition I-4.20. Rappelons que, pour un modèle régulier $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$, la positivité des densités μ -p.p. permet de définir la *log-vraisemblance*

$$\ell(\theta, x) := \log q_\theta(x)$$

et son gradient $\nabla \ell(\theta, x)$ (bien défini par régularité du modèle), appelé *fonction score* ou simplement *score de Fisher*.

Pour un modèle régulier, l'information de Fisher $\mathbb{I}(\theta)$ (définition I-4.22) apparaît naturellement comme la matrice de covariance asymptotique de l'estimateur du maximum de vraisemblance.

Théorème II-3.3 (Loi asymptotique de l'estimateur du M.V.). *Supposons que le modèle $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ est régulier (voir définition I-4.20). Soit $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ un estimateur consistant vérifiant, pour tout $\theta_0 \in \Theta$,*

$$\sqrt{n} \nabla \ell_n(\hat{\theta}_n^{\text{MV}}) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0, \quad \text{où } \ell_n(\theta) := n^{-1} \sum_{i=1}^n \ell(\theta, X_i).$$

Alors, pour tout $\theta_0 \in \Theta$,

$$\sqrt{n}(\hat{\theta}_n^{\text{MV}} - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \mathbb{I}^{-1}(\theta_0)).$$

Démonstration. Il s'agit d'une application directe du théorème II-2.20, en notant que, par la Proposition I-4.23 :

$$\mathbb{I}(\theta) = \int \nabla \ell(\theta, x) \nabla^T \ell(\theta, x) \mathbb{Q}_\theta(dx) = - \int H_\ell(\theta, x) \mathbb{Q}_\theta(dx), \quad (\text{II-3.2})$$

cette matrice étant en outre inversible. $\square$

Exemple II-3.4. Soit $\{X_k, k \in \mathbb{N}\}$ une suite de variables aléatoires réelles indépendantes de densité par rapport à la mesure de Lebesgue donnée par

$$q(\theta, x) = \begin{cases} (\theta_1 + \theta_2)^{-1} e^{-x/\theta_1} & x > 0 \\ (\theta_1 + \theta_2)^{-1} e^{x/\theta_2} & x \leq 0, \end{cases}$$

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 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

$$\begin{aligned} \nabla \ell_n(\theta) &= \left(-\frac{1}{\theta_1 + \theta_2} + \frac{T_{1,n}}{\theta_1^2}, -\frac{1}{\theta_1 + \theta_2} + \frac{T_{2,n}}{\theta_2^2} \right), \\ H_{\ell_n}(\theta) &= \begin{pmatrix} \frac{1}{(\theta_1 + \theta_2)^2} - \frac{2T_{1,n}}{\theta_1^3} & \frac{1}{(\theta_1 + \theta_2)^2} \\ \frac{1}{(\theta_1 + \theta_2)^2} & \frac{1}{(\theta_1 + \theta_2)^2} - \frac{2T_{2,n}}{\theta_2^3} \end{pmatrix}. \end{aligned}$$

Nous vérifions aisément que les équations de vraisemblance admettent une unique solution, qui correspond à l'estimateur du maximum de vraisemblance, donné par

$$\hat{\theta}_n^{\text{MV}} = (\sqrt{T_{1,n} T_{2,n}} + T_{1,n}, \sqrt{T_{1,n} T_{2,n}} + T_{2,n})$$

efficacité
asymptotique

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 $\theta \in \Theta$,

$$\begin{aligned} T_{1,n} &\xrightarrow{\mathbb{P}_\theta - \text{prob}} \frac{\theta_1^2}{\theta_1 + \theta_2} \\ T_{2,n} &\xrightarrow{\mathbb{P}_\theta - \text{prob}} \frac{\theta_2^2}{\theta_1 + \theta_2} \end{aligned}$$

et donc $\hat{\theta}_n^{\text{MV}} \xrightarrow{\mathbb{P}_\theta - \text{prob}} (\theta_1, \theta_2)$. Une application directe du théorème II-3.3 montre que

$$\sqrt{n}(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, \mathbb{I}^{-1}(\theta))$$

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-3.3 Efficacité asymptotique

Nous avons étudié au chapitre précédent la construction d'estimateurs basés sur la maximisation d'un critère ,

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} n^{-1} \sum_{i=1}^n m(\theta, X_i), \quad m : \Theta \times \mathcal{X} \rightarrow \mathbb{R}$$

ou sur la résolution d'un système d'équations

$$n^{-1} \sum_{i=1}^n \psi(\theta, X_i) = 0, \quad \psi : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^d .$$

Nous avons montré que, sous des hypothèses de régularité, la suite d'estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$ était (faiblement ou fortement) consistante et asymptotiquement normale

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, V(\theta)) \quad (\text{II-3.3})$$

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.

Nous nous plaçons dans cette section dans le cas de la dimension 1, avec $\Theta \subset \mathbb{R}$ pour simplifier. Les extensions au cas multidimensionnel se font de la même manière que pour la section II-3.2. On se restreint ici 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 quantité $v(\theta) > 0$, appelée variance asymptotique, telle que

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, v(\theta)) .$$

On suppose de plus :

H II-3.5. 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-3.5**. 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-3.6. Si $\{\hat{\theta}_{n,1}, n \in \mathbb{N}\}$ et $\{\hat{\theta}_{n,2}, n \in \mathbb{N}\}$ sont deux suites d'estimateurs asymptotiquement normaux de variances asymptotiques respectives $v_1(\theta)$ et $v_2(\theta)$ et vérifiant **H II-3.5**, on dit que $\hat{\theta}_{n,1}$ est plus efficace que $\hat{\theta}_{n,2}$ si pour tout $\theta \in \Theta$,

$$v_1(\theta) \leq v_2(\theta)$$

et si de plus, il existe un point $\tilde{\theta} \in \Theta$ tel que

$$v_1(\tilde{\theta}) < v_2(\tilde{\theta}).$$

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-3.7. 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 . $\diamond$

Efficacité asymptotique du maximum de vraisemblance

Dans cette section, on considère une expérience statistique régulière (au sens de la Définition I-4.20), et on se restreint à la classe des Z-estimateurs, qui contient en particulier les M-estimateurs.

Théorème II-3.8 (Efficacité asymptotique de l'E.M.V. dans la classe des Z-estimateurs). Soit $(X, \mathcal{X}, \{q_\theta : \theta \in \Theta\})$ un modèle statistique régulier et $\hat{\theta}_n$ 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

$$v_\psi(\theta) = \frac{\int \psi^2(\theta, x) \mathbb{Q}_\theta(dx)}{(\int \psi'(\theta, x) \mathbb{Q}_\theta(dx))^2} \geq \frac{1}{\mathbb{I}(\theta)}.$$

Corollaire II-3.9. Dans un modèle régulier, l'estimateur du maximum de vraisemblance est asymptotiquement efficace parmi les Z-estimateurs réguliers.

Démonstration. Puisque $\mathbb{E}_\theta[\psi(\theta, X)] = 0$ pour tout $\theta \in \Theta^\circ$, la dérivée est nulle, ce qui donne

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \psi'(\theta, x) q(\theta, x) \mu(dx) + \int_{\mathbb{R}} \psi(\theta, x) \frac{\partial q}{\partial \theta}(\theta, x) \mu(dx) \\ &= \int_{\mathbb{R}} \psi'(\theta, x) q(\theta, x) \mu(dx) + \int_{\mathbb{R}} \psi(\theta, x) \dot{\ell}(\theta, x) q(\theta, x) \mu(dx), \end{aligned}$$

c'est-à-dire

$$\int \psi'(\theta, x) \mathbb{Q}_\theta(dx) = - \int \psi(\theta, x) \dot{\ell}(\theta, x) \mathbb{Q}_\theta(dx).$$

□

En appliquant l'inégalité de Cauchy-Schwarz, on obtient

$$\left(\int \psi'(\theta, x) \mathbb{Q}_\theta(dx) \right)^2 \leq \left(\int \psi(\theta, x)^2 \mathbb{Q}_\theta(dx) \right) \left(\int (\dot{\ell}(\theta, x))^2 \mathbb{Q}_\theta(dx) \right),$$

c'est-à-dire

$$v_\psi(\theta)^{-1} = \frac{(\int \psi'(\theta, x) \mathbb{Q}_\theta(dx))^2}{\int \psi(\theta, x)^2 \mathbb{Q}_\theta(dx)} \leq \int (\dot{\ell}(\theta, X))^2 \mathbb{Q}_\theta(dx) = \mathbb{I}(\theta).$$

La conjecture de Fisher

En 1922, Fisher conjectura que, pour un modèle régulier, dans un sens comparable avec celui donné à la I-4.20,

- (i) l'estimateur du maximum de vraisemblance converge et a pour variance asymptotique $\frac{1}{\mathbb{I}(\theta)}$.

Fisher,
programme
de

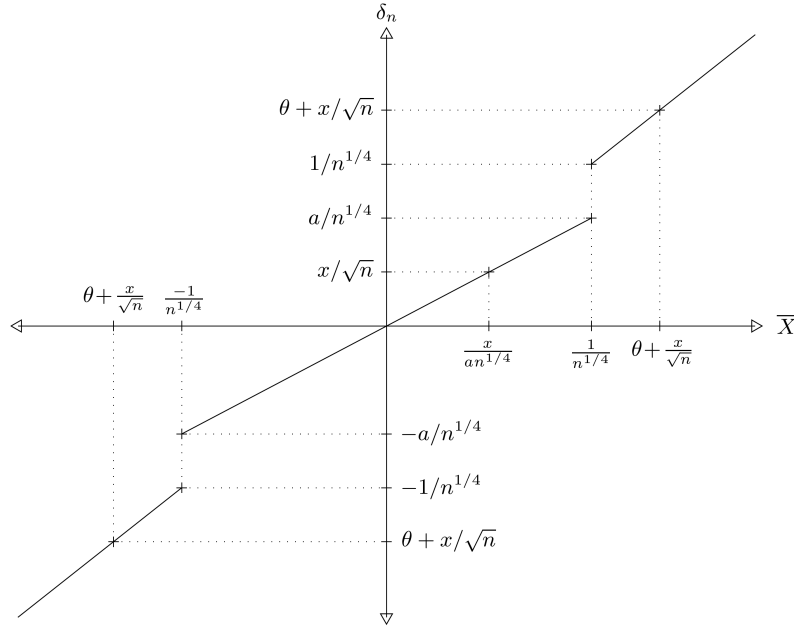


FIGURE II-3.1 – Estimateur de Hodges-Lehmann

- (ii) si, pour une suite d'estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$, on a $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, v(\theta))$, 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.

Soit $\{X_k, k \in \mathbb{N}\}$ une suite de variables i.i.d. distribuées suivant la loi $N(\theta, 1)$. Posons $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$. Définissons l'estimateur δ_n , donné par

$$\delta_n := \begin{cases} \bar{X}_n, & |\bar{X}_n| \geq 1/n^{1/4}, \\ a\bar{X}_n, & |\bar{X}_n| < 1/n^{1/4}, \end{cases}$$

où a est une constante de $]0, 1[$.

Déterminons la distribution de $\sqrt{n}(\delta_n - \theta)$. Supposons tout d'abord que $\theta < 0$. Fixons x et considérons

$$\mathbb{P}_\theta(\sqrt{n}(\delta_n - \theta) \leq x) = \mathbb{P}_\theta(\delta_n \leq \theta + x/\sqrt{n}).$$

Comme $\theta + x/\sqrt{n} \rightarrow \theta < 0$ et $-1/n^{1/4} \rightarrow 0$, pour n suffisamment grand, $\theta + x/\sqrt{n} < -1/n^{1/4}$, et donc

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(\sqrt{n}(\delta_n - \theta) \leq x) = \lim_{n \rightarrow \infty} \mathbb{P}_\theta(\bar{X}_n \leq \theta + x/\sqrt{n}) = \Phi(x),$$

où Φ désigne la fonction de répartition d'une gaussienne centrée réduite. Par conséquent, $\sqrt{n}(\delta_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, 1)$. Un calcul en tout point similaire montre que $\sqrt{n}(\delta_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, 1)$ pour $\theta > 0$.

Supposons maintenant que $\theta = 0$. Fixons x et considérons

$$\mathbb{P}_0(\sqrt{n}\delta_n \leq x) = \mathbb{P}_0(\delta_n \leq x/\sqrt{n}).$$

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 $\mathbb{P}_0$, $\sqrt{n}\bar{X}_n \sim N(0, 1)$; cela entraîne que $\mathbb{1}_{\{|\sqrt{n}\bar{X}_n| \geq n^{1/4}\}} \xrightarrow{\mathbb{P}_0\text{-prob}} 0$, puis, par le lemme de Slutsky, (voir lemme IV-2.33) on a

$$\sqrt{n}\delta_n \xrightarrow{\mathbb{P}_0} N(0, a^2).$$

Par conséquent, nous avons, pour tout $\theta \in \mathbb{R}$,

$$\sqrt{n}(\delta_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, \sigma^2(\theta)), \quad (\text{II-3.4})$$

où

$$\sigma^2(\theta) = \begin{cases} 1, & \theta \neq 0 \\ a^2, & \theta = 0 \end{cases}. \quad (\text{II-3.5})$$

Cet estimateur est “super efficace” car sa variance asymptotique pour $\theta = 0$ est inférieure à $1/\mathbb{I}(\theta) = 1$.

Comme $\sqrt{n}(\bar{X}_n - \theta) \sim N(0, 1)$, (II-3.4) semble suggérer que δ_n pourrait être asymptotiquement un meilleur estimateur que $\bar{X}_n$, au moins lorsque n est grand. Pour comprendre ce phénomène, nous allons évaluer le risque de ces deux estimateurs pour la perte quadratique. Comme $R(\theta, \bar{X}_n) = \mathbb{E}_\theta[(\bar{X}_n - \theta)^2] = 1/n$, $nR(\theta, \bar{X}_n) = 1$. Un calcul élémentaire montre que

$$nR(\theta, \delta_n) \rightarrow \begin{cases} 1, & \theta \neq 0 \\ a^2, & \theta = 0 \end{cases},$$

comme le suggère d’ailleurs (II-3.5). La comparaison de δ_n et $\bar{X}_n$ 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

$$(\delta_n - \theta_n)^2 \geq \left(\frac{1-a}{2n^{1/4}}\right)^2 = \frac{(1-a)^2}{4\sqrt{n}}.$$

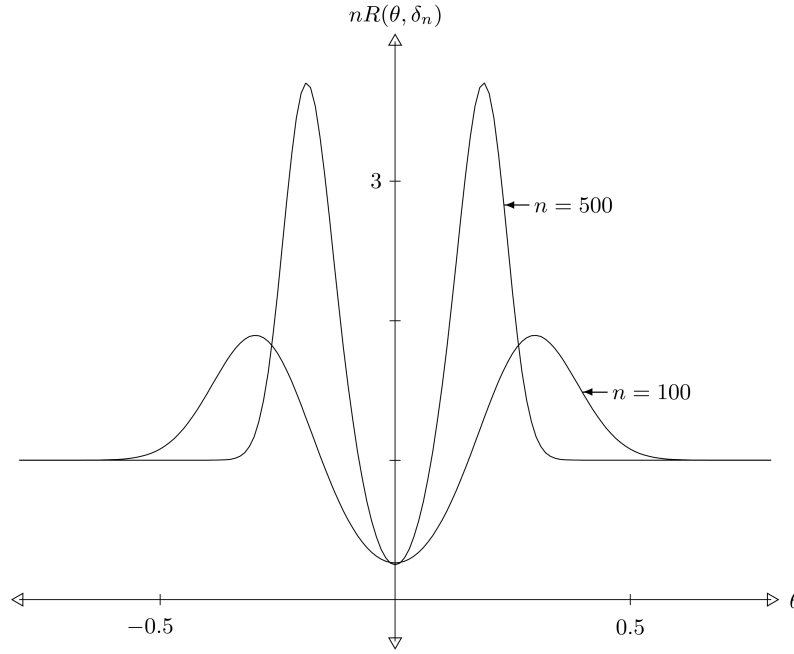
Nous en déduisons

$$nR(\theta_n, \delta_n) \geq n \frac{(1-a)^2}{4\sqrt{n}} = \frac{(1-a)^2}{4} \sqrt{n} \xrightarrow{n \rightarrow \infty} \infty.$$

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 $\bar{X}_n$ à 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. 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.

Conclusion

1. Concernant la notion de modèle régulier, par souci de simplicité, nous nous sommes restreints à un jeu d’hypothèses assez fortes. On peut étendre significativement les hypothèses de régularité.
2. La comparaison asymptotique d’estimateurs reste une notion fragile et *ad-hoc*. Un point de vue alternatif est la recherche d’uniformité en le paramètre (approche minimax).

FIGURE II-3.2 – $n\mathbb{E}_\theta[(\delta_n - \theta)^2]$ pour $n = 100$ et $n = 500$, $a = 1/2$

II-3.4 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-3.10. *Supposons le modèle $(X, \mathcal{X}, \{q_\theta, \theta \in \Theta\})$ régulier. Soit $\{\hat{\theta}_n, n \in \mathbb{N}\}$ un estimateur asymptotiquement normal. Alors,*

$$\nabla \ell_n(\hat{\theta}_n) = \nabla \ell_n(\theta_0) - \mathbb{I}(\theta_0)(\hat{\theta}_n - \theta_0) + o_P(n^{-1/2}),$$

où $\nabla \ell_n(\theta) := n^{-1} \sum_{i=1}^n \nabla \ell(\theta, X_i)$.

Démonstration. Nous avons

$$\begin{aligned} \nabla \ell_n(\hat{\theta}_n) &= n^{-1} \sum_{i=1}^n \nabla \ell(\hat{\theta}_n, X_i) \\ &= n^{-1} \sum_{i=1}^n \nabla \ell(\theta_0, X_i) + n^{-1} \sum_{i=1}^n H_\ell(\theta_0, X_i)(\hat{\theta}_n - \theta_0) \\ &\quad + n^{-1} \sum_{i=1}^n \left[\int_0^1 \{H_\ell(\theta_0 + w(\hat{\theta}_n - \theta_0), X_i) - H_\ell(\theta_0, X_i)\} dw \right] (\hat{\theta}_n - \theta_0) \\ &= n^{-1} \sum_{i=1}^n \nabla \ell(\theta_0, X_i) - \mathbb{I}_n(\theta_0)\{\hat{\theta}_n - \theta_0\} + \Delta_n(\hat{\theta}_n - \theta_0), \end{aligned}$$

où

$$\Delta_n := - \int_0^1 \{\mathbb{I}_n(\theta_0 + w(\hat{\theta}_n - \theta_0)) - \mathbb{I}_n(\theta_0)\} dw.$$

Par conséquent, en utilisant la proposition II-4.9, nous avons, pour tout $\varepsilon > 0$,

$$\mathbb{P}_{\theta_0}(\|\Delta_n\| \geq \varepsilon) \leq \mathbb{P}_{\theta_0}(\|\hat{\theta}_n - \theta_0\| \geq \delta_n) + \mathbb{P}_{\theta_0} \left(\sup_{\|t\| \leq \delta_n} \|\mathbb{I}_n(\theta_0 + t) - \mathbb{I}_n(\theta_0)\| \geq \varepsilon \right) \xrightarrow{n \rightarrow \infty} 0,$$

ce qui montre que $\Delta_n \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0$. Comme $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est asymptotiquement normal, nous avons, en utilisant le lemme IV-2.33, que

$$\sqrt{n}\Delta_n(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0,$$

où 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 $\theta_0 \in \Theta$, $\mathbb{I}_n(\theta_0) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \mathbb{I}(\theta_0)$ et donc, en utilisant le lemme IV-2.33,

$$\sqrt{n}\{\mathbb{I}_n(\theta_0) - \mathbb{I}(\theta_0)\}(\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0. \quad \square$$

Théorème II-3.11 (Score de Fisher). *Supposons le modèle $(\mathbf{X}, \mathcal{X}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ 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 $\theta_0 \in \Theta$,

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \mathbb{I}^{-1}(\theta_0)). \quad (\text{II-3.6})$$

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-3.6).

Démonstration. Le lemme II-3.10 montre que

$$\nabla \ell_n(\hat{\theta}_n) = \nabla \ell_n(\theta_0) - \mathbb{I}(\theta_0)(\hat{\theta}_n - \theta_0) + o_P(n^{-1/2}). \quad (\text{II-3.7})$$

Comme $\sqrt{n}\nabla \ell_n(\theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \mathbb{I}(\theta_0))$ (noter que par la proposition I-4.23, $\mathbb{E}_{\theta_0}[\nabla \ell(\theta_0, X)] = 0$), nous avons $\sqrt{n}\nabla \ell_n(\theta_0) = o_P(1)$ (conséquence de lemme IV-2.58). Nous avons de même, comme $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est asymptotiquement normal, $\sqrt{n}(\hat{\theta}_n - \theta_0) = o_P(1)$. Nous en déduisons que

$$\nabla \ell_n(\hat{\theta}_n) = o_P(n^{-1/2}). \quad (\text{II-3.8})$$

Comme la fonction $A \mapsto A^{-1}$ est continue en toute matrice A_0 inversible, le théorème IV-2.6 et la proposition II-4.9 montrent que

$$\mathbb{I}_n^{-1}(\hat{\theta}_n) = \mathbb{I}_n^{-1}(\theta_0) + o_P(1). \quad (\text{II-3.9})$$

En combinant (II-3.8) et (II-3.9) nous obtenons donc

$$\begin{aligned} \tilde{\theta}_n &= \hat{\theta}_n + \mathbb{I}_n^{-1}(\hat{\theta}_n) \nabla \ell_n(\hat{\theta}_n) \\ &= \hat{\theta}_n + \mathbb{I}_n^{-1}(\theta_0) \nabla \ell_n(\hat{\theta}_n) + o_P(n^{-1/2}) \\ &= \hat{\theta}_n + \mathbb{I}_n^{-1}(\theta_0) \{\nabla \ell_n(\theta_0) - \mathbb{I}(\theta_0)(\hat{\theta}_n - \theta_0)\} + o_P(n^{-1/2}) \end{aligned}$$

Nous en déduisons que

$$\tilde{\theta}_n - \theta_0 = \mathbb{I}^{-1}(\theta_0) \nabla \ell_n(\theta_0) + o_P(n^{-1/2}), \quad (\text{II-3.10})$$

ce qui montre (II-3.6). $\square$

Exemple II-3.12. 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 $X_1, \dots, X_n$, 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 $\int |f_1(x) - f_2(x)| dx > 0$. Soit $\{X_k, k \in \mathbb{N}\}$ une suite de variables aléatoires réelles de densité $\{q(\theta, x) : \theta \in \Theta := [0, 1]\}$, où

$$q(\theta, x) = \theta f_1(x) + (1 - \theta) f_2(x).$$

La fonction de vraisemblance de

$$L_n(\theta, X_1, \dots, X_n) = \prod_{i=1}^n (\theta f_1(X_i) + (1 - \theta) f_2(X_i)),$$

de sorte que pour tout $\theta \in]0, 1[$

$$\ell_n(\theta, X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, X_i) \quad (\text{II-3.11})$$

où $\ell(\theta, x)$ est le score

$$\ell(\theta, x) = \frac{f_1(x) - f_2(x)}{\theta f_1(x) + (1 - \theta) f_2(x)} = \frac{f_1(x) - f_2(x)}{\theta \{f_1(x) - f_2(x)\} + f_2(x)}. \quad (\text{II-3.12})$$

Le score $\theta \mapsto \ell(\theta, x)$ est une fonction décroissante sur $[0, 1]$. Nous allons montrer que pour tout $\theta_0 \in [0, 1]$ nous avons

$$\int \ell(0, x) \mathbb{Q}_{\theta_0}(dx) > 0,$$

(cette quantité étant éventuellement infinie). Pour que cette quantité soit bien définie, nous devons tout d'abord montrer que $\int \{\ell(0, x)\}^- \mathbb{Q}_{\theta_0}(dx) < \infty$. Pour tout $\theta_0 \in]0, 1[$,

$$\begin{aligned} \int \frac{\{f_1(x) - f_2(x)\}^-}{f_2(x)} \mathbb{Q}_{\theta_0}(dx) &= \int \frac{\{f_1(x) - f_2(x)\}^-}{f_2(x)} \{\theta_0 f_1(x) + (1 - \theta_0) f_2(x)\} dx \\ &\leq \theta_0 \int \{f_1(x) - f_2(x)\}^- dx + (1 - \theta_0) \int \{f_1(x) - f_2(x)\}^- dx. \end{aligned}$$

Comme $\int \{f_1(x) - f_2(x)\} dx = 0$, nous avons

$$\int \{f_1(x) - f_2(x)\}^+ dx = \int \{f_1(x) - f_2(x)\}^- dx,$$

ce qui implique que $\int \{f_1(x) - f_2(x)\}^- dx = (1/2) \int |f_1(x) - f_2(x)| dx$. Nous obtenons donc

$$\int \frac{\{f_1(x) - f_2(x)\}^-}{f_2(x)} \mathbb{Q}_{\theta_0}(dx) \leq \frac{1}{2} \int |f_1(x) - f_2(x)| dx.$$

Un calcul élémentaire montre que

$$\begin{aligned} \int \ell(0, X_1) \mathbb{Q}_{\theta_0}(dx) &= \int \frac{f_1(x) - f_2(x)}{f_2(x)} \{\theta_0 f_1(x) + (1 - \theta_0) f_2(x)\} dx \\ &= \theta_0 \int \frac{\{f_1(x) - f_2(x)\}^2}{f_2(x)} dx, \end{aligned}$$

où nous avons utilisé que $\int \{f_1(x) - f_2(x)\} dx = 0$. De façon tout à fait similaire, nous pouvons montrer que $\int \{\ell(1, x)\}^+ \mathbb{Q}_{\theta_0}(dx) < \infty$ et que

$$\int \ell(1, x) \mathbb{Q}_{\theta_0}(dx) = -(1 - \theta_0) \int \frac{\{f_1(x) - f_2(x)\}^2}{f_1(x)} dx < 0.$$

En appliquant la loi des grands nombres (Théorème IV-2.19), nous avons donc, pour tout $\theta_0 \in]0, 1[$ que

$$\begin{aligned} n^{-1} \sum_{k=1}^n \ell(0, X_k) &\xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \int \ell(0, x) \mathbb{Q}_{\theta_0}(dx) > 0 \\ n^{-1} \sum_{k=1}^n \ell(1, X_k) &\xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \int \ell(1, x) \mathbb{Q}_{\theta_0}(dx) < 0 \end{aligned}$$

Notons que $\int \ell(0, x) \mathbb{Q}_{\theta_0}(dx)$ peut être égal à $+\infty$ et $\int \ell(1, x) \mathbb{Q}_{\theta_0}(dx)$ peut être égal à $-\infty$. Par conséquent, pour tout $\theta_0 \in \Theta$, l'équation de vraisemblance (II-3.11) admet une solution unique, qui correspond à un maximum, notée $\hat{\theta}_n^{\text{MV}}$. On vérifie aisément que les conditions de régularité (Définition I-4.20) sont satisfaites. La quantité d'information de Fisher est donnée, pour tout $\theta \in \Theta$, par

$$\mathbb{I}(\theta) = \frac{1}{\theta(1-\theta)} \left[1 - \int \frac{f_1(x)f_2(x)}{\theta f_1(x) + (1-\theta)f_2(x)} dx \right] \quad \diamond$$

On remarque que l'information de Fisher est maximale quand $f_1(x)f_2(x) = 0$ 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 $\int_{\mathbb{R}} (F_1(x) - F_2(x))^2 dx < +\infty$, 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

$$\theta \mapsto \int_{\mathbb{R}} (\hat{F}_n(x) - F_\theta(x))^2 dx,$$

où

$$F_\theta(x) := \theta F_1(x) + (1-\theta)F_2(x), \quad \hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x}.$$

En dérivant par rapport à la variable θ , on obtient

$$\int_{\mathbb{R}} (\hat{F}_n(x) - F_\theta(x)) (F_1(x) - F_2(x)) dx = 0,$$

d'où

$$\hat{\theta}_n = \frac{\int_{\mathbb{R}} (\hat{F}_n(x) - F_2(x)) (F_1(x) - F_2(x)) dx}{\int_{\mathbb{R}} (F_1(x) - F_2(x))^2 dx}.$$

En utilisant le T.L.C. (Théorème IV-2.39), on peut montrer que $\hat{\theta}_n$ est asymptotiquement normal. Alors l'estimateur modifié

$$\tilde{\theta}_n = \hat{\theta}_n - \frac{\ell'_n(\hat{\theta}_n, X_1, \dots, X_n)}{\ddot{\ell}_n(\hat{\theta}_n, X_1, \dots, X_n)}$$

où

$$\ell'_n(\theta, X_1, \dots, X_n) := -\frac{1}{n} \sum_{i=1}^n \frac{(f_1(X_i) - f_2(X_i))^2}{(\theta f_1(X_i) + (1-\theta)f_2(X_i))^2}$$

est asymptotiquement efficace, et sa variance asymptotique est l'information de Fisher du modèle

$$\mathbb{I}(\theta) = \int_{\mathbb{R}} \frac{(f_1(x) - f_2(x))^2}{\theta f_1(x) + (1-\theta)f_2(x)} dx.$$

Chapitre II-4

Tests asymptotiques

II-4.1 Introduction

Dans tout ce chapitre, n désigne un entier non nul, α un réel de l'ouvert $]0, 1[$, $(X, \mathcal{X}, \{\mathbb{Q}_\theta, \theta \in \Theta\})$ un modèle statistique où $\Theta \subset \mathbb{R}^d$. On notera $Z = X^{\mathbb{N}}$ l'espace des suites à valeurs dans X et $Z_n = X^n$ sa projection canonique, $\mathcal{Z} = \mathcal{X}^{\otimes \mathbb{N}}$ la tribu produit et $\mathcal{Z}_n = \mathcal{X}^{\otimes n}$ sa projection canonique et on notera $\{\mathbb{P}_\theta, \theta \in \Theta\} = \{\mathbb{Q}_\theta^{\otimes \mathbb{N}}, \theta \in \Theta\}$ la famille des lois produits.

Chaque fois qu'on supposera la famille $\{\mathbb{Q}_\theta, \theta \in \Theta\}$ dominée, μ sera la mesure dominante, $q(\theta, \cdot)$ la densité de $\mathbb{Q}_\theta$ par rapport à la mesure μ , $\ell(\theta, \cdot) = \log q(\theta, \cdot)$, $\ell_n(\theta) = n^{-1} \sum_{i=1}^n \log q(\theta, X_i)$, $\mathbb{I}(\theta)$ sera l'information de Fisher au point $\theta \in \Theta$ la matrice de covariance du score $x \mapsto \nabla \ell(\theta, x)$ ou opposé de l'espérance de la Hessienne $H_\ell(\cdot)(\theta)$ (lorsque ces quantités sont bien définies).

On notera toujours $Z = \{X_i, i \in \mathbb{N}\}$ le processus canonique et $Z_n = \{X_i, i \in \{1, \dots, n\}\}$ sa projection canonique. Tous les résultats limites seront obtenus dans la limite $n \rightarrow \infty$.

Dans ce chapitre, on va montrer comment utiliser des méthodes asymptotiques pour calibrer des tests, en obtenant d'abord la loi limite de statistiques d'intérêt et en estimant les paramètres inconnus de cette loi. Nous verrons également l'intérêt pratique du point de vue asymptotique permettant d'approcher des fonctions difficilement calculables par des approximations stochastiques naturelles.

Nous verrons ensuite comment comparer théoriquement des tests en étudiant leur puissance sous des alternatives locales (Section ??).

La définition suivante formalise mathématiquement les concepts de base de la théorie asymptotique des tests.

Définition II-4.1. *Considérons le test*

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

Soit $\{\phi_n, n \in \mathbb{N}\}$ une suite de tests, i.e. pour tout $n \in \mathbb{N}^*$, $\phi_n : X^n \mapsto [0, 1]$. On dit que la suite de tests $\{\phi_n, n \in \mathbb{N}\}$ est de niveau asymptotique $\alpha \in [0, 1]$ lorsque

$$\text{pour tout } \theta \in \Theta_0, \quad \limsup_{n \rightarrow \infty} \mathbb{E}_\theta[\phi_n(X_1, \dots, X_n)] \leq \alpha.$$

On dit que la suite de tests $\{\phi_n, n \in \mathbb{N}\}$ est consistante lorsque

$$\text{pour tout } \theta \in \Theta_1, \quad \liminf_{n \rightarrow \infty} \mathbb{E}_\theta[\phi_n(X_1, \dots, X_n)] = 1.$$

On rappelle qu'un test est utilisé pour prendre une décision au vue des données. Cette décision est prise en tirant indépendamment des observations $(X_1, \dots, X_n)$ une loi de Bernoulli de paramètre $\phi_n(X_1, \dots, X_n)$.

L'ensemble $\{\phi_n(X_1, \dots, X_n) = 1\}$ est l'ensemble de rejet du test, celui sur lequel on rejette H_0 pour prendre la décision H_1 , l'ensemble $\{\phi_n(X_1, \dots, X_n) = 0\}$ est l'ensemble d'acceptation sur lequel on accepte H_0 . L'ensemble $\{0 < \phi_n(X_1, \dots, X_n) < 1\}$ est l'ensemble sur lequel la décision ne relève pas seulement de l'observation (on utilise une randomisation supplémentaire).

Rappelons la philosophie des tests qui consiste d'abord à garantir un niveau donné pour un test, ce qui amène à trancher les situations litigieuses en faveur de H_0 . La consistance est dans ce cadre une propriété minimale assurant un comportement raisonnable du test sous l'hypothèse alternative. Comme pour l'estimation, la consistance sera systématiquement vérifiée par les tests que nous discuterons et il faudra d'autres outils pour aborder la question délicate de l'optimalité d'un test.

Exemple II-4.2 (Modèle de translation et d'échelle). Soit F une fonction de répartition sur $\mathbb{R}$. Pour $\Theta = \mathbb{R} \times \mathbb{R}_+^*$, notons $\mathbb{Q}_\theta$ la loi de fonction de répartition

$$F_\theta(x) = F\left(\frac{x - \mu}{\sigma}\right), \quad \text{pour tout } x \in \mathbb{R}.$$

Soit $\{X_k, k \in \mathbb{N}\}$ un échantillon du modèle $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{Q}_\theta, \theta \in \Theta\})$. On s'intéresse au test

$$H_0 : \mu = 0, \quad \text{contre} \quad H_1 : \mu \neq 0.$$

Supposons tout d'abord que $F = \Phi$ où Φ est la fonction de répartition de la gaussienne centrée réduite $N(0, 1)$

$$\Phi(x) = \int_{-\infty}^x e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}, \quad \text{pour tout } x \in \mathbb{R}.$$

Pour tout $\theta = (\mu, \sigma) \in \Theta$, $\mathbb{Q}_\theta$ est la loi $N(\mu, \sigma^2)$. Le t -test de l'Exemple I-3.24 est défini pour tout $n \in \mathbb{N}^*$ comme le test pur égal à 1 si et seulement si

$$\sqrt{n} \frac{|\bar{X}_n|}{S_n} > t_{1-\alpha/2}(n-1), \quad \text{où} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Autrement dit, pour tout $n \in \mathbb{N}^*$, on considère le test

$$\phi_n(X_1, \dots, X_n) = \mathbb{1}_{\{\sqrt{n}|\bar{X}_n|/S_n > t_{1-\alpha/2}(n-1)\}}.$$

C'est un test de taille α , i.e. pour tout $n \in \mathbb{N}^*$ et tout $\theta \in \Theta$, $\mathbb{E}_\theta[\phi_n(X_1, \dots, X_n)] = \alpha$. Il est donc bien entendu de niveau asymptotique α . D'autre part, pour tout $\theta = (\mu, \sigma)$ tel que $\mu \neq 0$,

$$\mathbb{E}_\theta[\phi_n(X_1, \dots, X_n)] = \mathbb{P}_\theta\left(\sqrt{n} \frac{|\bar{X}_n|}{S_n} > t_{1-\alpha/2}(n-1)\right) \geq \mathbb{P}_\theta\left(\sqrt{n} \frac{|\mu|}{S_n} - \sqrt{n} \frac{|\bar{X}_n - \mu|}{S_n} > t_{1-\alpha/2}(n-1)\right).$$

Comme $\sqrt{n}(\bar{X}_n - \mu)/S_n$ converge en loi vers la gaussienne $N(0, 1)$, que $t_{1-\alpha/2}(n-1)$ converge vers le quantile d'ordre $1 - \alpha/2$ de la loi $N(0, 1)$ et que S_n converge en probabilité vers σ , cette dernière probabilité tend vers 1, donc le test est consistant.

Supposons maintenant que F est une fonction de répartition sur $\mathbb{R}$ telle que $\int x^2 F(dx) < \infty$ et $\int x F(dx) = 0$. Le dernier raisonnement s'applique toujours, donc le t -test est toujours consistant dans ce cadre général. De plus, si $\theta = (0, \sigma) \in \Theta_0$, on a toujours $t_{1-\alpha/2}(n-1) \rightarrow z_{1-\alpha/2}$ et $\sqrt{n}\bar{X}_n/S_n$ converge en loi vers $N(0, 1)$, donc

$$\mathbb{E}_\theta[\phi_n(X_1, \dots, X_n)] = \mathbb{P}_\theta\left(\sqrt{n} \frac{|\bar{X}_n|}{S_n} > t_{1-\alpha/2}(n-1)\right) \rightarrow \alpha.$$

Le t -test est donc toujours de niveau asymptotique α . ◇

Dans le modèle (non paramétrique) de translation et échelle que nous venons de discuter, il a suffi d'utiliser la construction exacte dans le modèle Gaussien pour obtenir un test asymptotiquement bien calibré. Alors que la construction d'un test dont le niveau est exactement contrôlé par α est impossible dans ce modèle général, la construction asymptotique a été élémentaire. C'est le premier avantage du point de vue asymptotique que nous allons développer plus avant dans la section suivante. Remarquons déjà qu'on aurait pu tout aussi bien remplacer $t_{1-\alpha/2}(n-1)$ par n'importe quelle suite convergeant vers $z_{1-\alpha/2}$ (y compris donc par $z_{1-\alpha/2}$) et S_n par n'importe quel estimateur consistant de σ sans changer le résultat.

II-4.2 Tests d'adéquation

Dans cette section, nous présentons deux tests d'adéquation. Soit $\theta_0 \in \Theta$, un test d'adéquation est un test pour les hypothèses.

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0.$$

Autrement dit, H_0 est réduit au singleton $\{\mathbb{Q}_{\theta_0}\}$ et $H_1 = \{\mathbb{Q}_{\theta}, \theta \neq \theta_0\}$ et on veut savoir si les données sont en adéquation avec la loi $\mathbb{Q}_{\theta_0}$.

II-4.2.1 Test du χ^2

Supposons tout d'abord que l'ensemble des valeurs possibles est fini $X = \{0, 1, \dots, M\}$. Chaque loi $\mathbb{Q}_{\theta}$ est déterminée par la donnée d'un vecteur $q_{\theta} = (\mathbb{Q}_{\theta}(\{1\}), \dots, \mathbb{Q}_{\theta}(\{M\}))^T$. On suppose que, pour tout $\theta \in \Theta$, $q_{\theta} \in \Delta = \{(u_1, \dots, u_M)^T \in (\mathbb{R}_+^*)^M : \sum_{i=1}^M u_i < 1\}$. Chaque coordonnée $q_{\theta,i} = \mathbb{Q}_{\theta}(\{i\})$ est estimée par la fréquence d'apparition de la classe i ,

$$\hat{q}_{n,i} = \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{X_j=i\}}, \quad \text{pour tout } i \in \{1, \dots, M\}.$$

Notons $T_n(\theta)$ le vecteur de $\mathbb{R}^M$ de coordonnées

$$\frac{\hat{q}_{n,i} - q_{\theta,i}}{\sqrt{q_{\theta,i}}}, \quad i \in \{1, \dots, M\}.$$

Le résultat suivant donne le pivot asymptotique du test d'adéquation du χ^2 .

Proposition II-4.3. *Supposons que $q_{\theta} \in \Delta$ pour tout $\theta \in \Theta$. Alors la suite de fonctions $\{n \|T_n(\cdot)\|^2, n \in \mathbb{N}\}$ est asymptotiquement pivotale,*

$$n \|T_n(\theta)\|^2 \xrightarrow{\mathbb{P}_{\theta}} \chi^2(M-1), \quad \text{pour tout } \theta \in \Theta.$$

Démonstration. Pour tout $\theta \in \Theta$, $\sqrt{n}T_n(\theta)$ se réécrit $\sqrt{n}T_n(\theta) = n^{-1/2} \sum_{j=1}^n Y_j(\theta)$, où $Y_j(\theta) \in \mathbb{R}^M$ est le vecteur $M \times 1$ dont les coordonnées sont définies par

$$Y_j^{(i)}(\theta) = \frac{\mathbb{1}_{\{X_j=i\}} - q_{\theta,i}}{\sqrt{q_{\theta,i}}}, \quad i \in \{1, \dots, M\}.$$

Les vecteurs $\{Y_j(\theta), j \in \mathbb{N}\}$ sont, sous la probabilité $\mathbb{P}_{\theta}$, indépendants et de même loi. Elles vérifient $\mathbb{E}_{\theta}[Y_j(\theta)] = 0$ et, en notant $\sqrt{q_{\theta}} = (\sqrt{\mathbb{Q}_{\theta}(\{1\})}, \dots, \sqrt{\mathbb{Q}_{\theta}(\{M\})})^T$, on vérifie que

$$\Sigma(\theta) := \mathbb{E}_{\theta}[Y_j(\theta)Y_j^T(\theta)] = I_M - \sqrt{q_{\theta}}\sqrt{q_{\theta}}^T.$$

Par le Théorème de la Limite Centrale (Théorème IV-2.39), on a donc

$$\sqrt{n}T_n(\theta) \xrightarrow{\mathbb{P}_{\theta}} N(0, \Sigma(\theta)), \quad \text{pour tout } \theta \in \Theta.$$

Par le théorème IV-2.30, on a donc $n \|T_n(\theta)\|^2 \xrightarrow{\mathbb{P}_{\theta}} \|N(0, \Sigma(\theta))\|^2$. La covariance limite $\Sigma(\theta)$ est la matrice de la projection orthogonale sur l'orthogonal de la droite engendrée par $\sqrt{p_{\theta}}$, donc sur un espace de dimension $M-1$. Le théorème de Cochran (Proposition IV-5.20) permet de conclure la preuve. $\square$

Le test d'adéquation du χ^2 consiste à rejeter H_0 lorsque $n \|T_n(\theta_0)\|^2 > \chi_{1-\alpha}^2(M-1)$, c'est à dire que le test du χ^2 d'adéquation est défini, pour tout $n \geq 1$, par :

$$\phi_n(X_1, \dots, X_n) = \mathbb{1}_{\{n \|T_n(\theta_0)\|^2 > \chi_{1-\alpha}^2(M-1)\}}. \quad (\text{II-4.1})$$

Proposition II-4.4. *Le test d'adéquation du χ^2 de*

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0 .$$

défini par (II-4.1) est un test consistant de niveau asymptotique α .

Démonstration. La proposition II-4.3 montre que

$$\mathbb{E}_{\theta_0}[\phi_n(X_1, \dots, X_n)] \rightarrow \alpha .$$

Le test du χ^2 est donc de niveau asymptotique α . D'autre part, par la loi des grands nombres, pour tout θ ,

$$\|T_n(\theta_0)\|^2 \xrightarrow{\mathbb{P}_{\theta_0-\text{prob}}} \sum_{i=1}^M \frac{(q_{\theta,i} - q_{\theta_0,i})^2}{q_{\theta_0,i}} .$$

Donc, si $q_\theta \neq q_{\theta_0}$, $n \|T_n(\theta_0)\|^2 \xrightarrow{\mathbb{P}_{\theta_0-\text{prob}}} +\infty$, ce qui démontre la consistance du test. $\square$

Supposons maintenant que $(X, \mathcal{X})$ est un espace mesurable quelconque (plus nécessairement fini), et soit $\{\mathbb{Q}_\theta, \theta \in \Theta\}$ une famille de lois de probabilité sur $(X, \mathcal{X})$. On souhaite encore faire le test d'adéquation

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0 .$$

Pour cela, on considère $A_0, \dots, A_M$ une partition de X en éléments de $\mathcal{X}$. On suppose que pour tout $i \in \{1, \dots, M\}$ et $\theta \in \Theta$,

$$\mathbb{Q}_\theta(A_i) > 0 .$$

Le vecteur $q_\theta = (\mathbb{Q}_\theta(A_1), \dots, \mathbb{Q}_\theta(A_M))^T$ appartient donc à l'espace Δ , chaque coordonnée $q_{\theta,i} = \mathbb{Q}_\theta(A_i)$ peut être estimée par

$$\hat{q}_{n,i} = n^{-1} \sum_{j=1}^n \mathbb{1}_{A_i}(X_j)$$

et on peut encore construire $T_n(\theta)$ le vecteur de coordonnées

$$\frac{\hat{q}_{n,i} - q_{\theta,i}}{\sqrt{q_{\theta,i}}}, \quad i \in \{1, \dots, M\} .$$

On vérifie immédiatement que la proposition II-4.3 s'applique encore, de sorte que le test d'adéquation du χ^2 dans ce cadre général, défini par (II-4.1) est encore de niveau asymptotique α . En revanche, ce test n'est consistant que si le vecteur q_θ permet d'identifier la loi $\mathbb{Q}_\theta$ car d'après la proposition II-4.3, la puissance du test sera toujours majorée par α si θ est tel que $q_\theta = q_{\theta_0}$.

Le test du χ^2 d'adéquation peut donc s'étendre de manière naturelle à des espaces généraux, mais la puissance de cette version générale dépend beaucoup du choix de la partition $A_0, \dots, A_M$. Si celle-ci est mal choisie, par exemple, si elle n'est pas assez riche, le test peut même ne pas être consistant. On peut être tenté de prendre une partition très fine (M grand). Il faut toutefois être prudent car le contrôle du niveau et de la puissance du test ne sont valides qu'asymptotiquement. L'approximation asymptotique peut devenir médiocre si le nombre d'éléments de la partition M croît trop rapidement par rapport à la taille de l'échantillon n .

Le choix d'un nombre approprié d'éléments de la partition M , et plus généralement d'une bonne partition, est une question délicate qui dépasse le cadre de ce cours. En revanche, lorsque $X \subset \mathbb{R}$ il existe une alternative au test du χ^2 que nous allons maintenant présenter.

II-4.2.2 Test de Kolmogorov-Smirnov

On suppose dans cette section que $X \subset \mathbb{R}$ est muni de la tribu des Boréliens. Pour tout $\theta \in \Theta$, on note pour tout $t \in \mathbb{R}$,

$$F_\theta(t) = \mathbb{Q}_\theta([-\infty, t]), \quad \widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}}.$$

Le théorème de Glivenko-Cantelli (Théorème IV-4.5) montre que $\{\widehat{F}_n, n \in \mathbb{N}\}$ est une suite consistante d'estimateur de la fonction de répartition $\theta \mapsto F_\theta$, i.e. pour tout $\theta \in \Theta$ et $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(D_n(\theta) \geq \varepsilon) = 0 \quad \text{où} \quad D_n(\theta) = \sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F_\theta(t)|.$$

Le théorème de Kolmogorov, dont la preuve dépasse le cadre de ce cours, assure que la suite $\{\sqrt{n}D_n(\theta), n \in \mathbb{N}\}$ est asymptotiquement pivotale.

Théorème II-4.5 (Kolmogorov). Pour tout $\theta \in \Theta$, $D_n(\theta) \xrightarrow{\mathbb{P}_\theta} K$, où K est la loi de Kolmogorov, i.e. pour tout $\theta \in \Theta$ et $t \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(\sqrt{n}D_n(\theta) \leq t) = K(t)$$

où $K(t) = 1 + 2 \sum_{k=1}^{+\infty} (-1)^k e^{-k^2 t^2}$.

Une version non-asymptotique de ce test donnée par le Théorème IV-4.6.

Théorème II-4.6. Pour tout $\theta \in \Theta$, $n \geq 1$ et $t \in \mathbb{R}_+$,

$$\mathbb{P}_\theta(\sqrt{n}D_n(\theta) > t) \leq 2e^{-2t^2}$$

Ce résultat est remarquable car il fournit, pour tout $n \geq 1$, une borne qui coïncide avec l'asymptotique exacte au premier ordre.

Comme dans le cadre du test du χ^2 , nous allons maintenant utiliser la suite de pivots asymptotiques $\{\sqrt{n}D_n(\cdot), n \in \mathbb{N}\}$ pour construire le test d'adéquation de Kolmogorov-Smirnov. Ce test consiste à rejeter H_0 lorsque $\sqrt{n}D_n(\theta_0) > K^{-1}(1 - \alpha)$, c'est à dire que le test de Kolmogorov-Smirnov d'adéquation est défini, pour tout $n \geq 1$, par

$$\phi_n(X_1, \dots, X_n) = \mathbb{1}_{\{\sqrt{n}D_n(\theta_0) > K^{-1}(1 - \alpha)\}}. \quad (\text{II-4.2})$$

Proposition II-4.7. Le test d'adéquation de Kolmogorov-Smirnov de

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0.$$

défini par (II-4.2) est un test consistant de niveau asymptotique α .

Démonstration. Le théorème II-4.5 montre que

$$\mathbb{E}_{\theta_0}[\phi_n(X_1, \dots, X_n)] \rightarrow \alpha.$$

Le test de Kolmogorov-Smirnov est donc de niveau asymptotique α . D'autre part, par le théorème de Glivenko-Cantelli (Théorème IV-4.5), pour tout $\theta \in \Theta$,

$$D_n(\theta) \xrightarrow{\mathbb{P}_{\theta}-\text{prob}} \sup_{t \in \mathbb{R}} |F_{\theta}(t) - F_{\theta_0}(t)| .$$

Donc, si $\mathbb{Q}_{\theta} \neq \mathbb{Q}_{\theta_0}$, $\sqrt{n}D_n(\theta_0) \xrightarrow{\mathbb{P}_{\theta}-\text{prob}} +\infty$, ce qui démontre la consistance de la suite de tests. $\square$

Le test de Kolmogorov-Smirnov requiert, pour être mis en place en pratique, le calcul de $\sqrt{n}D_n(\theta_0)$ donc celui de $\sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F_{\theta_0}(t)|$. Pour évaluer ce supremum, il suffit de remarquer que, comme F_{θ_0} est croissante et $\widehat{F}_n$ est constante sur les intervalles $[X_{i:n}, X_{i+1:n}[$, $i = 0, \dots, n-1$, avec $X_{0:n} = -\infty$, où $X_{1:n} < \dots < X_{n:n}$ sont les statistiques d'ordre (voir Section IV-4.3). Nous avons donc

$$\sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F_{\theta_0}(t)| = \max_{i=1, \dots, n} \left[|\max(\widehat{F}_n(X_{i:n}-) - F_{\theta_0}(X_{i:n}-), |\widehat{F}_n(X_{i:n}) - F_{\theta_0}(X_{i:n})|) \right] .$$

où pour $t \in \mathbb{R}$,

$$\widehat{F}_n(t-) = n^{-1} \sum_{i=1}^n \mathbb{1}_{]-\infty, t[}(X_i) .$$

II-4.3 Les tests de Wald, de Rao et du rapport de vraisemblance

Le point de vue asymptotique est efficace dans le cadre paramétrique de ce cours pour calibrer différents tests et régions de confiance fondés sur l'objet central de ce cours, la vraisemblance. Cette section donne quelques exemples d'application de ce principe général. On considère un modèle statistique $(X, \mathcal{X}, \{q_{\theta}, \theta \in \Theta\})$ régulier (voir définition I-4.20). On note $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ une suite d'estimateurs maximisant pour tout n (de façon éventuellement) approchée la vraisemblance construite sur $(X_1, \dots, X_n)$. Nous supposons dans toute cette section que $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ est une suite consistante d'estimateurs vérifiant, pour tout $\theta_0 \in \Theta$,

$$\sqrt{n} \nabla \ell_n(\hat{\theta}_n^{\text{MV}}) \xrightarrow{\mathbb{P}_{\theta_0}-\text{prob}} 0, \quad \text{où } \ell_n(\theta) := n^{-1} \sum_{i=1}^n \ell(\theta, X_i) . \quad (\text{II-4.3})$$

où $\ell(\theta, x) = \log q(\theta, x)$ pour tout $\theta \in \Theta, x \in X$.

II-4.3.1 Remarques préliminaires

Proposition II-4.8. *Supposons le modèle statistique $(X, \mathcal{X}, \{q_{\theta}, \theta \in \Theta\})$ soit régulier (Définition I-4.20). Soit $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ suite consistante d'estimateurs satisfaisant (II-4.3). Pour tout $\theta_0 \in \Theta$, nous avons*

$$\sqrt{n} \mathbb{I}^{1/2}(\theta)(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathbb{P}_{\theta}} N(0, \mathbb{I}_d) , \quad (\text{II-4.4})$$

où $\mathbb{I}^{1/2}(\theta)$ est une racine carrée de la matrice $\mathbb{I}(\theta)$ (c'est à dire $\mathbb{I}^{1/2}(\theta) \mathbb{I}^{1/2}(\theta) = \mathbb{I}(\theta)$). En particulier, pour tout $\theta \in \Theta$,

$$n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathbb{P}_{\theta}} \chi^2(d) . \quad (\text{II-4.5})$$

Démonstration. En appliquant le théorème II-3.3, on a

$$\sqrt{n}(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathbb{P}_{\theta}} N(0, \mathbb{I}^{-1}(\theta)) .$$

La Proposition II-4.8 assure que la suite de fonctions

$$\theta := n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\text{MV}} - \theta)$$

est asymptotiquement pivotale (voir la définition II-1.15). On peut utiliser ce pivot pour construire un test d'adéquation en posant

$$\phi_n(X_1, \dots, X_n) = \mathbb{I}_{\{n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\text{MV}} - \theta) > \chi_{1-\alpha}^2(d)\}} \quad . \quad (\text{II-4.6})$$

La proposition II-4.8 assure que ce test est de niveau asymptotique α . De plus, comme la suite d'estimateurs $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ est consistante, nous avons, pour tout $\theta \neq \theta_0$,

$$n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} +\infty ,$$

ce qui garantit la consistance du test (II-4.6) pour les hypothèses

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0 .$$

On peut aussi appliquer ce pivot pour construire des régions de confiance pour θ . Considérons l'ensemble (aléatoire)

$$S_{n,\alpha} := \{\theta \in \Theta : n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\text{MV}} - \theta) \leq \chi_{1-\alpha}^2(d)\} . \quad (\text{II-4.7})$$

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

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta}(\theta \in S_{n,\alpha}) = 1 - \alpha .$$

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 que ces constructions soient parfaitement légitimes, elles ne sont guère utilisées. L'information de Fisher est en effet souvent difficile à calculer de façon analytique même lorsqu'on dispose d'une forme explicite, c'est en général une fonction compliquée du paramètre rendant la détermination de $S_{n,\alpha}$ ou de la région de rejet de $\phi_n, S_{n,\alpha}^c$, délicate.

L'un des autres avantages du point de vue asymptotique, que nous allons maintenant développer, est de pouvoir contourner cette difficulté en remplaçant l'information de Fisher au point θ par une approximation stochastique élémentaire à calculer. 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 $\theta \mapsto \mathbb{I}(\theta)$ soit continue. Si la suite d'estimateurs $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ est consistante, ceci implique que $\mathbb{I}(\hat{\theta}_n^{\text{MV}}) \xrightarrow{\mathbb{P}_{\theta} - \text{prob}} \mathbb{I}(\theta)$. Par conséquent, en appliquant le théorème II-3.3 et le lemme IV-2.33, nous obtenons, pour tout $\theta \in \Theta$,

$$\sqrt{n} \mathbb{I}^{1/2}(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathbb{P}_{\theta}} \mathcal{N}(0, \mathbf{I}_d) ,$$

ce qui implique que la suite de fonctions

$$\theta \mapsto n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta)$$

est asymptotiquement pivotale. En particulier, la suite de tests

$$\phi_n(X_1, \dots, X_n) = \mathbb{I}_{\{n(\hat{\theta}_n^{\text{MV}} - \theta_0)^T \mathbb{I}(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta_0) \geq \chi_{1-\alpha}^2(d)\}}$$

est consistante de niveau asymptotique α pour le test d'hypothèses

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0 .$$

De plus $\tilde{S}_{n,\alpha}$ est un ellipsoïde de confiance pour θ , où

$$\tilde{S}_{n,\alpha} = \{\theta \in \Theta, n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta) \leq \chi_{1-\alpha}^2(d)\} . \quad (\text{II-4.8})$$

Pour évaluer les pivots $\theta \mapsto n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\theta)(\hat{\theta}_n^{\text{MV}} - \theta)$ et $\theta \mapsto n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta)$ il faut calculer la matrice d'information de Fisher. Celle-ci est parfois délicate à obtenir, mais on peut toujours en calculer une approximation. Pour tout $\theta \in \Theta^o$ posons

$$\mathbb{I}_n(\theta) := -n^{-1} \sum_{i=1}^n \mathbf{H}_{\ell}(\theta, X_i) . \quad (\text{II-4.9})$$

Proposition II-4.9. Supposons que le modèle $(X, \mathcal{X}, \{q_\theta, \theta \in \Theta\})$ soit régulier (Définition I-4.20).

(i) Pour toute suite $\{\delta_n, n \in \mathbb{N}\}$ positive telle que $\lim_{n \rightarrow \infty} \delta_n = 0$ et tout $\theta_0 \in \Theta^o$,

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\theta_0} \left[\sup_{\|t\| \leq \delta_n} \|\mathbb{I}_n(\theta_0 + t) - \mathbb{I}_n(\theta_0)\| \right] = 0. \quad (\text{II-4.10})$$

(ii) Si $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est une suite consistante d'estimateurs de θ , alors, pour tout $\theta_0 \in \Theta^o$,

$$\mathbb{I}_n(\hat{\theta}_n) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} \mathbb{I}(\theta_0).$$

Démonstration. (i) Pour tout $\delta > 0$ tel que $B(\theta_0, \delta) \subset \mathcal{V}(\theta_0)$,

$$\sup_{t \in B(\theta_0, \delta)} \|\mathbb{H}_\ell(\theta_0 + t, x) - \mathbb{H}_\ell(\theta_0, x)\| \leq 2g(x).$$

D'autre part, comme $\theta \mapsto \ell(\theta, x)$ est deux fois continûment différentiable au point θ_0 pour μ -presque tout $x \in X$, nous avons $\lim_{\delta \rightarrow 0} R_\delta(x) = 0$ où

$$R_\delta(x) := \sup_{\|t\| \leq \delta} \|\mathbb{H}_\ell(\theta_0 + t, x) - \mathbb{H}_\ell(\theta_0, x)\|.$$

Le théorème de convergence dominée montre que $\lim_{\delta \rightarrow 0} \mathbb{E}_{\theta_0}[R_\delta(X)] = 0$. Par conséquent, pour toute suite $\{\delta_n, n \in \mathbb{N}\}$ telle que $\lim_{n \rightarrow \infty} \delta_n = 0$, on a

$$\mathbb{E}_{\theta_0} \left[\sup_{\|t\| \leq \delta_n} \|\mathbb{I}_n(\theta_0 + t) - \mathbb{I}_n(\theta_0)\| \right] \leq \mathbb{E}_{\theta_0} \left[n^{-1} \sum_{i=1}^n R_{\delta_n}(X_i) \right] = \mathbb{E}_{\theta_0}[R_{\delta_n}(X_1)] \rightarrow_{n \rightarrow \infty} 0.$$

(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

$$\mathbb{I}_n(\theta_0) - \mathbb{I}(\theta_0) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0.$$

Comme la suite d'estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est consistante, il existe une suite $\{\delta_n, n \in \mathbb{N}\}$ telle que $\lim_{n \rightarrow \infty} \delta_n = 0$ et $\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\|\hat{\theta}_n - \theta_0\| \geq \delta_n) = 0$. Par conséquent, nous avons pour tout $\varepsilon > 0$, en utilisant (i),

$$\begin{aligned} & \mathbb{P}_{\theta_0}(\|\mathbb{I}_n(\hat{\theta}_n) - \mathbb{I}_n(\theta_0)\| \geq \varepsilon) \\ & \leq \mathbb{P}_{\theta_0}(\|\hat{\theta}_n - \theta_0\| \geq \delta_n) + \mathbb{P}_{\theta_0} \left(\sup_{\|t\| \leq \delta_n} \|\mathbb{I}_n(\theta_0 + t) - \mathbb{I}_n(\theta_0)\| \geq \varepsilon \right) \rightarrow_{n \rightarrow \infty} 0. \end{aligned}$$

On en déduit que $\mathbb{I}_n(\hat{\theta}_n) - \mathbb{I}(\theta_0) \xrightarrow{\mathbb{P}_{\theta_0} - \text{prob}} 0$. □

II-4.3.2 Tests de Wald

Soit $(X, \mathcal{X}, \{q_\theta, \theta \in \Theta\})$ un modèle statistique régulier. Soit $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ une suite consistante d'estimateurs satisfaisant (II-4.3). On appelle *Information de Fisher observée* la quantité

$$\mathbb{I}_n(\hat{\theta}_n^{\text{MV}}) := -\frac{1}{n} \sum_{i=1}^n \mathbb{H}_\ell(\hat{\theta}_n^{\text{MV}}, X_i). \quad (\text{II-4.11})$$

Proposition II-4.10 (Test d'adéquation de Wald). Supposons que le modèle statistique $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ soit régulier (Définition I-4.20). Soit $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ suite consistante d'estimateurs satisfaisant (II-4.3). Pour tout $\theta \in \Theta^o$ posons

$$G_n^{\text{Wald}}(\theta) = n(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}_n(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta)$$

On a pour tout $\theta \in \Theta^o$,

$$G_n^{\text{Wald}}(\theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(d).$$

La suite de tests définis par

$$\phi_n^{\text{Wald}}(X_1, \dots, X_n) = \mathbb{I}_{\{G_n^{\text{Wald}}(\theta_0) > \chi_{1-\alpha}^2(d)\}} \quad , \quad (\text{II-4.12})$$

est consistant de niveau asymptotique α pour le test d'hypothèses

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0$$

Posons

$$\check{S}_{n,\alpha} := \left\{ \theta \in \Theta : G_n^{\text{Wald}}(\theta) \leq \chi_{1-\alpha}^2(d) \right\} \quad , \quad (\text{II-4.13})$$

Pour tout $\theta \in \Theta$, $\lim_{n \rightarrow \infty} \mathbb{P}_\theta(\theta \in \check{S}_{n,\alpha}) = 1 - \alpha$, i.e. la suite $\{\check{S}_{n,\alpha} : n \in \mathbb{N}\}$ définit une suite de régions de confiance de couverture asymptotique $1 - \alpha$.

Démonstration. En combinant la Proposition II-4.8-(II-4.4) et la proposition II-4.9, nous obtenons donc que, pour tout $\theta_0 \in \Theta^o$,

$$\sqrt{n} \mathbb{I}_n^{1/2}(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, I_d) \quad . \quad (\text{II-4.14})$$

La preuve découle directement de ce résultat. $\square$

Le test (II-4.12) est appelé *test de Wald*. On peut généraliser le test de Wald pour pouvoir tester des hypothèses composites.

Nous dirons qu'une fonction $\mathbf{h} = (h_1, \dots, h_r)^T, \mathbb{R}^d \rightarrow \mathbb{R}^r, r \leq d$, définit une *contrainte régulière d'ordre r* si

- (i) $\mathbf{h}$ est continûment différentiable sur Θ^o ,
- (ii) pour tout $\theta \in \Theta^o$, la Jacobienne $\mathbf{J}_\mathbf{h}(\theta)$ soit de rang r .

Nous considérons le test d'hypothèses

$$H_0 : \mathbf{h}(\theta) = 0, \quad \text{contre} \quad H_1 : \mathbf{h}(\theta) \neq 0,$$

Théorème II-4.11 (Test de Wald). Supposons que le modèle statistique $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ soit régulier (Définition I-4.20). Soit $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ une suite consistante d'estimateurs satisfaisant (II-4.3).

Soit $\mathbf{h}$ une contrainte régulière d'ordre r . Supposons de plus que pour tout $\theta \in \Theta^o$ la matrice $V(\theta) = \mathbf{J}_g(\theta) \mathbb{I}(\theta)^{-1} \mathbf{J}_g(\theta)^T$ est inversible et que la fonction $\theta \mapsto V^{-1}(\theta)$ est continue.

Posons $\Theta_0 = \{\theta \in \Theta, g(\theta) = 0\}$. Alors le test de Wald défini par

$$\phi_n^{g, \text{Wald}}(X_1, \dots, X_n) = \mathbb{I}_{\{n\mathbf{h}(\hat{\theta}_n^{\text{MV}})^T V^{-1}(\hat{\theta}_n^{\text{MV}}) \mathbf{h}(\hat{\theta}_n^{\text{MV}}) > \chi_{1-\alpha}^2(r)\}} \quad ,$$

est une suite consistante de tests de niveau asymptotique α des hypothèses

$$H_0 : \mathbf{h}(\theta) = 0, \quad \text{contre} \quad H_1 : \mathbf{h}(\theta) \neq 0$$

Démonstration. Le théorème IV-2.47 (méthode Delta) assure que, pour tout $\theta \in \Theta^o$,

$$\sqrt{n}\{\mathbf{h}(\hat{\theta}_n^{\text{MV}}) - \mathbf{h}(\theta)\} \xrightarrow{\mathbb{P}_\theta} \mathbf{N}(0, V(\theta)) .$$

Il s'en suit alors que

$$n\{\mathbf{h}(\hat{\theta}_n^{\text{MV}}) - \mathbf{h}(\theta)\}^T V^{-1}(\theta) \{\mathbf{h}(\hat{\theta}_n^{\text{MV}}) - \mathbf{h}(\theta)\} \xrightarrow{\mathbb{P}_\theta} \chi^2(q), \quad \text{pour tout } \theta \in \Theta^o .$$

Comme V^{-1} est continue et $\hat{\theta}_n^{\text{MV}}$ consistant, $V^{-1}(\hat{\theta}_n^{\text{MV}}) \xrightarrow{\mathbb{P}_\theta - \text{prob}} V^{-1}(\theta)$ donc, par le lemme de Slutsky (Lemme IV-2.33),

$$G_n^{\text{g,Wald}}(\theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(q), \quad \text{pour tout } \theta \in \Theta^o .$$

où

$$G_n^{\text{g,Wald}}(\theta) = n\{\mathbf{h}(\hat{\theta}_n^{\text{MV}}) - \mathbf{h}(\theta)\}^T V^{-1}(\hat{\theta}_n^{\text{MV}}) \{\mathbf{h}(\hat{\theta}_n^{\text{MV}}) - \mathbf{h}(\theta)\} .$$

□

Bien entendu, on a aussi que $\left\{ \theta \in \Theta, G_n^{\text{g,Wald}}(\theta) \leq \chi_{1-\alpha}^2(q) \right\}$ est une région de confiance de niveau asymptotique $1 - \alpha$.

II-4.3.3 Tests du score de Rao

La mise en oeuvre des tests de Wald nécessitent l'évaluation d'un estimateur du maximum de vraisemblance (approché). On peut se passer de cette évaluation en considérant le test du score de Rao.

Proposition II-4.12 (Test d'adéquation du score de Rao). *Supposons que le modèle statistique $(\mathcal{X}, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ est régulier (Définition I-4.20). Posons*

$$G_n^{\text{Rao}}(\theta) = n \nabla \ell_n(\theta)^T \mathbb{I}_n^{-1}(\theta) \nabla \ell_n(\theta) .$$

Pour tout $\theta \in \Theta^o$, nous avons

$$G_n^{\text{Rao}}(\theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(d) .$$

Posons pour $\theta_0 \in \Theta$,

$$\phi_n^{\text{Rao}}(X_1, \dots, X_n) = \mathbb{I}_{\{G_n^{\text{Rao}}(\theta_0) > \chi_{1-\alpha}^2(d)\}}, \quad \text{pour tout } n \geq 1 ,$$

La suite de tests $\{\phi_n^{\text{Rao}}, n \in \mathbb{N}\}$ est consistante de niveau asymptotique α pour le test d'hypothèses

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0 .$$

Démonstration. Dans un modèle régulier, on a

$$\mathbb{E}_\theta[\nabla \ell(\theta, X)] = 0, \quad \mathbb{I}(\theta) = \mathbb{E}_\theta[\nabla \ell(\theta, X) \nabla \ell(\theta, X)^T] \quad \text{pour tout } \theta \in \Theta ,$$

donc

$$n^{1/2} \nabla \ell_n(\theta) \xrightarrow{\mathbb{P}_\theta} \mathbf{N}(0, \mathbb{I}(\theta)) .$$

Finalement, $G_n^{\text{Rao}}(\theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(d)$.

□

Dans le cadre d'hypothèses multiples, on a le résultat suivant.

Théorème II-4.13. Supposons que le modèle statistique $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ soit régulier (Définition I-4.20).

Soit $\mathbf{h}$ une contrainte régulière d'ordre r . On suppose qu'il existe une suite d'estimateurs

$$\bar{\theta}_n^{\text{MV}, \Theta_0} = \arg \max_{\theta \in \Theta_0} \ell_n(\theta) .$$

Alors, le test du score de Rao défini par

$$\phi_n^{\text{Rao}}(X_1, \dots, X_n) = \mathbb{1}_{\{n \nabla \ell_n(\bar{\theta}_n^{\text{MV}, \Theta_0})^T \mathbb{I}_n(\bar{\theta}_n^{\text{MV}, \Theta_0})^{-1} \nabla \ell_n(\bar{\theta}_n^{\text{MV}, \Theta_0}) > \chi_{1-\alpha}^2(r)\}} \quad n \geq 1 ,$$

est consistant et de niveau asymptotique α pour les hypothèses

$$H_0 : \mathbf{h}(\theta) = 0, \quad \text{contre} \quad H_1 : \mathbf{h}(\theta) \neq 0 .$$

Démonstration. La preuve de ce résultat repose sur la méthode Delta et le lemme de Slutsky en remarquant que $\bar{\theta}_n^{\text{MV}, \Theta_0}$ est un estimateur consistant de θ pour tout $\theta \in \Theta_0$. Le détail de la preuve est laissé en exercice. $\square$

Pour un test d'hypothèses multiples, le test de Rao nécessite donc le calcul de l'estimateur du maximum de vraisemblance $\bar{\theta}_n^{\text{MV}, \Theta_0}$ sur le modèle contraint $\{P_\theta, \theta \in \Theta_0\}$ alors que le test de Wald nécessitait le calcul de l'E.M.V. $\hat{\theta}_n^{\text{MV}}$ sur le modèle non-contraint.

II-4.3.4 Tests du rapport de vraisemblance

Comme les autres, ces tests reposent sur l'obtention d'un pivot asymptotique, le résultat de probabilité correspondant est connu sous le nom de théorème de Wilks.

Théorème II-4.14 (Test d'adéquation du rapport de vraisemblance). Supposons que le modèle statistique $(X, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ soit régulier (Définition I-4.20). Soit $\{\hat{\theta}_n^{\text{MV}}, n \in \mathbb{N}\}$ suite consistante d'estimateurs satisfaisant (II-4.3). Posons

$$G_n^{\text{RV}}(\theta) = 2n \ell_n(\hat{\theta}_n^{\text{MV}}) - 2n \ell_n(\theta) .$$

La suite de fonctions $\theta \mapsto G_n^{\text{RV}}(\theta)$ est asymptotiquement pivotale, i.e. pour tout $\theta \in \Theta^o$,

$$G_n^{\text{RV}}(\theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(d)$$

La suite de tests définie par

$$\phi_n^{\text{RV}}(X_1, \dots, X_n) = \mathbb{1}_{\{G_n^{\text{RV}}(\theta_0) \geq \chi_{1-\alpha}^2(d)\}} ,$$

est consistante et de niveau asymptotique α pour le test

$$H_0 : \theta = \theta_0, \quad \text{contre} \quad H_1 : \theta \neq \theta_0 .$$

La région

$$\check{S}_{n, \alpha} := \{\theta \in \Theta : G_n^{\text{RV}}(\theta) < \chi_{1-\alpha}^2(d)\} \quad (\text{II-4.15})$$

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

Démonstration. Comme le modèle est régulier, $\nabla \ell_n(\hat{\theta}_n^{\text{MV}}) = 0$, donc, en utilisant la formule de Taylor-Lagrange pour la log-vraisemblance en $\hat{\theta}_n^{\text{MV}}$, il existe un point θ_n^* du segment $[\theta, \hat{\theta}_n^{\text{MV}}]$ tel que

$$\ell_n(\theta) = \ell_n(\hat{\theta}_n^{\text{MV}}) - \frac{1}{2}(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}_n(\theta_n^*)(\hat{\theta}_n^{\text{MV}} - \theta). \quad (\text{II-4.16})$$

Pour tout $\theta \in \Theta^o$, on a $(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}_n(\theta)(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(d)$ et $\mathbb{I}_n(\theta_n^*) - \mathbb{I}_n(\hat{\theta}_n^{\text{MV}}) \xrightarrow{\mathbb{P}_{\theta-\text{prob}}} \mathbf{0}_d$, donc

$$2n\ell_n(\hat{\theta}_n^{\text{MV}}) - 2n\ell_n(\theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(d). \quad (\text{II-4.17})$$

□

Le test du rapport de vraisemblance s'étend comme les tests de Wald et de Rao à des hypothèses composites.

Théorème II-4.15 (Théorème de Wilks). *Supposons que le modèle statistique $(\mathcal{X}, \mathcal{X}, \{q_\theta \cdot \mu, \theta \in \Theta\})$ soit régulier (Définition I-4.20).*

Soit $\mathbf{h}$ une contrainte régulière d'ordre r . On note $\hat{\theta}_n^{\text{MV}} = \arg \max_{\theta \in \Theta} \ell_n(\theta)$ et $\bar{\theta}_n^{\text{MV}, \Theta_0} = \arg \max_{\theta \in \Theta_0} \ell_n(\theta)$. Alors la suite de tests définie par

$$\phi_n^{\text{Wilks}}(X_1, \dots, X_n) = \mathbb{1}_{\{2n\{\ell_n(\hat{\theta}_n^{\text{MV}}) - \ell_n(\bar{\theta}_n^{\text{MV}, \Theta_0})\} > \chi_{1-\alpha}^2(r)\}} \quad n \geq 1,$$

est consistante et de niveau asymptotique α pour les hypothèses

$$H_0 : \mathbf{h}(\theta) = 0, \quad \text{contre} \quad H_1 : \mathbf{h}(\theta) \neq 0.$$

La dénomination "test du rapport de vraisemblance" vient du fait que $2n\{\ell_n(\hat{\theta}_n^{\text{MV}}) - \ell_n(\bar{\theta}_n^{\text{MV}, \Theta_0})\} = -2n \log \Lambda_n$, où

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} L_n(\theta)}{\sup_{\theta \in \Theta} L_n(\theta)}.$$

Λ_n est le rapport de la vraisemblance maximale dans le modèle contraint et de la vraisemblance maximale dans le modèle non-contraint. Le test revient donc à rejeter l'hypothèse nulle lorsque ce rapport de vraisemblance est inférieur un certain seuil.

Exemple II-4.16 (Loi de Cauchy). Soit $\{X_k, k \in \mathbb{N}\}$ une suite i.i.d. de variables aléatoires de loi de Cauchy de paramètre de position $\theta \in \mathbb{R}$ et de paramètre d'échelle 1, i.e. $f(\theta, x) = \pi^{-1}\{1 + (x - \theta)^2\}^{-1}$. Le score est donné par

$$\dot{\ell}(\theta, x) = \frac{2(x - \theta)}{1 + (x - \theta)^2}.$$

Comme $\dot{\ell}(\theta, x) = O(x^{-1})$ quand $x \rightarrow \pm\infty$, la quantité $\mathbb{E}_\theta[(\dot{\ell}(\theta, x))^2] < \infty$. De plus, cette quantité ne dépend pas de $\theta \in \mathbb{R}$; l'information de Fisher est constante et est égale à 0.5. Dans ce cas, les régions de confiance associées à (II-4.7) et (II-4.8) coïncident.

Le score est deux fois différentiable et sa dérivée seconde est donnée par

$$\ddot{\ell}(\theta, x) = \frac{2(x - \theta)^2 - 2}{(1 + (x - \theta)^2)^2}. \quad \diamond$$

Cette fonction est continue et pour tout compact $K \subset \mathbb{R}$, $\sup_{\theta \in K} |\ddot{\ell}(\theta, x)| = O(x^{-2})$ quand $x \rightarrow \infty$. L'hypothèse I-4.20 est satisfaite. La Figure II-4.16 illustre la construction de la région de confiance $\check{S}_{n, \alpha}$ définie par (II-4.15).

Exemple II-4.17 (Loi de Poisson). Supposons que $\{X_k, k \in \mathbb{N}\}$ est une suite de variables aléatoires i.i.d. de loi de Poisson de paramètre $\theta \in \Theta = \mathbb{R}_+^*$. Nous avons, pour tout $x \in \mathbb{N}$,

$$\ell(\theta, x) = x \log(\theta) - \theta \log(x!) , \quad \dot{\ell}(\theta, x) = \frac{x}{\theta} - \log(x!), \quad \text{et} \quad \ddot{\ell}(\theta, x) = -\frac{x}{\theta^2}.$$

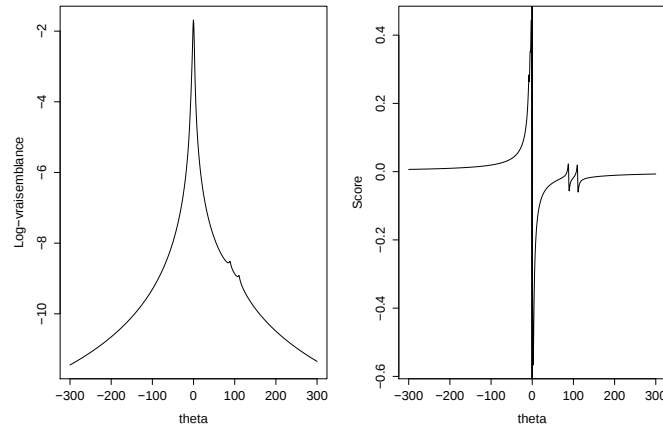


FIGURE II-4.1 – Figure de gauche : Log-vraisemblance de $n = 25$ observations simulées suivant une loi de Cauchy centrée. Figure de droite : dérivée par rapport au paramètre de la log-vraisemblance. On remarquera la présence de deux maxima locaux, bien séparés du maximum global, qui est au voisinage de 0.

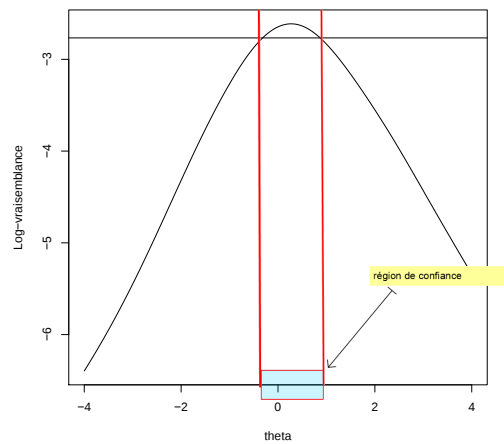


FIGURE II-4.2 – Construction de la région $\check{S}_{n,\alpha}$ pour $\alpha = 0.95$ définie par (II-4.15) pour l'estimation du paramètre de position d'une loi de Cauchy. La log-vraisemblance est obtenue à partir de $n = 25$ observations indépendantes de la loi de Cauchy centrée

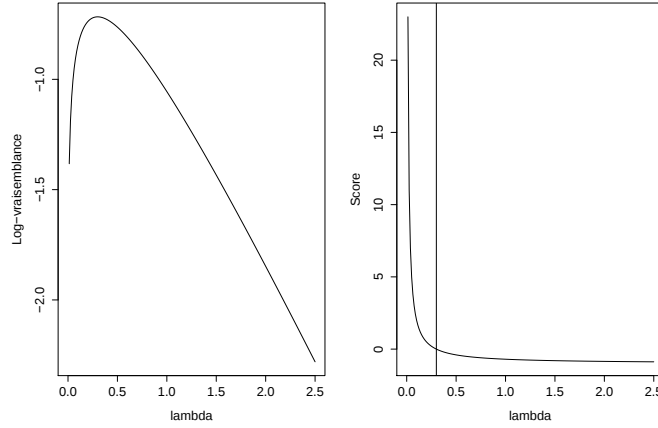


FIGURE II-4.3 – Log vraisemblance et score pour l'estimation du paramètre d'une loi de Poisson

La log-vraisemblance a donc pour expression

$$\theta \mapsto \ell_n(\theta) = n^{-1} \sum_{i=1}^n \ell(\theta, X_i) = \frac{\bar{X}_n}{\theta} - \theta - n^{-1} \log \left(\prod_{i=1}^n X_i! \right)$$

où $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$ est la moyenne empirique. Cette fonction est strictement concave et admet un maximum unique au point $\bar{X}_n$ et donc $\hat{\theta}_n^{\text{MV}} = \bar{X}_n$ (voir Figure II-4.3). L'information de Fisher est donnée ici par $\mathbb{I}(\theta_0) = 1/\theta_0$ et la région de confiance (II-4.7) par

$$\begin{aligned} S_{n,\alpha} &= \left\{ \theta > 0 : \sqrt{n/\theta} |\hat{\theta}_n^{\text{MV}} - \theta| < z_{1-\alpha/2} \right\} \\ &= \left\{ \theta > 0 : \theta^2 - 2\bar{X}_n\theta + \bar{X}_n^2 < z_{1-\alpha/2}^2 \theta/n \right\}. \end{aligned} \quad (\text{II-4.18})$$

Dans ce cas particulier, cette région de confiance est un intervalle $S_{n,\alpha} = [\hat{\theta}_n^-, \hat{\theta}_n^+]$, où

$$\hat{\theta}_n^\pm = \bar{X}_n + \frac{z_{1-\alpha/2}^2}{2n} \pm \sqrt{\left(\bar{X}_n + \frac{z_{1-\alpha/2}^2}{2n} \right)^2 - \bar{X}_n^2}.$$

Si nous substituons dans (II-4.18) l'information de Fisher $\mathbb{I}(\theta)$ par $\mathbb{I}(\hat{\theta}_n^{\text{MV}})$ (voir la construction (II-4.8)), nous obtenons l'intervalle de confiance donné par

$$\left[\bar{X}_n - z_{1-\alpha/2} \sqrt{\bar{X}_n/n}, \bar{X}_n + z_{1-\alpha/2} \sqrt{\bar{X}_n/n} \right]. \quad (\text{II-4.19})$$

Remarquons que rien ne garantit que la borne inférieure de cet intervalle de confiance soit positive.

Dans cet exemple, l'information de Fisher évaluée au point $\hat{\theta}_n^{\text{MV}}$ coïncide avec l'information de Fisher observée :

$$\mathbb{I}_n(\hat{\theta}_n^{\text{MV}}) = -n^{-1} \sum_{i=1}^n \ddot{\ell}(\hat{\theta}_n^{\text{MV}}, X_i) = n^{-1} \sum_{i=1}^n \frac{X_i}{\bar{X}_n^2} = \frac{\bar{X}_n}{\bar{X}_n^2} = \mathbb{I}(\bar{X}_n).$$

par conséquent l'intervalle de confiance (II-4.13), basé sur l'information de Fisher observé coïncide avec (II-4.19). Finalement, la région de confiance basée sur le développement local de la log-vraisemblance (II-4.15) est donnée par

$$\check{S}_{n,\alpha} = \left\{ \theta > 0 : \theta - \bar{X}_n \log(\theta/\bar{X}_n) - \bar{X}_n < \frac{z_{1-\alpha/2}^2}{2n} \right\}. \quad (\text{II-4.20})$$

◇

Cet ensemble définit un intervalle car la fonction $\theta \mapsto \theta - \bar{X}_n \log(\theta/\bar{X}_n) - \bar{X}_n$ est convexe. Toutefois, les extrémités de cet intervalle ne peuvent pas être calculées explicitement.

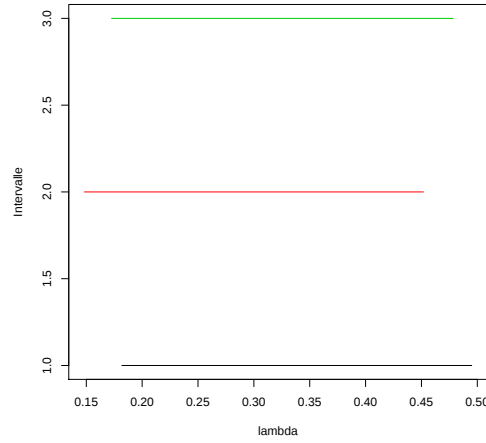


FIGURE II-4.4 – Intervalles de confiance pour le paramètre d’une loi de Poisson d’intensité 0.25. Haut : intervalle de confiance (II-4.18). Milieu : intervalle de confiance (II-4.19). Bas : Intervalle de confiance (II-4.20)

Exemple II-4.18 (Loi Gamma). Soit $\{X_k, k \in \mathbb{N}\}$ une suite i.i.d. de variables aléatoires distribuées suivant une loi Gamma de paramètres $\theta = (\alpha, \beta) \in \Theta = \mathbb{R}_+^* \times \mathbb{R}_+^*$ donnée par

$$f(\theta, x) = \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp(-x/\beta). \quad (\text{II-4.21})$$

La log-densité et le score sont donnés par

$$\begin{aligned} \ell(\theta, x) &= (\alpha - 1) \log(x) - \alpha \log(\beta) - \log \Gamma(\alpha) - x/\beta, \\ \nabla \ell(\theta, x) &= \begin{bmatrix} \log(x) - \log(\beta) - \psi(\alpha) \\ -\alpha/\beta + x/\beta^2 \end{bmatrix} \end{aligned}$$

où $\psi(\alpha) = \{\log(\Gamma(\alpha))\}'$ est la fonction digamma. La matrice d’information de Fisher est donnée par

$$\mathbb{I}(\alpha, \beta) = -\mathbb{E}_\theta[\mathbf{H}_\ell(\theta, X)] = \begin{bmatrix} \psi'(\alpha) & \beta^{-1} \\ \beta^{-1} & \alpha\beta^{-2} \end{bmatrix}.$$

Nous avons représenté dans la fig. II-4.5 la log-vraisemblance de l’observation (figure de gauche). Sur la figure de droite, nous avons aussi représenté la forme locale de la vraisemblance au voisinage de son maximum que nous avons superposé avec le développement quadratique.

$$\theta \mapsto \ell_n(\hat{\theta}_n^{\text{MV}}) - \frac{1}{2}(\hat{\theta}_n^{\text{MV}} - \theta)^T \mathbb{I}_n(\hat{\theta}_n^{\text{MV}})(\hat{\theta}_n^{\text{MV}} - \theta). \quad (\text{II-4.22})$$

◇

Nous avons visualisé sur la fig. II-4.6 l’ellipsoïde de confiance de probabilité asymptotique de couverture 0.95 (pour $n = 100$ observations) calculé en utilisant (II-4.8) et en utilisant (II-4.13) : les deux régions de confiance sont quasiment confondues.

II-4.3.5 Extension aux M et Z -estimateurs

Les constructions de la section précédente peuvent être étendues aux cas des M - et Z -estimateurs généraux. On donne le résultat pour un Z -estimateur.

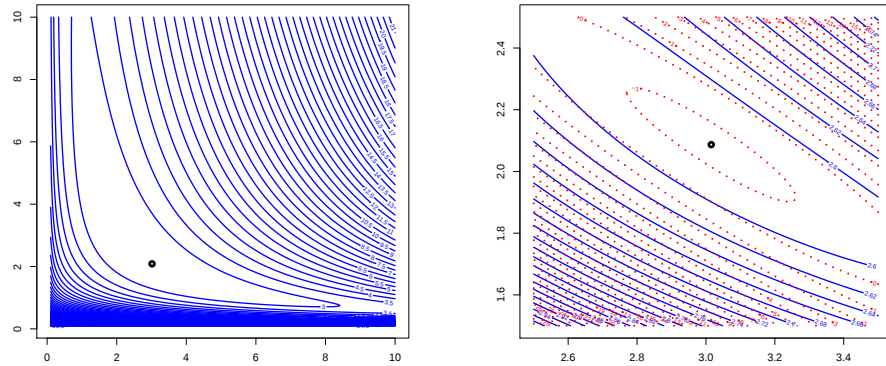


FIGURE II-4.5 – Figure de gauche : Lignes de niveau de la log-vraisemblance de $n = 100$ observations d'une loi Gamma de paramètre $(\alpha_0, \beta_0) = (3, 2)$ (le maximum est indiqué par un point). Figure de droite : Approximation quadratique locale de la log-vraisemblance

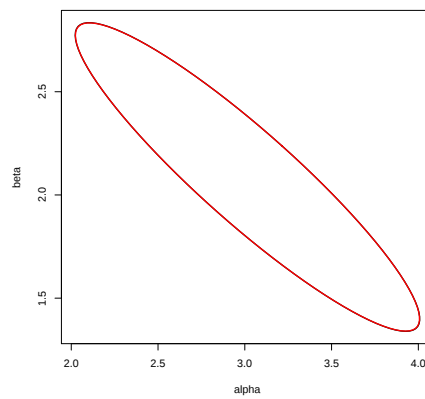


FIGURE II-4.6 – Ellipsoïdes de confiance de probabilité de couverture asymptotique 0.95

Proposition II-4.19 (Z-Test d'adéquation). Supposons que le modèle $(P_\theta)_{\theta \in \Theta}$ vérifie les hypothèses II-2.17 et soit $T : \cup X^n \rightarrow \mathbb{R}^d$ tel que l'estimateur $\hat{\theta}_n = T(X_1, \dots, X_n)$ vérifie

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(\hat{\theta}_n, X_i) \xrightarrow{\mathbb{P}_{\theta-\text{prob}}} 0, \quad \text{pour tout } \theta \in \Theta.$$

Soit $v_{n,\psi}$ une suite de matrices inversibles $v_{n,\psi}$ telles que

$$v_{n,\psi} \xrightarrow{\mathbb{P}_{\theta-\text{prob}}} v_\psi(\theta), \quad \text{pour tout } \theta \in \Theta,$$

où

$$v_\psi(\theta) = [J_\psi(\theta)]^{-1} \mathbb{E}_\theta[\psi(\theta, X) \psi(\theta, X)^T] [J_\psi(\theta)]^{-T}.$$

Alors

$$G_{n,\psi}(\theta) = n(\hat{\theta}_n - \theta)^T v_{n,\psi}^{-1} (\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_\theta} \chi^2(d).$$

En particulier, le test d'adéquation

$$\phi_\psi(X_1, \dots, X_n) = \mathbb{1}_{\{G_{n,\psi}(\theta_0) > \chi^2_{1-\alpha}(d)\}}$$

est consistant et de niveau asymptotique α .

Démonstration. D'après le Théorème II-2.18,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathbb{P}_\theta} N(0, v_\psi(\theta)).$$

La proposition est alors une conséquence du lemme IV-2.33. □

On conclut par l'extension naturelle de ce test à des hypothèses multiples.

Théorème II-4.20. Soit $g = (g_1, \dots, g_q)^T$ une fonction $\mathbb{R}^d \rightarrow \mathbb{R}^q$ continûment différentiable sur Θ telle que $J_g(\theta)$ est de rang q . Soit $\Theta_0 = \{\theta \in \Theta, g(\theta) = 0\}$ et supposons que V^{-1} est continue, où $V(\theta) = J_g(\theta) v_\psi(\theta) J_g(\theta)^T$. Supposons que le modèle $(P_\theta)_{\theta \in \Theta}$ vérifie les hypothèses II-2.17 et soit $T : \cup X^n \rightarrow \mathbb{R}^d$ tel que l'estimateur $\hat{\theta}_n = T(X_1, \dots, X_n)$ vérifie

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(\hat{\theta}_n, X_i) \xrightarrow{\mathbb{P}_{\theta-\text{prob}}} 0, \quad \text{pour tout } \theta \in \Theta.$$

Alors le test des hypothèses

$$H_0 : \theta \in \Theta_0, \quad \text{contre} \quad H_1 : \theta \in \Theta \setminus \Theta_0,$$

défini par

$$\phi_\psi(X_1, \dots, X_n) = \mathbb{1}_{\{ng(\hat{\theta}_n)^T V^{-1}(\hat{\theta}_n) g(\hat{\theta}_n) > \chi^2_{1-\alpha}(q)\}},$$

est un test consistant de niveau asymptotique α .

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 $Y = \{0, 1\}$ ou $Y = \{-1, +1\}$ (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 supposons que l'espace des attributs est un sous-ensemble $X \subset \mathbb{R}^d$ (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 $(\Omega, \mathcal{F}, \mathbb{P})$ 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 $A \in \mathcal{B}(\mathbb{R}^d)$,

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

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 $A \in \mathcal{B}(\mathbb{R}^d)$ (en supposant pour éviter les trivialités que $p \in]0, 1[$),

$$\mu_i(A) = \mathbb{P}(X \in A, Y = i) / \mathbb{P}(Y = i), \quad \forall i \in \{0, 1\}.$$

Nous avons donc, pour tout $A \in \mathcal{B}(\mathbb{R}^d)$

$$\mu(A) = p\mu_0(A) + (1 - p)\mu_1(A).$$

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 $p_i = d\mu_i/d\mu$, $i \in \{0, 1\}$, la densité de μ_i par rapport à μ . Nous avons, pour tout $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\mu_i(A) = \int \mathbb{1}_A(x) p_i(x) \mu(dx), \quad i \in \{0, 1\}.$$

La loi jointe du couple (X, Y) admet donc une densité $h : \mathbb{R}^d \times \{0, 1\} \rightarrow \mathbb{R}_+$ 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 $x \in \{x \in \mathbb{R}^d : pp_0(x) + (1 - p)p_1(x) > 0\}$ 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 $x \in \mathbb{R}^d$,

$$\eta(x) = h_{Y|X}(1|x) = \frac{pp_1(x)}{(1 - p)p_0(x) + pp_1(x)} . \quad (\text{III-1.1})$$

Cette fonction est appelée *probabilité a posteriori*.

De façon générale, soit λ une mesure σ -finie sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Supposons que la loi du couple (X, Y) admet une densité $h_{X,Y}$ par rapport à $\mu \otimes c$, où c est la mesure de comptage sur $\{0, 1\}$. Pour toute fonction $f : \mathbb{R}^d \times \{0, 1\} \rightarrow \mathbb{R}_+$, nous avons par définition

$$\mathbb{E}[f(X, Y)] = \iint f(x, y)h_{X,Y}(x, y)\lambda(\mathrm{d}x)c(\mathrm{d}y) = \int f(x, 0)h_{X,Y}(x, 0)\lambda(\mathrm{d}x) + \int f(x, 1)h_{X,Y}(x, 1)\lambda(\mathrm{d}x) .$$

La loi marginale des attributs X admet une densité par rapport à mesure λ donnée par

$$h_X(x) = \int h_{X,Y}(x, y)c(\mathrm{d}y) = h_{X,Y}(x, 0) + h_{X,Y}(x, 1) .$$

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

$$\eta(x) = \frac{h_{X,Y}(x, 1)}{h_X(x)} . \quad (\text{III-1.2})$$

Par convention, si $h_X(x) = 0$, nous posons $\eta(x) = 0$. Nous avons, pour tout $f : \mathbb{R}^d \times \{0, 1\} \rightarrow \mathbb{R}_+$,

$$\begin{aligned} \mathbb{E}[f(X, Y)] &= \iint f(x, y)h_{X,Y}(x, y)\lambda(\mathrm{d}x)c(\mathrm{d}y) \\ &= \int_{\{x \in \mathbb{R}^d : h_X(x) \neq 0\}} h_X(x) \int f(x, y) \frac{h_{X,Y}(x, y)}{h_X(x)} c(\mathrm{d}y) \lambda(\mathrm{d}x) \\ &= \int_{\{x \in \mathbb{R}^d : h_X(x) \neq 0\}} h_X(x) \{f(x, 1)\eta(x) + f(x, 0)(1 - \eta(x))\} \lambda(\mathrm{d}x) . \end{aligned}$$

Définition III-1.1 (Classifieur, probabilité d'erreur). Une fonction $g : \mathbb{R}^d \rightarrow \mathcal{Y}$ (où $\mathcal{Y} = \{-1, +1\}$ ou $\mathcal{Y} = \{0, 1\}$) 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

$$R(g) = \mathbb{P}(g(X) \neq Y) . \quad (\text{III-1.3})$$



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 $\mathbf{x} = (1, x^{(1)}, \dots, x^{(d)})$ est le vecteur des attributs, $\mathbf{a} = (a_0, \dots, a_d)$ est le vecteur de *poids synaptiques* et σ est la fonction d'activation que nous prenons ici égale à la fonction de Heaviside

$$\text{heaviside}(x) = \begin{cases} 0 & x \leq 0 \\ +1 & x > 0 \end{cases}. \quad (\text{III-1.4})$$

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 ([Rosenblatt, 1960, Rosenblatt, 1962]) 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

$$\psi(\mathbf{x}) = c_0 + \sum_{i=1}^d c_i x^{(i)}. \quad (\text{III-1.5})$$

Le perceptron monocouche est un cas particulier d'un réseau de neurones sans couche cachée (appelé aussi *perceptron multicouche* ou *réseau connexionniste* ; voir [Anthony and Bartlett, 1999, Goodfellow et al., 2016]).

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 a 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

$$u_i = \sigma(\psi_i(\mathbf{x})), \quad (\text{III-1.6})$$

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 $\lim_{z \rightarrow -\infty} \sigma(z) = 0$, $\lim_{z \rightarrow +\infty} \sigma(z) = +1$. Des exemples de fonctions d'activation incluent le *signe*, $\sigma(x) = \text{signe}(x)$, la fonction logistique (ou sigmoïde)

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (\text{III-1.7})$$

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

$$\psi(\mathbf{u}) = c_0 + \sum_{i=1}^k c_i u_i, \quad \mathbf{u} = (u_1, \dots, u_k),$$

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

$$g(\mathbf{x}) = \text{heaviside} \left(c_0 + \sum_{i=1}^k c_i \sigma(\psi_i(\mathbf{x})) \right). \quad (\text{III-1.8})$$

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. $\diamond$

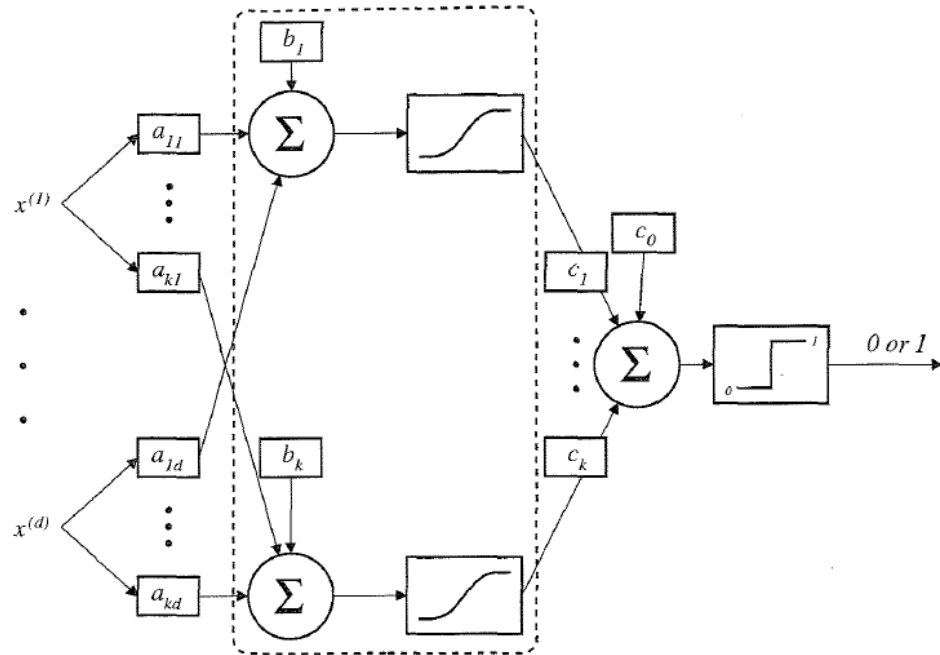


FIGURE III-1.2 – Un réseau neuronal avec une couche cachée.

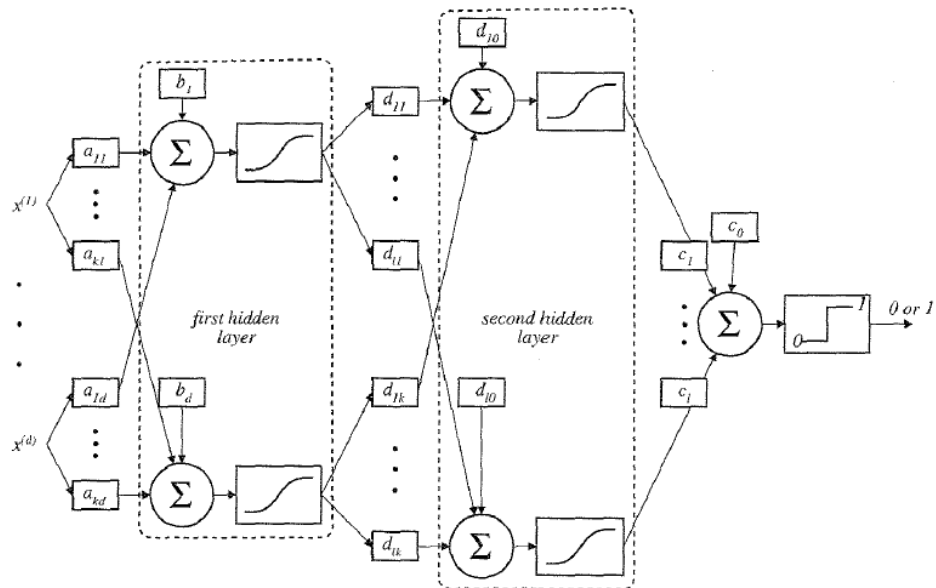


FIGURE III-1.3 – Un réseau neuronal avec deux couches cachées.

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.*

$$R^* = \inf_{g \in \text{mesurable}} R(g) . \quad (\text{III-1.9})$$

Une règle de classification qui atteint le risque bayésien est appelée classifieur bayésien.

Nous allons montrer que toute règle de classification qui satisfait

$$g^*(x) = \begin{cases} 1 & \text{si } \eta(x) > 1/2 \\ 0 & \text{autrement,} \end{cases}$$

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 $\lambda_{\text{Leb}} \otimes c$ 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

$$\begin{aligned} \mathbb{P}(g^*(X) \neq Y) &= \mathbb{P}(g^*(X) = 1, Y = 0) + \mathbb{P}(g^*(X) = 0, Y = 1) \\ &= \mathbb{P}(X \geq 0, Y = 0) + \mathbb{P}(X < 0, Y = 1) \\ &= \int_0^1 h_{X,Y}(x, 0) \lambda_{\text{Leb}}(dx) + \int_{-1}^0 h_{X,Y}(x, 1) \lambda_{\text{Leb}}(dx) = \frac{0.1}{2} + \frac{0.1}{2} = 0.1 . \end{aligned} \quad \diamond$$

Nous sommes maintenant en mesure d'énoncer notre premier résultat.

Théorème III-1.6. *Pour tout classifieur $g : \mathbb{R}^d \rightarrow \{0, 1\}$,*

$$\mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y) .$$

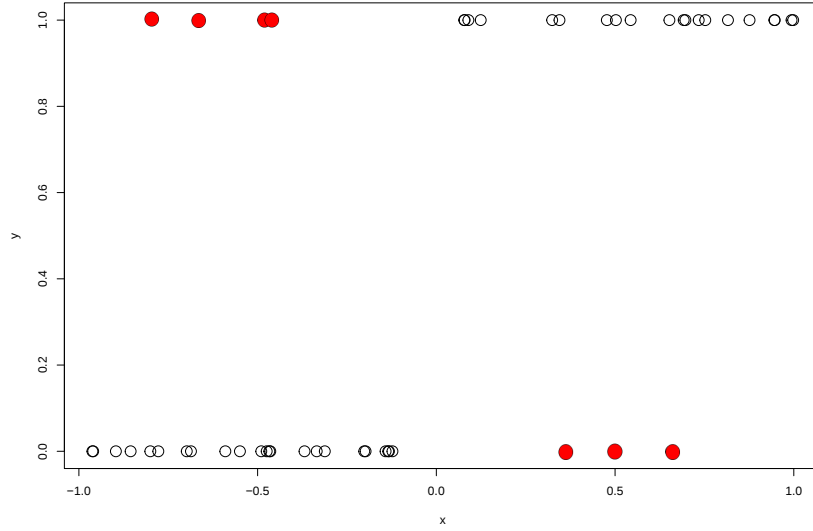


FIGURE III-1.4 – 40 échantillons engendrés par le modèle de l’Exemple III-1.5. En rouge, les points mal classés par le classifieur bayésien.

Démonstration. La probabilité d’erreur conditionnelle de toute règle de classification g peut être exprimée comme suit

$$\begin{aligned} \mathbb{P}(g(X) \neq Y | X) &= 1 - \mathbb{P}(Y = g(X) | X) \\ &= 1 - \{\mathbb{P}(Y = 1, g(X) = 1 | X) + \mathbb{P}(Y = 0, g(X) = 0 | X)\} \\ &= 1 - \{\mathbb{1}_{\{g(X)=1\}} \eta(X) + \mathbb{1}_{\{g(X)=0\}} (1 - \eta(X))\} , \end{aligned}$$

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

$$\begin{aligned} &\mathbb{P}(g(X) \neq Y | X) - \mathbb{P}(g^*(X) \neq Y | X) \\ &= \eta(X)(\mathbb{1}_{\{g^*(X)=1\}} - \mathbb{1}_{\{g(X)=1\}}) + (1 - \eta(X))(\mathbb{1}_{\{g^*(X)=0\}} - \mathbb{1}_{\{g(X)=0\}}) \\ &= (2\eta(X) - 1)(\mathbb{1}_{\{g^*(X)=1\}} - \mathbb{1}_{\{g(X)=1\}}) \geq 0 , \end{aligned} \tag{III-1.10}$$

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 . $\square$

La preuve ci-dessus montre que

$$R(g) = 1 - \mathbb{E} [\mathbb{1}_{\{g(X)=1\}} \eta(X) + \mathbb{1}_{\{g(X)=0\}} (1 - \eta(X))] , \tag{III-1.11}$$

et en particulier

$$R^* = 1 - \mathbb{E} [\mathbb{1}_{\{\eta(X) > 1/2\}} \eta(X) + \mathbb{1}_{\{\eta(X) \leq 1/2\}} (1 - \eta(X))] = \mathbb{E} [\min(\eta(X), 1 - \eta(X))] .$$

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 $g : \mathbb{R}^d \rightarrow [0, 1]$ est une fonction mesurable

$$\mathbb{E}[(\eta(X) - Y)^2] \leq \mathbb{E}[(g(X) - Y)^2].$$

En effet,

$$\begin{aligned} \mathbb{E}[(g(X) - Y)^2 | X] &= \mathbb{E}[(g(X) - \eta(X) + \eta(X) - Y)^2 | X] \\ &= (g(X) - \eta(X))^2 + 2(g(X) - \eta(X))\mathbb{E}[\eta(X) - Y | X] + \mathbb{E}[(\eta(X) - Y)^2 | X] \\ &= (g(X) - \eta(X))^2 + \mathbb{E}[(\eta(X) - Y)^2 | X] \geq \mathbb{E}[(\eta(X) - Y)^2 | X]. \end{aligned} \quad \diamond$$

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 $\mathcal{C}$.

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

$$\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{g(X_i) \neq Y_i\}}.$$

La loi des grands nombres montre que, pour toute règle de prédiction g , le risque empirique $\hat{R}_n(g)$ converge en probabilité vers le risque $L(g)$.

Exemple III-1.9 (Classifieurs linéaires). Nous supposons ici $Y = \{-1, 1\}$ et considérons l'ensemble des classifieurs linéaires sur $\mathbb{R}^d$,

$$\mathcal{C} = \left\{ \mathbf{x} \mapsto g(\mathbf{x}) = \text{signe}(\mathbf{w}^T \mathbf{x}) : \mathbf{w} \in \mathbb{R}^d \right\},$$

où par convention $\text{signe}(x) = 2\text{heaviside}(x) - 1$. Dans ce cas, le classifieur qui minimise le risque empirique est obtenu en déterminant un vecteur de poids qui minimise le risque empirique

$$\mathbf{w}_n^* \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\text{signe}(\mathbf{w}^T X_i) \neq Y_i\}}.$$

Ce problème n'admet pas de façon générale une unique solution. Le classifieur minimisant le risque empirique est alors donné par $g_n(\mathbf{x}) = \text{heaviside}(\mathbf{w}_n^T \mathbf{x})$.

L'algorithme du perceptron travaille directement sur le vecteur des poids synaptiques $\mathbf{a}$ 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

$$\sum_{i=1}^n \mathbb{1}_{\{\text{signe}(\mathbf{w}_*^T X_i) \neq Y_i\}} = 0.$$

On peut déterminer $\mathbf{w}_*$ en utilisant l'algorithme du perceptron.

La version stochastique 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. $\diamond$

III-1.4 Sur-apprentissage

Plus la classe $\mathcal{C}$ est grande, plus petite est la quantité $\inf_{g \in \mathcal{C}} R(g)$. Toutefois, choisir une classe $\mathcal{C}$ trop grande ne permet pas nécessairement de réduire le risque d'un minimiseur du risque empirique.

Prenons un exemple extrême : supposons que $\mathcal{C}$ 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 avons 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). $\diamond$

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

$$\mathbb{E}[R(g_n)] = \mathbb{P}(g_n(X; D_n) \neq Y).$$

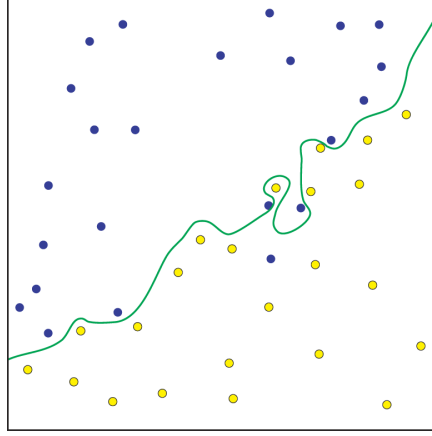


FIGURE III-1.5 – Exemple de sur-apprentissage. La frontière de décision du classifieur s’adapte afin d’étiqueter correctement toutes les données d’apprentissage, alors que le classifieur Bayésien optimal est un hyperplan.

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* :

$$\mathbb{E}[R(g_n)] - R^* . \quad (\text{III-1.12})$$

L’excès de risque peut se décomposer en deux termes :

— le *risque d’estimation*

$$\mathbb{E}[R(g_n)] - \inf_{g \in \mathcal{C}} R(g) , \quad (\text{III-1.13})$$

— le *risque d’approximation*

$$\inf_{g \in \mathcal{C}} R(g) - R^* . \quad (\text{III-1.14})$$

Le risque d’approximation quantifie la perte associée au choix de la classe $\mathcal{C}$. Il ne dépend pas des données d’apprentissage, mais seulement du choix de la famille de classifieurs $\mathcal{C}$ 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 $\mathcal{C}$. Il dépend de la classe $\mathcal{C}$ (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 $\mathcal{C}$ est petite (le cas extrême étant que $\mathcal{C}$ contient uniquement une règle de classification), alors l’erreur d’estimation est faible (dans le cas extrême où $\mathcal{C}$ est un singleton, cette erreur est nulle), mais l’erreur d’approximation sera très importante. Si la classe $\mathcal{C}$ est grande (le cas extrême étant que $\mathcal{C}$ 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

$$g^*(x) = \mathbb{1}_{[-1/2, 1/2]}(x) .$$

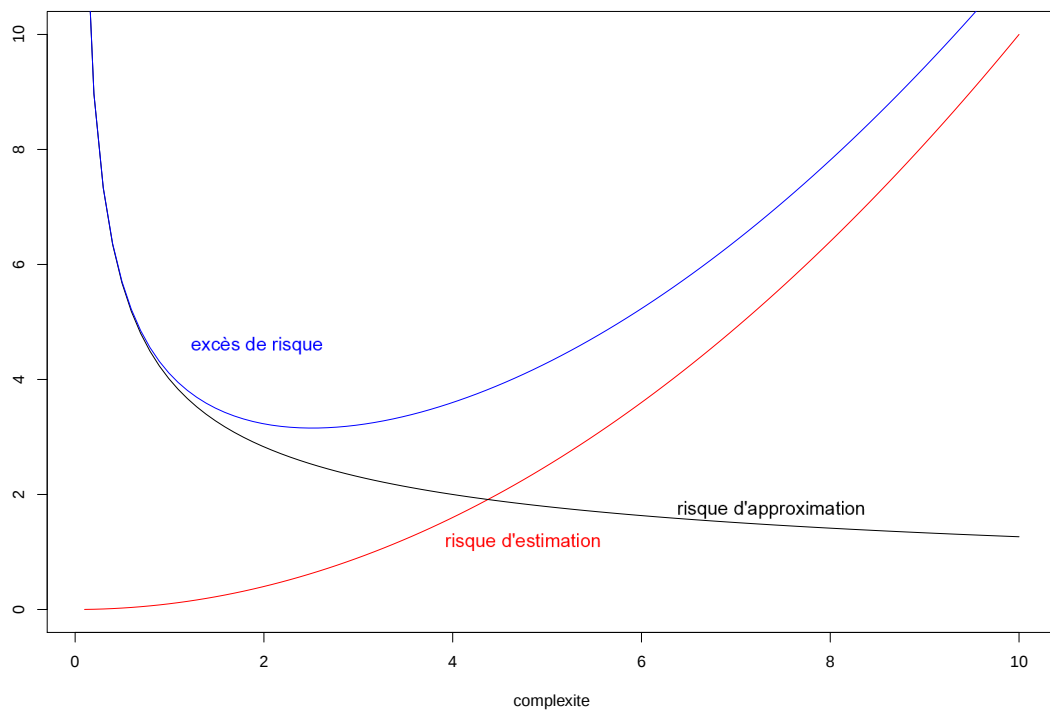


FIGURE III-1.6 – Illustration du compromis entre le risque d'approximation et le risque d'estimation

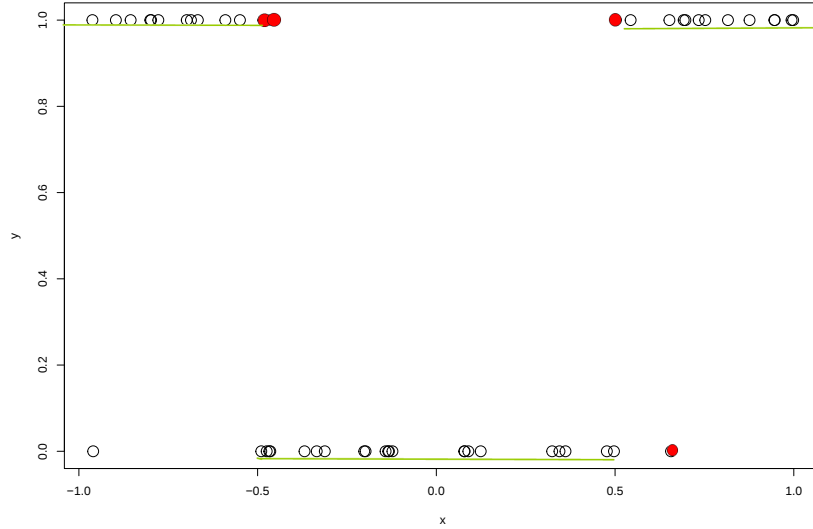


FIGURE III-1.7 – 50 échantillons engendrés par le modèle de l’Exemple III-1.5. En rouge les points mal classés par le classifieur bayésien

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

$$\mathcal{C}_R = \left\{ g(x) = \text{heaviside}(b + ax) : (a, b) \in \mathbb{R}^2 \right\}.$$

Nous déterminons les poids a, b en minimisant le risque empirique

$$\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{g(X_i) \neq Y_i\}}.$$

Nous considérons maintenant la classe d’estimateurs quadratiques

$$\mathcal{C}_Q = \left\{ g(x) = \text{heaviside}(b + a_1x + a_2x^2) : (a_1, a_2, b) \in \mathbb{R}^3 \right\}.$$

◇

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 $\mathcal{C}$ est trop petite pour contenir une bonne approximation du classifieur de Bayes.

Deux approches sont utilisées pour réaliser ce compromis :

- Restreindre la classe $\mathcal{C}$ 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.

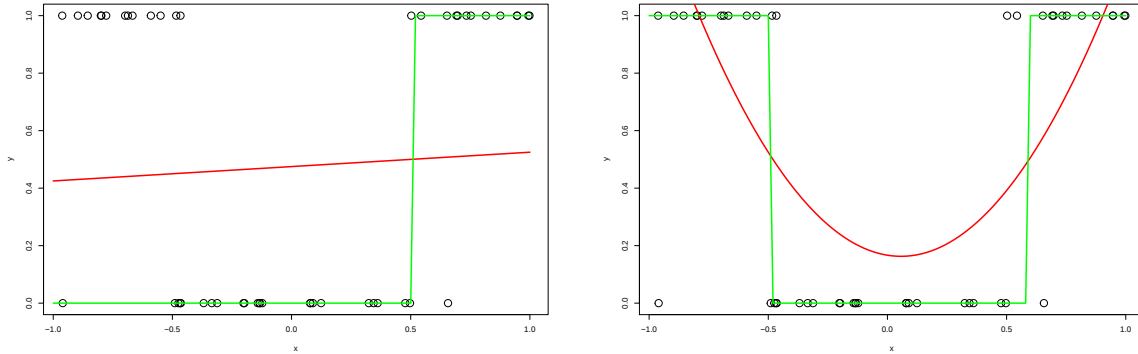


FIGURE III-1.8 – Panneau de gauche : classifieur linéaire ; Panneau de droite : classifieur quadratique

- 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 $\{\mathcal{C}_k\}_{k=1}^K$, dont la “complexité” croît avec l’indice k : par exemple $\mathcal{C}_k$ 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 ($\mathcal{C}_1$ est la classe des discriminateurs linéaires, $\mathcal{C}_2$ la classe des discriminateurs quadratiques, etc..). La méthode du crible procède en deux étapes.

Le première étape consiste à calculer pour chaque classe $\mathcal{C}_k$ une règle de classification $\hat{g}_k$ (par exemple, la règle minimisant le risque empirique sur cette classe $\hat{g}_k^* = \arg \min_{g \in \mathcal{C}_k} \hat{R}_n(g)$). Nous disposons ainsi pour chaque classe $\mathcal{C}_k$ 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 $\{\mathcal{C}_k : k \in \mathbb{N}\}$. 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 $\hat{g}_k^* = \arg \min_{g \in \mathcal{C}_k} \hat{R}_m(g)$, où $\hat{R}_m(g) = m^{-1} \sum_{i=1}^m \mathbb{1}_{\{g(X_i) \neq Y_i\}}$). 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 minimisant l’erreur “hold-out”

$$\hat{g}_n^{**} \in \arg \min_k \hat{R}_V(\hat{g}_k^*).$$

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 l'on 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

$$R(\hat{g}_n) \xrightarrow{\mathbb{P}\text{-prob}} R^* . \quad (\text{III-1.15})$$

où R^* est le risque bayésien (voir Définition III-1.4-Eq.(III-1.9)).

Remarquons que la condition (III-1.15) équivaut à

$$\lim_{n \rightarrow \infty} \mathbb{E}[R(\hat{g}_n)] = R^* , \quad (\text{III-1.16})$$

En effet, pour tout $\delta > 0$, en appliquant l'inégalité de Markov,

$$\mathbb{P}(R(\hat{g}_n) - R^* \geq \delta) \leq \delta^{-1} \{\mathbb{E}[R(\hat{g}_n)] - R^*\} \rightarrow_{n \rightarrow \infty} 0 .$$

D'autre part, puisque $R^* \leq R(\hat{g}_n) \leq 1$,

$$\begin{aligned} \mathbb{E}[R(\hat{g}_n)] - R^* &= \mathbb{E}[\{R(\hat{g}_n) - R^*\} \mathbb{1}_{\{R(\hat{g}_n) - R^* \geq \delta\}}] + \mathbb{E}[\{R(\hat{g}_n) - R^*\} \mathbb{1}_{\{R(\hat{g}_n) - R^* < \delta\}}] \\ &\leq \mathbb{P}(R(\hat{g}_n) - R^* \geq \delta) + \delta \end{aligned}$$

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 $\mathcal{C}$ 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 [Valiant, 1984]. 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 ; [Valiant, 1984]). Soit $\varepsilon > 0$, $\delta \in]0, 1[$ et $\mathcal{C}$ une famille de règles de classification. Nous dirons qu'une règle $g_n \in \mathcal{C}$ est Probablement Approximativement Correcte avec une précision ε et une confiance $1 - \delta$ (où plus simplement, g_n est (ε, δ) -PAC) si

$$\mathbb{P} \left(R(g_n) - \inf_{g \in \mathcal{C}} R(g) \geq \varepsilon \right) \leq \delta, \quad (\text{III-2.1})$$

où, pour $g \in \mathcal{C}$, $R(g) = \mathbb{P}(g(X) \neq Y)$ est le risque de classification.

De façon équivalente, la règle de classification g_n est (ε, δ) -PAC si

$$R(g_n) - \inf_{g \in \mathcal{C}} R(g) \leq \varepsilon \quad \text{avec une probabilité } 1 - \delta. \quad (\text{III-2.2})$$

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 $\mathcal{C}$ une classe de règles de classification. Notons par $\hat{g}_n^*$ une règle de classification de $\mathcal{C}$ qui minimise le risque empirique dans $\mathcal{C}$. Le résultat suivant montre que l'erreur conditionnelle de $\hat{g}_n$ est contrôlée par la déviation uniforme $\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|$ sur la classe $\mathcal{C}$.

Théorème III-2.2. Soit $\hat{g}_n^* \in \mathcal{C}$ un minimiseur du risque empirique sur $\mathcal{C}$:

$$\hat{R}_n(\hat{g}_n^*) \leq \hat{R}_n(g) \quad \text{pour tout } g \in \mathcal{C}.$$

Alors, l'excès de risque ainsi que la différence entre le risque empirique et le risque de $\hat{g}_n^*$ satisfont

$$R(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g) \leq 2 \sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|, \quad (\text{III-2.3})$$

$$|\hat{R}_n(\hat{g}_n^*) - R(\hat{g}_n^*)| \leq \sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|. \quad (\text{III-2.4})$$

Démonstration. Pour tout $\varepsilon > 0$, il existe $g_\varepsilon \in \mathcal{C}$ tel que $R(g_\varepsilon) - \varepsilon \leq \inf_{g \in \mathcal{C}} R(g)$. Par conséquent,

$$\begin{aligned} R(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g) &= R(\hat{g}_n^*) - \hat{R}_n(\hat{g}_n^*) + \hat{R}_n(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g) \\ &\leq R(\hat{g}_n^*) - \hat{R}_n(\hat{g}_n^*) + \hat{R}_n(g_\varepsilon) - R(g_\varepsilon) + \varepsilon, \end{aligned}$$

où on a également utilisé que, par définition $\hat{R}_n(\hat{g}_n^*) \leq \hat{R}_n(g_\varepsilon)$. On en déduit finalement

$$R(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g) \leq 2 \sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| + \varepsilon.$$

L'inégalité (III-2.3) en découle immédiatement car $\varepsilon > 0$ est arbitraire. L'inégalité (III-2.4) est trivialement satisfaite. $\square$

Nous voyons que $\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|$ fournit des bornes supérieures pour deux quantités d'intérêt simultanément :

1. Une borne supérieure pour l'excès de risque de $\hat{g}_n^*$:

$$R(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g).$$

2. Une borne supérieure pour l'erreur $|\hat{R}_n(\hat{g}_n^*) - R(\hat{g}_n^*)|$ commise en estimant le risque $R(\hat{g}_n^*)$ de $\hat{g}_n^*$ par son risque empirique $\hat{R}_n(\hat{g}_n^*)$.

Ainsi, l'estimateur du minimum du risque empirique $\hat{R}_n(\hat{g}_n^*)$ est biaisé de façon optimiste, mais son biais est borné par $\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|$. Il est donc utile de disposer de bornes pour cette quantité.

Pour chaque $g \in \mathcal{C}$, la variable aléatoire $n\hat{R}_n(g)$ 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 $\mathcal{C}$ est de cardinal fini et que $\min_{g \in \mathcal{C}} R(g) = 0$. Ceci implique en particulier que l'estimateur bayésien est un élément de la classe $\mathcal{C}$.

Théorème III-2.3. *Supposons que le cardinal de la classe $\mathcal{C}$, $|\mathcal{C}|$, est fini et que $\min_{g \in \mathcal{C}} R(g) = 0$. Notons $\hat{g}_n^* \in \arg \min_{g \in \mathcal{C}} \hat{R}_n(g)$ un minimiseur du risque empirique. Alors pour tout $n \in \mathbb{N}$ et $\varepsilon > 0$,*

$$\mathbb{P}(R(\hat{g}_n^*) \geq \varepsilon) \leq |\mathcal{C}| e^{-n\varepsilon}. \quad (\text{III-2.5})$$

Démonstration. Soit $g^* \in \arg \min_{g \in \mathcal{C}} R(g)$ le classifieur bayésien. Par hypothèse, $R(g^*) = 0$. Comme $\mathbb{E}[\hat{R}_n(g^*)] = 0$ et $\hat{R}_n(g^*) \geq 0$, nous avons $\mathbb{P}(\hat{R}_n(g^*) = 0) = 1$. Comme $\hat{R}_n(\hat{g}_n^*) \leq \hat{R}_n(g^*)$, nous avons donc aussi $\mathbb{P}(\hat{R}_n(\hat{g}_n^*) = 0) = 1$. Ainsi, $R(\hat{g}_n^*) \geq \varepsilon$ seulement s'il existe $g \in \mathcal{C}$ tel que $R(g) \geq \varepsilon$ et $\hat{R}_n(g) = 0$. Autrement dit

$$\mathbb{P}(R(\hat{g}_n^*) \geq \varepsilon) \leq \mathbb{P}\left(\bigcup_{g \in \mathcal{C}, R(g) \geq \varepsilon} \{\hat{R}_n(g) = 0\}\right).$$

Par la borne d'union, on a donc

$$\mathbb{P}(R(\hat{g}_n^*) \geq \varepsilon) \leq \sum_{g \in \mathcal{C}, R(g) \geq \varepsilon} \mathbb{P}(\hat{R}_n(g) = 0). \quad (\text{III-2.6})$$

La variable aléatoire $n\hat{R}_n(g)$ est une variable binomiale de probabilité de succès $R(g)$. Nous avons donc $\mathbb{P}(\hat{R}_n(g) = 0) = (1 - R(g))^n$ et (III-2.6) implique donc que

$$\mathbb{P}(R(\hat{g}_n^*) \geq \varepsilon) \leq \sum_{g \in \mathcal{C}, R(g) \geq \varepsilon} (1 - \varepsilon)^n \leq |\mathcal{C}| (1 - \varepsilon)^n.$$

En remarquant que $(1 - \varepsilon)^n \leq e^{-n\varepsilon}$, on en déduit (III-2.5). $\square$

Corollaire III-2.4. *Supposons que le cardinal de la classe $\mathcal{C}$, $|\mathcal{C}|$, est fini et que $\min_{g \in \mathcal{C}} R(g) = 0$. 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*

$$n(\varepsilon, \delta) = \frac{\log |\mathcal{C}| + \log(1/\delta)}{\varepsilon}.$$

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*

$$\forall \varepsilon > 0, \quad \mathbb{P}(Z \geq \varepsilon) \leq C e^{-\alpha \varepsilon}. \quad (\text{III-2.7})$$

Alors,

$$\mathbb{E}[Z] \leq \frac{1 + \log(C)}{\alpha}.$$

Démonstration. Comme Z est à valeurs dans $\mathbb{R}_+$,

$$\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z > t) dt.$$

Pour tout $u > 0$, en utilisant $\mathbb{P}(Z > t) \leq 1$, on a alors

$$\mathbb{E}[Z] \leq u + C \int_u^\infty e^{-\alpha t} dt \leq u + \frac{C e^{-\alpha u}}{\alpha}.$$

En appliquant cette borne avec $u = (\log C)/\alpha$, nous obtenons le résultat désiré. $\square$

Corollaire III-2.6. *Supposons que le cardinal de la classe $\mathcal{C}$, $|\mathcal{C}|$, est fini et que $\min_{g \in \mathcal{C}} R(g) = 0$. Alors*

$$\mathbb{E}[R(\hat{g}_n^*)] \leq \frac{1 + \log(|\mathcal{C}|)}{n}.$$

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 $\mathcal{C}$. 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 $g \in \mathcal{C}$, $n\hat{R}_n(g)$ 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-3.9 et le Corollaire IV-3.10 pour la preuve de ce résultat).

Si $Z_1, \dots, Z_n$ 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 $t > 0$,

$$\mathbb{P}(\bar{Z}_n - p \geq t) \leq \exp(-2nt^2).$$

$$\mathbb{P}(\bar{Z}_n - p \leq -t) \leq \exp(-2nt^2).$$

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 $\mathcal{C}$, $|\mathcal{C}|$ soit fini. Alors nous avons pour tout $n \in \mathbb{N}$ et $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{g \in \mathcal{C}} \hat{R}_n(g) - R(g) > \varepsilon\right) \leq |\mathcal{C}| e^{-2n\varepsilon^2}, \quad (\text{III-2.8})$$

$$\mathbb{P}\left(\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| > \varepsilon\right) \leq 2|\mathcal{C}| e^{-2n\varepsilon^2}. \quad (\text{III-2.9})$$

Démonstration. Notons $\mathcal{C} = \{g_i\}_{i=1}^{|\mathcal{C}|}$. La borne de l'union montre que

$$\mathbb{P}\left(\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| > \varepsilon\right) \leq \sum_{i=1}^{|\mathcal{C}|} \mathbb{P}(|\hat{R}_n(g_i) - R(g_i)| > \varepsilon).$$

Les inégalités (III-2.8) et (III-2.9) découlent du Corollaire IV-3.10 et du fait les variables $\{\mathbb{1}_{\{g(X_i) \neq Y_i\}}\}_{i=1}^n$ sont distribuées suivant une loi de Bernoulli. $\square$

La différence entre (III-2.9) et l'inégalité de Hoeffding (Théorème IV-3.9) réside dans le terme $|\mathcal{C}|$ du membre de droite. Ce terme garantit que la borne soit valable simultanément pour toutes les règles de classification de $\mathcal{C}$.

Les bornes de déviation uniforme obtenues dans le Théorème III-2.7 ont de nombreuses conséquences. Elles montrent en particulier que, pour tout $\delta \in]0, 1[$,

$$\mathbb{P}\left(\forall g \in \mathcal{C}, \quad R(g) \leq \widehat{R}_n(g) + C(\mathcal{C}, n, \delta)\right) \geq 1 - \delta, \quad (\text{III-2.10})$$

où

$$C(\mathcal{C}, n, \delta) = \sqrt{\frac{\log |\mathcal{C}| + \log(1/\delta)}{2n}}.$$

Ainsi, le risque de chaque classifieur est contrôlé uniformément sur la classe $\mathcal{C}$ 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 $1 - \delta$,

$$R(\hat{g}_n^*) \leq \widehat{R}_n(\hat{g}_n^*) + C(\mathcal{C}, n, \delta).$$

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

$$\mathbb{P}\left(R(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g) \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{g \in \mathcal{C}} |\widehat{R}_n(g) - R(g)| \geq \varepsilon/2\right) \leq 2|\mathcal{C}|e^{-n\varepsilon^2/2}. \quad (\text{III-2.11})$$

On peut déduire de cette inégalité une borne PAC agnostique.

Corollaire III-2.8. *Supposons que le cardinal de la classe $\mathcal{C}$, $|\mathcal{C}|$ 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*

$$n(\varepsilon, \delta) = \frac{2\{\log(1/\delta) + \log(2|\mathcal{C}|)\}}{\varepsilon^2}.$$

Notons que $\log(|\mathcal{C}|)$ est aussi le nombre d'éléments binaires nécessaires pour spécifier une fonction g particulière dans $\mathcal{C}$. 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 $\delta \in]0, 1[$,

$$\mathbb{P}\left(R(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g) \leq \sqrt{\frac{\log(2|\mathcal{C}|) + \log(1/\delta)}{2n}}\right) \geq 1 - \delta. \quad (\text{III-2.12})$$

Il est aussi intéressant de déterminer une borne de

$$\mathbb{E}\left[\sup_{g \in \mathcal{C}} |\widehat{R}_n(g) - R(g)|\right].$$

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*

$$\forall \varepsilon > 0, \quad \mathbb{P}(Z \geq \varepsilon) \leq Ce^{-\alpha\varepsilon^2}.$$

Alors

$$\mathbb{E}[Z] \leq \sqrt{\frac{1 + \log(C)}{\alpha}}.$$

Démonstration. On a par hypothèse,

$$\forall \varepsilon > 0, \quad \mathbb{P}(Z^2 \geq \varepsilon) \leq C e^{-\alpha \varepsilon}.$$

Donc, d'après (III-2.7)

$$\mathbb{E}[Z^2] \leq \frac{1 + \log(C)}{\alpha}.$$

On conclut par l'inégalité de Cauchy-Schwarz. $\square$

Théorème III-2.10. *Supposons que le cardinal de l'ensemble des classifieurs $\mathcal{C}$ soit fini. Alors,*

$$\mathbb{E} \left[\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| \right] \leq \sqrt{\frac{1 + \log(2|\mathcal{C}|)}{2n}}, \quad (\text{III-2.13})$$

$$\mathbb{E}[R(\hat{g}_n)] \leq \inf_{g \in \mathcal{C}} R(g) + 2\sqrt{\frac{1 + \log(2|\mathcal{C}|)}{2n}}. \quad (\text{III-2.14})$$

Il est possible d'améliorer les constantes numériques dans ce résultat en combinant le Lemme IV-3.8 et l'inégalité de Pisier (voir le Théorème IV-3.13)

III-2.4 Une application : classification par histogramme

Soit $m \in \mathbb{N}$ et $\{Q_1, \dots, Q_m\}$ une partition finie de l'ensemble des attributs $X \subset \mathbb{R}^d$. Par exemple, si $X = [0, 1]^d$, nous pouvons diviser chaque axe en q intervalles de longueur égales et considérer la famille des hypercubes

$$Q_{i_1, \dots, i_d} = \prod_{j=1}^d [(i_j - 1)/q, i_j/q], (i_1, \dots, i_d) \in \{1, \dots, q\}^d.$$

On a alors $m = q^d$. Les histogrammes sur la partition $\{Q_1, \dots, Q_m\}$ sont les classifieurs de la forme

$$\mathcal{C} = \left\{ g : X \rightarrow \{0, 1\} : g(x) = \sum_{i=1}^m d_i \mathbb{1}_{Q_i}(x), (d_1, \dots, d_m) \in \{0, 1\}^m \right\}.$$

Pour tout $i \in \{1, \dots, m\}$, 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

$$\begin{aligned} \hat{R}_n(g) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{g(X_i) \neq Y_i\}} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \mathbb{1}_{\{g(X_i) \neq Y_i, X_i \in Q_j\}} \\ &= \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Q_j}(X_i) \left\{ 1 - \mathbb{1}_{\{d_j\}}(Y_i) \right\}. \end{aligned}$$

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)) :

$$\mathbb{P} \left(R(\hat{g}_n^*) - \inf_{g \in \mathcal{C}} R(g) \leq \sqrt{\frac{(m+1)\log(2) + \log(1/\delta)}{2n}} \right) \geq 1 - \delta ,$$

$$\mathbb{E}[R(\hat{g}_n)] - R^* \leq 2\sqrt{\frac{m\log(2) + \ln(2e)}{2n}} .$$

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 $\mathcal{C}$ 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 $\mathcal{C}$ (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 $\mathcal{C}$ 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 $\mathcal{C}$. 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 $\mathcal{C}$, valable indépendamment de la distribution des exemples (en supposant simplement que les exemples d'apprentissage sont indépendants). L'analyse PAC ("Probably Approximately 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 [Vapnik, 2000]. Ces idées ont débouché non seulement sur une méthode d'évaluation de l'erreur $L(\hat{g}_n) - \inf_{g \in \mathcal{C}} R(g)$ (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 d'une inégalité de concentration pour des fonctions de variables aléatoires indépendantes plus générales que la somme.

Soit X un ensemble borélien et $g : X^n \rightarrow \mathbb{R}$ une fonction mesurable (de n variables). Nous allons maintenant obtenir des bornes de déviation de $g(X_1, \dots, X_n)$ par rapport à sa moyenne lorsque les variables aléatoires $X_1, \dots, X_n$ 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'ambiguïté, pour tout $1 \leq m \leq n$, $X_{mn} = (X_m, \dots, X_n)$.

Définition III-3.1 (Différences bornées). La fonction mesurable $g : X^n \rightarrow \mathbb{R}$ est à différences bornées, si pour tout $i \in \{1, \dots, n\}$,

$$\sup_{x_1, \dots, x_n, x'_i \in X^{n+1}} |g(x_1, \dots, x_n) - g(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i,$$

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 $(X_1, \dots, X_n)$ n variables aléatoires indépendantes à valeurs dans un ensemble $X \in \mathcal{B}(\mathbb{R}^d)$. Soit $g : X^n \rightarrow \mathbb{R}$ une fonction mesurable à différences bornées

(voir Définition III-3.1). Alors pour tout $t > 0$,

$$\mathbb{P}\{g(X_{1n}) - \mathbb{E}[g(X_{1n})] \geq t\} \leq \exp\left(-2t^2 / \sum_{i=1}^n c_i^2\right) \quad (\text{III-3.1})$$

et

$$\mathbb{P}\{g(X_{1n}) - \mathbb{E}[g(X_{1n})] \leq -t\} \leq \exp\left(-2t^2 / \sum_{i=1}^n c_i^2\right). \quad (\text{III-3.2})$$

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-3.8.

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 $s > 0$,

$$\mathbb{E}[e^{sV} | Z] \leq e^{s^2 c^2 / 8}.$$

Démonstration. (du Théorème III-3.2) Posons $G_n = g(X_{1n})$, $V = G_n - \mathbb{E}[G_n]$ et définissons

$$V_i = \mathbb{E}[G_n | X_{1i}] - \mathbb{E}[G_n | X_{1i-1}], \quad i = 1, \dots, n;$$

avec la convention $\mathbb{E}[G_n | X_{10}] = \mathbb{E}[G_n]$. Par construction, $V = \sum_{i=1}^n V_i$. Introduisons

$$\forall i \in \{1, \dots, n\}, \quad H_i(X_{1i}) = \mathbb{E}[G_n | X_{1i}] = \int \cdots \int g(X_{1i}, x_{i+1n}) \prod_{j=i+1}^n \mu_j(dx_j).$$

En notant μ_i la distribution de X_i et en utilisant l'indépendance de X_{1n} , nous obtenons

$$\forall i \in \{1, \dots, n\}, \quad V_i = H_i(X_{1i}) - \int H_i(X_{1i-1}, x_i) \mu_i(dx_i).$$

Définissons les variables aléatoires

$$W_i = \sup_{u \in X} \left(H_i(X_{1i-1}, u) - \int H_i(X_{1i-1}, x_i) \mu_i(dx_i) \right),$$

$$Z_i = \inf_u \left(H_i(X_{1i-1}, u) - \int H_i(X_{1i-1}, x_i) \mu_i(dx_i) \right).$$

Clairement, $Z_i \leq V_i \leq W_i$ presque-sûrement et, comme g est à différences bornées

$$W_i - Z_i = \sup_{u \in X} \sup_{v \in X} (H_i(X_{1i-1}, u) - H_i(X_{1i-1}, v))$$

$$= \sup_{u \in X} \sup_{v \in X} \int \cdots \int \{g(X_{1i-1}, u, x_{i+1n}) - g(X_{1i-1}, v, x_{i+1n})\} \prod_{j=i+1}^n \mu_j(dx_j) \leq c_i.$$

D'après le Lemme III-3.3, pour tout $i \in \{1, \dots, n\}$ et $s > 0$, nous avons

$$\mathbb{E}[e^{sV_i} | X_{1i-1}] \leq e^{s^2 c_i^2 / 8}.$$

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 $s > 0$,

$$\mathbb{E} \left[e^{s \sum_{i=1}^n V_i} \right] = \mathbb{E} \left[e^{s \sum_{i=1}^{n-1} V_i} \mathbb{E} \left[e^{s V_n} \mid X_{1:n-1} \right] \right] = e^{s^2 c_n^2 / 8} \mathbb{E} \left[e^{s \sum_{i=1}^{n-1} V_i} \right].$$

Par une induction élémentaire, nous obtenons donc, pour tout $n \in \mathbb{N}$ et $s > 0$.

$$\mathbb{E} \left[e^{s \sum_{i=1}^n V_i} \right] \leq e^{s^2 \sum_{i=1}^n c_i^2 / 8}.$$

Finalement, en utilisant la borne de Chernoff (Corollaire IV-3.7), pour tout $s > 0$, nous avons

$$\mathbb{P}(G_n - \mathbb{E}[G_n] \geq t) \leq \mathbb{E} \left[e^{s \sum_{i=1}^n V_i} \right] e^{-st} \leq e^{-st} e^{s^2 \sum_{i=1}^n c_i^2 / 8}.$$

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 $-g$. $\square$

L'inégalité de McDiarmid a une conséquence immédiate mais essentielle.

Corollaire III-3.4. Soit $\mathcal{C}$ une classe quelconque de règles de classification $g : \mathbb{R}^d \rightarrow \{0, 1\}$. Alors

$$\forall \varepsilon > 0, \quad \mathbb{P} \left(\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| - \mathbb{E} \left[\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| \right] > \varepsilon \right) \leq e^{-2n\varepsilon^2}.$$

$$\forall \varepsilon > 0, \quad \mathbb{P} \left(\left| \sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| - \mathbb{E} \left[\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| \right] \right| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2}.$$

Démonstration. Considérons les fonctions

$$g((x_1, y_1), \dots, (x_n, y_n)) = \sup_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{g(x_i) \neq y_i\}} - R(g),$$

$$h((x_1, y_1), \dots, (x_n, y_n)) = \sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{g(x_i) \neq y_i\}} - R(g) \right|.$$

g et h sont des fonctions à différences bornées sur $(X \times \{0, 1\})^n$ avec $c_i = 1/n$, pour tout $i \in \{1, \dots, n\}$. Nous concluons en appliquant le Théorème III-3.2. $\square$

La variable aléatoire $\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|$ 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

$$\sup_{g \in \mathcal{C}} |L_n(g) - R(g)| \leq \mathbb{E} \left[\sup_{g \in \mathcal{C}} |L_n(g) - R(g)| \right] + \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (\text{III-3.3})$$

Pour obtenir une borne de risque, il reste à borner l'espérance $\mathbb{E}[\sup_{g \in \mathcal{C}} |L_n(g) - R(g)|]$.

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 $\mathcal{C}$ des règles de classification est borné par N , alors pour tout $\varepsilon > 0$, nous avons

$$\mathbb{P} \left(\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| > \varepsilon \right) \leq 2Ne^{-2n\varepsilon^2},$$

$$\mathbb{E} \left[\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| \right] \leq \sqrt{\frac{\ln(2N)}{2n}}.$$

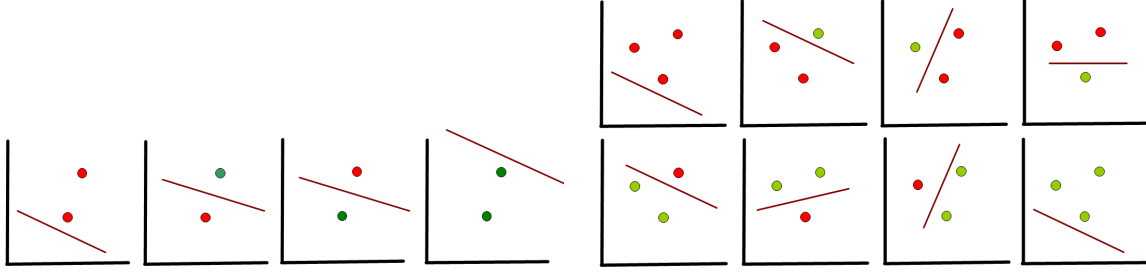


FIGURE III-3.1 – Pulvérisation une famille d’hyperplans en dimension 2 : panneau de gauche 2 points. Panneau de droite : 3 points

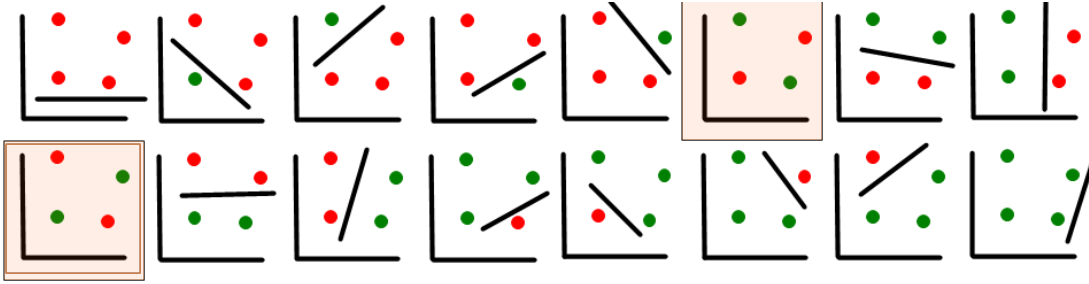


FIGURE III-3.2 – Pulvérisation d’un ensemble de quatre points par une famille d’hyperplans en dimension 2

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 $A \in \mathcal{B}(\mathbb{R}^d)$

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(X_i) .$$

Considérons une famille $\mathcal{A}$ d’ensembles de $\mathcal{B}(\mathbb{R}^d)$. L’objectif de ce paragraphe est l’étude de $\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$. L’inégalité de McDiarmid implique que pour tout $n \in \mathbb{N}$ et $t > 0$,

$$\mathbb{P} \left(\left| \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| - \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \right| > t \right) \leq 2e^{-2nt^2} .$$

Pour toute classe $\mathcal{A}$, l’écart maximal est concentré autour de sa moyenne. Nous allons maintenant montrer des inégalités pour

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right]$$

en fonction de certaines quantités décrivant la “complexité” combinatoire de la classe d’ensembles $\mathcal{A}$, ne dépendant pas de la distribution des observations. Etant donné n points $x_1, \dots, x_n$ de $\mathbb{R}^d$, définissons

$$N_{\mathcal{A}}(x_1, \dots, x_n) := |\{ \{x_1, \dots, x_n\} \cap A : A \in \mathcal{A} \}|$$

$N_{\mathcal{A}}(x_1, \dots, x_n)$ 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 $A \in \mathcal{A}$. On a toujours $N_{\mathcal{A}}(x_1, \dots, x_n) \leq 2^n$, et lorsque $N_{\mathcal{A}}(x_1, \dots, x_n) = 2^n$, on dit que la classe $\mathcal{A}$ *pulvérise* l’ensemble $\{x_1, \dots, x_n\}$.

Définition III-3.5 (Coefficient de pulvérisation de n points de la classe $\mathcal{A}$). Soit $\mathcal{A}$ une classe d'ensembles boréliens. Le coefficient de pulvérisation de n points de la classe $\mathcal{A}$ est donné par

$$\mathbb{S}_{\mathcal{A}}(n) = \max_{(x_1, \dots, x_n) \in \mathbb{R}^d} N_{\mathcal{A}}(x_1, \dots, x_n) .$$

Le théorème principal est la version suivante d'un résultat classique de Vapnik et Chervonenkis.

Théorème III-3.6 (Inégalité de Vapnik-Chervonenkis). Soient $X_1, \dots, X_n$ n variables aléatoires indépendantes, de même loi μ sur $\mathbb{R}^d$ et soit μ_n la mesure empirique correspondante. Alors, pour toute famille $\mathcal{A}$ d'ensembles boréliens de $\mathbb{R}^d$ et pour tout $\varepsilon > 0$, on a

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \sqrt{\frac{2 \log 2 \mathbb{S}_{\mathcal{A}}(2n)}{n}} .$$

Démonstration. Soit $(X'_1, \dots, X'_n)$ une copie indépendante de $(X_1, \dots, X_n)$ ($(X'_1, \dots, X'_n)$ est indépendant de $(X_1, \dots, X_n)$ et de même loi que $(X_1, \dots, X_n)$). Définissons n variables indépendantes de Rademacher : $\sigma_1, \dots, \sigma_n$ telles que $\mathbb{P}(\sigma_1 = -1) = \mathbb{P}(\sigma_1 = 1) = 1/2$ et indépendantes de (X_{1n}, X'_{1n}) . Alors, en notant $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\mu'_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(X'_i),$$

nous avons

$$\begin{aligned} \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] &= \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mathbb{E} [\mu_n(A) - \mu'_n(A) | X_{1n}]| \right] \\ &\leq \mathbb{E} \left[\sup_{A \in \mathcal{A}} \mathbb{E} [|\mu_n(A) - \mu'_n(A)| | X_{1n}] \right] \leq \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu'_n(A)| \right] \\ &= \frac{1}{n} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i)) \right| \right] , \end{aligned}$$

où nous avons utilisé que, pour tout $A \in \mathcal{A}$, 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

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{1}{n} \mathbb{E} \left[\mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i)) \right| \middle| X_{1n}, X'_{1n} \right] \right] .$$

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

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i \{ \mathbb{1}_A(x_i) - \mathbb{1}_A(x'_i) \} \right| \right] .$$

Pour tout (x_{1n}, x'_{1n}) , nous notons par $\hat{\mathcal{A}}(x_{1n}, x'_{1n}) \subset \mathcal{A}$ la collection d'ensembles de $\mathcal{A}$ vérifiant les conditions suivantes

— Pour tout $B, C \in \hat{\mathcal{A}}(x_{1n}, x'_{1n})$, $B \neq C$, nous avons

$$B \cap \{x_{1n}, x'_{1n}\} \neq C \cap \{x_{1n}, x'_{1n}\} .$$

Deux ensembles distincts de $\hat{\mathcal{A}}(x_{1n}, x'_{1n})$ ont des intersections différentes avec (x_{1n}, x'_{1n})

— Pour tout $A \in \mathcal{A}$, il existe $B \in \hat{\mathcal{A}}(x_{1n}, x'_{1n})$ 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 $A \in \mathcal{A}$ est présente dans $\hat{\mathcal{A}}(x_{1n}, x'_{1n})$ une fois et une seule.** Par définition du coefficient de pulvérisation de $2n$ points, nous avons $|\hat{\mathcal{A}}(x_{1n}, x'_{1n})| \leq \mathbb{S}_{\mathcal{A}}(2n)$, et

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i(\mathbb{1}_A(x_i) - \mathbb{1}_A(x'_i)) \right| \right] = \mathbb{E} \left[\max_{A \in \hat{\mathcal{A}}(x_{1n}, x'_{1n})} \left| \sum_{i=1}^n \sigma_i(\mathbb{1}_A(x_i) - \mathbb{1}_A(x'_i)) \right| \right].$$

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-3.8 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-3.13 (ou le Théorème III-2.7 et le Lemme III-2.9) implique immédiatement que

$$\mathbb{E} \left[\max_{A \in \hat{\mathcal{A}}} \left| \sum_{i=1}^n \sigma_i(\mathbb{1}_A(x_i) - \mathbb{1}_A(x'_i)) \right| \right] \leq \sqrt{2n \log 2 \mathbb{S}_{\mathcal{A}}(2n)}.$$

□

Remarque III-3.7. La forme originale de l'inégalité Vapnik-Chervonenkis est la suivante

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| > t \right) \leq 4 \mathbb{S}_{\mathcal{A}}(2n) e^{-nt^2/8}.$$

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 $\mathbb{S}_{\mathcal{A}}(2n)$ est la clé de la compréhension du comportement de $\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$. Les classes pour lesquelles $\mathbb{S}_{\mathcal{A}}(2n)$ croît à un taux sous-géométrique avec le nombre d'observations n permettent de "généraliser" dans le sens où $\mathbb{E}[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|]$ 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 $\mathcal{A}$ 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 $\mathcal{A}$ et $\mathcal{B}$ des classes d'ensembles boréliens de $\mathbb{R}^d$, et $n, m \geq 1$ des entiers. Alors

- (i) $\mathbb{S}_{\mathcal{A}}(n+m) \leq \mathbb{S}_{\mathcal{A}}(n) \mathbb{S}_{\mathcal{A}}(m)$;
- (ii) Si $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, alors $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n) + \mathbb{S}_{\mathcal{B}}(n)$;
- (iii) Si $\mathcal{C} = \{C = A^c : A \in \mathcal{A}\}$, alors $\mathbb{S}_{\mathcal{C}}(n) = \mathbb{S}_{\mathcal{A}}(n)$;
- (iv) Si $\mathcal{C} = \{C = A \cap B : A \in \mathcal{A} \text{ et } B \in \mathcal{B}\}$, alors $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n) \mathbb{S}_{\mathcal{B}}(n)$;
- (v) Si $\mathcal{C} = \{C = A \cup B : A \in \mathcal{A} \text{ et } B \in \mathcal{B}\}$, alors $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n) \mathbb{S}_{\mathcal{B}}(n)$;
- (vi) Si $\mathcal{C} = \{C = A \times B : A \in \mathcal{A} \text{ et } B \in \mathcal{B}\}$, alors $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n) \mathbb{S}_{\mathcal{B}}(n)$.

Démonstration. Les assertions (i), (ii), (iii), et (vi) découlent de façon immédiate de la définition.

Pour montrer (iv), prenons $x_{1n} = \{x_1, \dots, x_n\}$ et posons $N = |\mathcal{A}(x_{1n})| \leq \mathbb{S}_{\mathcal{A}}(n)$. Notons $A_1, A_2, \dots, A_N$ les N ensembles (deux à deux différents) de la forme $\{x_1, \dots, x_n\} \cap A$, $A \in \mathcal{A}$. Pour tout $i \in \{1, \dots, N\}$, nous avons $\mathbb{S}_{\mathcal{B}}(|A_i|) \leq \mathbb{S}_{\mathcal{B}}(n)$ et,

$$|\mathcal{C}(x_{1n})| \leq \sum_{i=1}^N \mathbb{S}_{\mathcal{B}}(|A_i|) \leq \mathbb{S}_{\mathcal{A}}(n) \mathbb{S}_{\mathcal{B}}(n).$$

(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 $\dim\text{-VC}(\mathcal{A})$ de la classe $\mathcal{A}$ d'ensembles boréliens est le plus grand entier n tel que

$$\mathbb{S}_{\mathcal{A}}(n) = 2^n.$$

Si $\mathbb{S}_{\mathcal{A}}(n) = 2^n$ pour tout n , nous posons $\dim\text{-VC}(\mathcal{A}) = \infty$.

Clairement, si $\mathbb{S}_{\mathcal{A}}(n) < 2^n$ pour certains n , alors pour tout $m > n$, $\mathbb{S}_{\mathcal{A}}(m) < 2^m$ (voir Théorème III-3.8-(i)), et donc la dimension $\dim\text{-VC}$ est toujours bien définie.

Exemple III-3.10 (Dim-VC de la classe des hyperplans linéaires). Considérons la classe $\mathcal{L}_d$ des sous-ensembles de $\mathbb{R}^d$ de la forme $\{x \in \mathbb{R}^d : a^T x \geq 0\}$ pour tout $a \in \mathbb{R}^d$. Nous allons montrer que $V \leq d$. Considérons tout d'abord l'ensemble $S = \{e_1, \dots, e_d\}$ des vecteurs canoniques de $\mathbb{R}^d$: pour $i \in \{1, \dots, d\}$, $e_{i,i} = 1$ et $e_{i,j} = 0$, si $1 \leq j \neq i \leq d$. Nous allons montrer que cet ensemble est pulvérisé par $\mathcal{L}_d$. En effet, pour tout $(y_1, \dots, y_d) \in \{-1, +1\}^d$, considérons l'ensemble

$$A_{y_1, \dots, y_d} = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^d y_i e_i^T x \geq 0 \right\}.$$

Un point e_i appartient à l'intersection $A_{y_1, \dots, y_d} \cap S$ si et seulement si $y_i = +1$. Par conséquent,

$$N_{\mathcal{L}_d}(e_1, \dots, e_d) \geq |\{S \cap A_{y_1, \dots, y_d} : (y_1, \dots, y_d) \in \{-1, +1\}^d\}| = 2^d.$$

Supposons maintenant qu'il existe un ensemble $\{x_1, \dots, x_{d+1}\}$ de $d+1$ pulvérisé par $\mathcal{L}_d$. Ceci implique que pour les 2^{d+1} sous ensembles possibles $\{S_k\}_{k=1}^{2^{d+1}}$ de $\{x_1, \dots, x_{d+1}\}$, il est possible de trouver un vecteur $\{w_k\}_{k=1}^{2^{d+1}}$ tels que

$$S_k = \{x_1, \dots, x_{d+1}\} \cap \left\{ x \in \mathbb{R}^d : w_k^T x \geq 0 \right\}.$$

Considérons la matrice obtenue en concaténant les sorties réelles de tous ces classifieurs

$$\mathbf{H} = \begin{bmatrix} w_1^T x_1 & w_2^T x_2 & \dots & w_{2^{d+1}}^T x_1 \\ \vdots & \ddots & & \vdots \\ w_1^T x_{d+1} & \dots & & w_{2^{d+1}}^T x_{d+1} \end{bmatrix} = \mathbf{XW}, \quad (\text{III-3.4})$$

où nous avons introduit $\mathbf{X} = [x_1, \dots, x_{d+1}]^T$ et $\mathbf{W} = [w_1, \dots, w_{2^{d+1}}]$. Considérons la matrice $\mathbf{G} = \text{signe}(\mathbf{H})$ où $\mathbf{G}_{i,j} = \text{signe}(\mathbf{H}_{i,j})$, $i \in \{1, \dots, d+1\}$, $j \in \{1, \dots, 2^{d+1}\}$. 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 $\mathbf{H} = \mathbf{XW}$, nous avons

$$\text{Rang}(\mathbf{H}) \leq \min(\text{Rang}(\mathbf{X}), \text{Rang}(\mathbf{W})) \leq d$$

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 $\dim\text{-VC}(\mathcal{L}_d) \leq d$ et comme nous avons déjà établi que $\dim\text{-VC}(\mathcal{L}_d) \geq d$, nous en déduisons $\dim\text{-VC}(\mathcal{L}_d) = d$. ◇

Exemple III-3.11 (Dimension VC de la classe des hyperplans affines). Considérons maintenant la classe $\mathcal{L}_f^a$ des sous-ensembles $\{x \in \mathbb{R}^d : a^T x + b \geq 0\}$ 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 $\mathcal{L}_d \subset \mathcal{L}_d^a$ et donc $d = \dim\text{-VC}(\mathcal{L}_d) \leq \dim\text{-VC}(\mathcal{L}_d^a)$. En fait, comme nous allons le montrer ci-dessous, $\dim\text{-VC}(\mathcal{L}_d^a) = d + 1$.

Nous établissons tout d'abord que $\dim\text{-VC}(\mathcal{L}_d^a) \geq d + 1$. Il suffit pour cela de construire un ensemble de $d + 1$ éléments de $\mathbb{R}^d$ qui soit pulvérisé par $\mathcal{L}_d^a$. Considérons à cet effet l'ensemble $S = \{e_1, \dots, e_d\}$ des vecteurs de la base canonique de $\mathbb{R}^d$ auquel nous adjoignons l'origine 0 de l'espace affine. $(y_1, \dots, y_{d+1}) \in \{-1, +1\}^{d+1}$, considérons l'ensemble

$$B_{y_1, \dots, y_{d+1}} = \left\{ x \in \mathbb{R}^d : f_y(x) := \sum_{i=1}^d (y_i - y_{d+1}) e_i^T x + y_{d+1} \geq 0 \right\}.$$

Notons que $f_{y_1, \dots, y_{d+1}}(e_j) = y_j$ pour $j \in \{1, \dots, d\}$ et que $f_{y_1, \dots, y_{d+1}}(0) = y_{d+1}$, nous avons donc

$$N_{S \cup \{0\}}(\mathcal{L}_d^a) \geq |\{(S \cup \{0\}) \cap B_{y_1, \dots, y_{d+1}} : (y_1, \dots, y_{d+1}) \in \{-1, +1\}^{d+1}\}| \geq 2^{d+1}.$$

La borne supérieure de la dimension VC est une conséquence directe du fait qu'un hyperplan affine est un hyperplan linéaire d'un espace $\mathbb{R}^{d+1}$. $\diamond$

Exemple III-3.12 (Dimension VC de la classe des hyper-rectangles). Nous nous intéressons maintenant à la classe $\mathcal{R}_d$ de tous les hyper-rectangles de $\mathbb{R}^d$ dont les côtés sont parallèles aux axes de coordonnées.

$$\mathcal{R}_d = \{[a_1, b_1] \times \dots \times [a_d, b_d] : -\infty < a_i \leq b_i < \infty, i \in \{1, \dots, d\}\}. \quad (\text{III-3.5})$$

Nous allons tout d'abord montrer que $\mathbb{S}_{\mathcal{R}_d}(2d) \geq 2^{2d}$. Nous notons comme précédemment l'ensemble $S = \{e_1, -e_1, e_2, -e_2, \dots, e_d, -e_d\}$ 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 $\mathcal{R}_d$. Le cardinal de cet ensemble est égal à 2^{2d} et il est élémentaire d'établir que cet ensemble pulvérise S . Par conséquent, $\mathbb{S}_{\mathcal{R}_d}(2d) \geq 2^{2d}$.

Considérons maintenant un ensemble quelconque de $2d + 1$ éléments de $\mathbb{R}^d$, $S = \{x_1, \dots, x_{2d+1}\}$. Pour $i \in \{1, \dots, d + 1\}$ et $j \in \{1, \dots, d\}$, nous notons $x_{i,j}$ la j -ème coordonnée du vecteur x_i . Nous notons

$$k_1 = \arg \min(x_{i,1}, i \in \{1, \dots, d + 1\}) \quad \text{et} \quad \ell_1 = \arg \max(x_{i,1}, i \in \{1, \dots, d + 1\} \setminus \{k_1\})$$

et par récurrence, pour $j \in \{2, \dots, d\}$

$$\begin{aligned} k_j &= \arg \min(x_{i,j}, i \in \{1, \dots, d + 1\} \setminus \{k_1, \ell_1, \dots, k_{j-1}, \ell_{j-1}\}) \\ \ell_j &= \arg \max(x_{i,j}, i \in \{1, \dots, d + 1\} \setminus \{k_1, \ell_1, \dots, k_{j-1}, \ell_{j-1}, k_j\}) \end{aligned}$$

Finalement, nous appelons $\{k_{d+1}\} = \{1, \dots, d + 1\} \setminus \{k_1, \ell_1, \dots, k_d, \ell_d\}$. Nous notons $y_i = x_{k_i}$ et $z_i = x_{\ell_i}$ pour $i \in \{1, \dots, d\}$. Par construction, pour $i \in \{1, \dots, d\}$,

$$y_{i,i} \leq x_{k_{d+1},i} \leq z_{i,i}.$$

ou, de façon équivalente, $x_{k_{d+1}} \in [y_{i,1}, z_{i,1}] \times \dots \times [y_{i,d}, z_{i,d}]$. Il n'est donc pas possible de construire un hyper-rectangle contenant les $2d$ points $\{y_1, z_1, \dots, y_d, z_d\}$ et qui ne contienne pas $x_{k_{d+1}}$. On en déduit que : $\mathbb{S}_{\mathcal{R}_d}(2d + 1) < 2^{2d+1}$. Par conséquent, $\dim\text{-VC}(\mathcal{R}_d) = 2d$. $\diamond$

Comme le montre le résultat de base suivant, la dimension $\dim\text{-VC}$ fournit une borne utile pour le coefficient d'éclatement d'une classe.

Théorème III-3.13 (Lemme de Sauer). Soit $\mathcal{A}$ une classe d'ensembles boréliens de dimension $\dim\text{-VC}$ $V = \dim\text{-VC}(\mathcal{A}) < \infty$. Alors, pour tout n ,

$$\mathbb{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^V \binom{n}{i}.$$

Démonstration. On dit que $B \subset \{0, 1\}^n$ pulvérise $S = \{s_1, \dots, s_m\} \subset \{1, \dots, n\}$ 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 $(x_1, \dots, x_n)$ tels que $|\mathcal{A}(x_{1n})| = \mathbb{S}_{\mathcal{A}}(n)$, et on pose $B_0 = \mathcal{A}(x_{1n})$.

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 $\mathbb{S}_{\mathcal{A}}(m) = 2^m$.

On va maintenant effectuer une transformation de l'ensemble B_0 qui préserve son cardinal. Pour $b = (b_1, \dots, b_n) \in B_0$, 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 $S = \{s_1, \dots, s_m\} \subset \{1, n\}$ 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 $c = (c_1, \dots, c_n)$, 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 $\mathcal{A}$ une classe d'ensembles boréliens de dimension $\dim\text{-VC } V < \infty$. Alors pour tout $n \in \mathbb{N}$,

$$\mathbb{S}_{\mathcal{A}}(n) \leq (n+1)^V,$$

et pour tout $n \geq V$,

$$\mathbb{S}_{\mathcal{A}}(n) \leq \left(\frac{ne}{V}\right)^V.$$

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

$$\left(\frac{V}{n}\right)^V \sum_{i=0}^V \binom{n}{i} \leq \sum_{i=0}^V \left(\frac{V}{n}\right)^i \binom{n}{i} \leq \sum_{i=0}^n \left(\frac{V}{n}\right)^i \binom{n}{i} = \left(1 + \frac{V}{n}\right)^n \leq e^V.$$

□

L'inégalité de Vapnik-Chervonenkis (Théorème III-3.6) montre que si $\mathcal{A}$ est une classe d'ensembles boréliens de dimension $V = \dim\text{-VC}(\mathcal{A})$, alors

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq 2 \sqrt{\frac{V \log(n+1) + \log 2}{n}},$$

et donc, si la classe d'ensembles boréliens $\mathcal{A}$ a une dimension $\dim\text{-VC}$ finie, $\mathbb{E}[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|]$ converge vers 0 à un taux $O(\sqrt{\log n}/n)$.

Lemme III-3.15. *Soit $\mathcal{G}$ 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*

$$\mathcal{A} = \left\{ \left\{ x \in \mathbb{R}^d : g(x) \geq 0 \right\} : g \in \mathcal{G} \right\}$$

est inférieure à m : $\dim\text{-VC}(\mathcal{A}) \leq m$.

Démonstration. Il suffit de montrer qu'un ensemble de cardinal $m+1$ ne peut être pulvérisé par les ensembles de $\mathcal{A}$. Soit $x_{1:m+1} = (x_1, \dots, x_{m+1}) \in \mathbb{R}^d$. On définit $\phi : \mathcal{G} \rightarrow \mathbb{R}^{m+1}$ par :

$$\phi(g) = (g(x_1), \dots, g(x_{m+1}))$$

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 $a = (a_1, \dots, a_{m+1})$ tel que, pour tout $g \in \mathcal{G}$

$$a_1 g(x_1) + \dots + a_{m+1} g(x_{m+1}) = 0$$

Si tous les coefficients $a_1, \dots, a_{m+1}$ sont positifs, alors il est clair que les ensembles de $\mathcal{A}$ ne peuvent pulvériser $\{x_1, x_{m+1}\}$, sinon on choisit $g \in \mathcal{G}$ telle que $g(x_i) < 0$ pour tout $i \in \{1, \dots, m+1\}$, et on aurait alors $a_1 g(x_1) + \dots + a_{m+1} g(x_{m+1}) < 0$.

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 $g \in \mathcal{G}$ qui soit telle que $\{x \in \mathbb{R}^d : g(x) > 0\}$ 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. $\square$

Corollaire III-3.16. (i) *Soit $\mathcal{A}$ la classe de toutes les boules de $\mathbb{R}^d$, i.e. les ensembles de la forme*

$$\left\{ x \in \mathbb{R}^d : \|x - a\|^2 \leq b \right\} \quad \text{avec } a \in \mathbb{R}^d \text{ et } b \in \mathbb{R}^d.$$

Alors $\dim\text{-VC}(\mathcal{A}) \leq d+2$.

(ii) *Soit $\mathcal{A}$ la classe de tous les ellipsoïdes de $\mathbb{R}^d$, i.e. les ensembles de la forme*

$$\left\{ x \in \mathbb{R}^d : x^T \Sigma^{-1} x \leq 1 \right\}, \quad \text{où } \Sigma \text{ est une matrice définie positive.}$$

Alors $\dim\text{-VC}(\mathcal{A}) \leq d(d+1)/2 + 1$.

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 : $\mathcal{C}$ est une classe de classifieurs contenant des fonctions de décision de la forme $g : \mathbb{R}^d \rightarrow \{0, 1\}$. Nous utilisons les données d'apprentissage $(X_1, Y_1), \dots, (X_n, Y_n)$ pour calculer l'erreur empirique $\hat{R}_n(g)$ pour tout $g \in \mathcal{C}$. Un minimiseur du risque empirique g_n^* est une règle de classification vérifiant

$$\hat{R}_n(\hat{g}_n^*) \leq \hat{R}_n(g) \quad \text{for all } g \in \mathcal{C}.$$

Nous notons par $L_{\mathcal{C}}$ le minimum du risque sur la classe $\mathcal{C}$, c'est-à-dire

$$L_{\mathcal{C}} = \inf_{g \in \mathcal{C}} R(g) .$$

(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

$$L(\hat{g}_n^*) - L_{\mathcal{C}} \leq 2 \sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)| .$$

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 $\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|$ est de la forme $\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$, où le rôle de la classe des ensembles $\mathcal{A}$ est maintenant joué par la classe $\tilde{\mathcal{A}}$ des ensembles d'erreurs, définis par

$$\left\{ \left\{ (x, y) \in \mathbb{R}^d \times \{0, 1\} : g(x) \neq y \right\} : g \in \mathcal{C} \right\} .$$

L'inégalité de Vapnik-Chervonenkis borne $\sup_{g \in \mathcal{C}} |\hat{R}_n(g) - R(g)|$ en fonction des coefficients d'éclatement (ou de la dimension dim-VC) de la classe $\tilde{\mathcal{A}}$ des ensembles d'erreurs.

Au lieu des ensembles d'erreurs, il est plus pratique de travailler avec des classes d'ensembles de la forme

$$\left\{ \left\{ x \in \mathbb{R}^d : g(x) = 1 \right\} : g \in \mathcal{C} \right\} .$$

Nous désignons la classe d'ensembles ci-dessus par $\mathcal{A}$. Le fait simple suivant montre que les classes $\tilde{\mathcal{A}}$ et $\mathcal{A}$ sont équivalentes d'un point de vue combinatoire.

Lemme III-3.17. *Pour chaque $n \in \mathbb{N}$ nous avons $\mathbb{S}_{\tilde{\mathcal{A}}}(n) = \mathbb{S}_{\mathcal{A}}(n)$, et donc les dimensions correspondantes dim-VC sont aussi égales : $V_{\tilde{\mathcal{A}}} = V_{\mathcal{A}}$.*

Démonstration. Soit N un entier positif. Nous allons montrer que pour tout ensemble de n points $\{(x_1, y_1), \dots, (x_n, y_n)\}$ de $\mathbb{R}^d \times \{0, 1\}$, si N ensembles de $\tilde{\mathcal{A}}$ sélectionnent N sous-ensembles de n paires, alors il existe N ensembles de $\mathcal{A}$ qui sélectionnent N sous-ensembles de $\{x_1, \dots, x_n\}$, et vice versa.

Plus précisément, considérons n paires $(x_1, 0), \dots, (x_m, 0), (x_{m+1}, 1), \dots, (x_n, 1)$. 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 $A \in \mathcal{A}$, l'ensemble correspondant

$$\bar{A} = A \times \{0\} \cup A^c \times \{1\} \in \tilde{\mathcal{A}}$$

choisit les paires $(x_1, 0), \dots, (x_k, 0), (x_{m+1}, 1), \dots, (x_{m+l}, 1)$, c'est-à-dire que l'ensemble de ces paires est l'intersection de $\tilde{\mathcal{A}}$ 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 $\{x_1, \dots, x_n\}$ le sous-ensemble $\{x_1, \dots, x_k, x_{m+l+1}, \dots, x_n\}$, et les deux sous-ensembles sont associés de façon biunivoque. Ceci prouve que $\mathbb{S}_{\tilde{\mathcal{A}}}(n) \leq \mathbb{S}_{\mathcal{A}}(n)$. Pour prouver la réciproque, remarquez que si A choisit un sous-ensemble de k points $\{x_1, \dots, x_k\} \subset \{x_1, \dots, x_n\}$, alors l'ensemble correspondant $\bar{A} \in \tilde{\mathcal{A}}$ choisit les paires avec les mêmes indices $\{(x_1, 0), \dots, (x_k, 0)\}$.

L'égalité des dimensions dim-VC découle de l'égalité des coefficients d'éclatement pour tout n . $\square$

Nous noterons dans la suite de l'exposé la valeur commune de $\mathbb{S}_{\tilde{\mathcal{A}}}(n)$ et $\mathbb{S}_{\mathcal{A}}(n)$ par $\mathbb{S}_{\mathcal{C}}(n)$, que nous appellerons le coefficient de pulvérisation de n -points de la famille de règles de classification $\mathcal{C}$. Il s'agit simplement du nombre maximum de façons différentes dont les points n peuvent être classés par les classifieurs dans la famille $\mathcal{C}$. De même, nous appellerons $V_{\tilde{\mathcal{A}}} = V_{\mathcal{A}}$ la dimension dim-VC de la famille de classifieurs $\mathcal{C}$, que nous noterons $V_{\mathcal{C}}$.

Nous disposons de tous les éléments nécessaires pour présenter les principales performances du minimiseur du risque empirique.

Corollaire III-3.18.

$$\mathbb{E}[L(\hat{g}_n^*) - L_{\mathcal{C}}] \leq \sqrt{2 \frac{\log(2\mathbb{S}_{\mathcal{C}}(2n))}{n}} .$$

Nous pouvons aisément déduire de ce résultat une borne pour $\mathbb{P}(L(\hat{g}_n^*) - L_{\mathcal{C}} > \varepsilon)$ en utilisant le Théorème III-3.2.

Quatrième partie

Outils probabilistes

Chapitre IV-1

Éléments de la théorie de la mesure et de l'intégration

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 [Garet and Kurtzmann, 2011, Gallouët and Herbin, 2013].

IV-1.1 Tribus et Mesurabilité

IV-1.1.1 Tribus

Définition IV-1.1 (Algèbre, Tribu). Soit E un ensemble et $\mathcal{E}$ un ensemble de parties de E . On dit que $\mathcal{E}$ est une algèbre sur E si elle vérifie les propriétés suivantes :

- $E \in \mathcal{E}$,
- $\mathcal{E}$ est stable par passage au complémentaire : pour tout $A \in \mathcal{E}$, on a $A^c := E \setminus A \in \mathcal{E}$,
- $\mathcal{E}$ est stable par réunion finie : pour A, B éléments de $\mathcal{E}$, $A \cup B \in \mathcal{E}$.

On dit que $\mathcal{E}$ est une tribu sur E (ou σ -algèbre) si elle vérifie les propriétés suivantes :

- $E \in \mathcal{E}$,
- $\mathcal{E}$ est stable par passage au complémentaire : pour tout $A \in \mathcal{E}$, on a $A^c \in \mathcal{E}$,
- $\mathcal{E}$ est stable par réunion dénombrable : pour toute suite $\{A_n, n \in \mathbb{N}\}$ d'éléments de $\mathcal{E}$, $\bigcup_{n=0}^{\infty} A_n \in \mathcal{E}$.

Remarquons que si $\mathcal{E}$ est une tribu sur E , alors $\emptyset \in \mathcal{E}$ 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 $\mathcal{E}$ alors $\bigcap_{n=0}^{\infty} A_n \in \mathcal{E}$ puisque cet ensemble est le complémentaire de $\bigcup_{n=0}^{\infty} A_n^c$. Par suite, $\mathcal{E}$ 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 $\mathcal{E}$ 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 IV-1.1 et des propriétés de stabilité par intersection et réunion finies ou dénombrables, on déduit les propriétés suivantes : Si $\mathcal{E}$ est une tribu, alors

- (i) Si $A, B \in \mathcal{E}$, alors $A \setminus B := A \cap B^c \in \mathcal{E}$,
- (ii) Si $A, B \in \mathcal{E}$ alors $A \Delta B := (A \setminus B) \cup (B \setminus A) \in \mathcal{E}$.

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

$$\liminf_{n \rightarrow \infty} A_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k, \quad \limsup_{n \rightarrow \infty} A_n := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

$\liminf_{n \rightarrow \infty} A_n$ est l'ensemble des éléments de E qui appartiennent à tous les A_n à partir d'un certain rang, et $\limsup_{n \rightarrow \infty} A_n$ 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 $\mathcal{E}$, alors

$$\liminf_n A_n \in \mathcal{E}, \quad \limsup_n A_n \in \mathcal{E}.$$

Il n'est souvent pas possible de décrire les éléments d'une tribu. On remarque que l'ensemble des parties de E , noté $\mathcal{P}(E)$, 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 $\mathcal{C} \subset \mathcal{P}(E)$, on peut considérer la plus petite tribu contenant $\mathcal{C}$.

Définition IV-1.2 (Tribu engendrée). Soit $\mathcal{C} \in \mathcal{P}(E)$. La tribu engendrée par $\mathcal{C}$ (sur E) est l'intersection de toutes les tribus sur E contenant $\mathcal{C}$. Cette tribu se note $\sigma(\mathcal{C})$.

Définition IV-1.3 (Tribu borélienne). Soit E un espace topologique et soit $\mathcal{O}$ la classe des ouverts de E . La tribu $\sigma(\mathcal{O})$ s'appelle la tribu borélienne de E et se note $\mathcal{B}(E)$.

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 $E = \mathbb{R}^d$ muni de sa topologie usuelle (associée à la distance euclidienne). Dans ce cas, $\mathcal{B}(\mathbb{R}^d)$ 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 $\mathcal{B}(\mathbb{R})$, on considérera

$$\begin{aligned} \mathcal{B}(\mathbb{R}^+) &:= \{A \in \mathcal{B}(\mathbb{R}), A \subset \mathbb{R}^+\}, \\ \mathcal{B}(\overline{\mathbb{R}}) &:= \sigma(\mathcal{B}(\mathbb{R}), \{+\infty\}, \{-\infty\}), \\ \mathcal{B}(\overline{\mathbb{R}^+}) &:= \sigma(\mathcal{B}(\mathbb{R}^+), \{+\infty\}). \end{aligned}$$

On prolonge à $\overline{\mathbb{R}^+}$ les relations d'ordre, l'addition et la multiplication de $\mathbb{R}^+$ en posant

$$\begin{aligned} &\text{pour tout } a \in \mathbb{R}^+ : a < +\infty. \\ &\text{pour tout } a \in \overline{\mathbb{R}^+} : a + (+\infty) = (+\infty) + a = +\infty. \\ &\text{pour tout } a \in \overline{\mathbb{R}^+} \setminus \{0\} : a(+\infty) = (+\infty)a = +\infty. \\ &0(+\infty) = (+\infty)0 = 0. \end{aligned}$$

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

$$\begin{aligned} &\text{pour tout } x \in \mathbb{R}, -\infty < x < +\infty. \\ &\text{pour tout } x \in]-\infty, +\infty] : x + (+\infty) = (+\infty) + x = +\infty. \\ &\text{pour tout } x \in [-\infty, +\infty[: x + (-\infty) = (-\infty) + x = -\infty. \\ &\text{pour tout } x \in [-\infty, +\infty] : x(+\infty) = +\infty \text{ si } x \geq 0 \text{ et } -\infty \text{ sinon.} \\ &\text{pour tout } x \in [-\infty, +\infty] : x(-\infty) = -\infty \text{ si } x \geq 0 \text{ et } +\infty \text{ sinon.} \end{aligned}$$

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.

IV-1.1.2 Classes monotones (ou λ -systèmes)

Définition IV-1.4 (Classe monotone, λ -système). Soit E un ensemble. Une famille $\mathcal{M}$ d'éléments de $\mathcal{P}(E)$ est appelée une classe monotone (ou un λ -système) sur E si

- (i) $E \in \mathcal{M}$,
- (ii) $A, B \in \mathcal{M}, A \subset B \implies B \setminus A \in \mathcal{M}$,
- (iii) pour toute suite $\{A_n, n \in \mathbb{N}\}$ d'éléments de $\mathcal{M}$ tels que $A_n \subset A_{n+1}$, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{M}$.

Une tribu est une classe monotone. L'intersection d'une famille quelconque de classes monotones est une classe monotone. Pour une famille quelconque $\mathcal{C}$ de sous-ensembles de E , il existe une *plus petite classe monotone contenant $\mathcal{C}$* qui est l'intersection de toutes les classes monotones contenant $\mathcal{C}$.

Lemme IV-1.5. Soit E un ensemble. Si $\mathcal{N}$ est une classe monotone sur E , stable par intersection finie, alors $\mathcal{N}$ est une tribu sur E .

Démonstration. Par définition d'une classe monotone, $E \in \mathcal{N}$. De plus, comme $\mathcal{N}$ est stable par différence, si $A \in \mathcal{N}$ alors $A^c \in \mathcal{N}$. Montrons la stabilité par union dénombrable.

Soit $\{B_k, k \in \mathbb{N}\}$ une suite d'éléments de $\mathcal{N}$; pour tout $n \in \mathbb{N}$, posons $A_n := \bigcup_{k=1}^n B_k \in \mathcal{N}$. Alors la stabilité par intersection finie et par passage au complémentaire impliquant la stabilité par union finie, on a $A_n \in \mathcal{N}$. Donc $\{A_n, n \in \mathbb{N}\}$ est une suite croissante d'éléments de $\mathcal{N}$. Par définition de la classe monotone, on a donc $\bigcup_{n=1}^{\infty} A_n \in \mathcal{N}$ ce qui entraîne $\bigcup_{k=1}^{\infty} B_k \in \mathcal{N}$ puisque $\bigcup_{n=1}^{\infty} A_n = \bigcup_{k=1}^{\infty} B_k \in \mathcal{N}$. $\square$

Théorème IV-1.6. Soit E un ensemble et $\mathcal{M}$ une classe monotone sur E . Soient $\mathcal{C} \subset \mathcal{M}$, une famille d'ensembles stable par intersection finie. Alors $\sigma(\mathcal{C}) \subset \mathcal{M}$.

Démonstration. Soit $\mathcal{S}$ la plus petite classe monotone contenant $\mathcal{C}$. Nous allons montrer que $\mathcal{S}$ est stable par intersection finie; nous conclurons par Lemme IV-1.5 que c'est une tribu, et c'est nécessairement la plus petite tribu sinon cela contredirait le fait que $\mathcal{S}$ est la plus petite classe monotone.

Pour $A \subset E$, on pose $\mathcal{D}_A := \{B \in \mathcal{S}, A \cap B \in \mathcal{S}\}$. Considérons le cas où $A \in \mathcal{S}$; nous allons montrer que dans ce cas, $\mathcal{D}_A$ est une classe monotone puisqu'elle vérifie les propriétés suivantes :

- (i) $E \in \mathcal{D}_A$. En effet, $E \in \mathcal{S}$ et $E \cap A = A \in \mathcal{S}$.
- (ii) si $B_1, B_2 \in \mathcal{D}_A$ avec $B_1 \subset B_2$, alors $B_2 \setminus B_1 \in \mathcal{D}_A$. En effet, on sait que $A \cap B_i \in \mathcal{S}$ et que $(A \cap B_1) \subset (A \cap B_2)$; par propriété des classes monotones, il vient

$$A \cap (B_2 \setminus B_1) = (A \cap B_2) \setminus (A \cap B_1) \in \mathcal{S}.$$

Par suite, $B_2 \setminus B_1 \in \mathcal{D}_A$.

- (iii) Soit une suite croissante $\{B_n, n \in \mathbb{N}\}$, d'éléments de $\mathcal{D}_A$; alors $\bigcup_n B_n \in \mathcal{D}_A$. En effet, la famille d'ensembles $\{A \cap B_n, n \in \mathbb{N}\}$ est aussi croissante, et chaque ensemble est élément de $\mathcal{S}$. Donc, par propriété des classes monotones,

$$A \cap \bigcup_n B_n = \bigcup_n (A \cap B_n) \in \mathcal{S}.$$

Dans le cas où $A \in \mathcal{C}$, alors puisque pour tout $B \in \mathcal{C}$ on a $A \cap B \in \mathcal{C} \subset \mathcal{S}$, il vient que $B \in \mathcal{D}_A$; donc $\mathcal{C} \subset \mathcal{D}_A$.

En combinant ces résultats, nous voyons que pour $A \in \mathcal{C} \subset \mathcal{S}$, la famille $\mathcal{D}_A$ est une classe monotone qui contient $\mathcal{C}$. Par suite, on a $\mathcal{S} \subset \mathcal{D}_A$.

Soit $A \in \mathcal{C}$ et $B \in \mathcal{S}$. Puisque $\mathcal{S} \subset \mathcal{D}_A$, il vient que $B \in \mathcal{D}_A$ et donc que $A \in \mathcal{D}_B$. On en déduit que $\mathcal{C} \subset \mathcal{D}_B$. Mais $\mathcal{D}_B$ étant une classe monotone, alors $\mathcal{S} \subset \mathcal{D}_B$.

Par conséquent, soit $B, B' \in \mathcal{S}$. Vu ce que l'on vient d'établir, $B' \in \mathcal{D}_B$ et donc $B \cap B' \in \mathcal{S}$. Donc $\mathcal{S}$ est stable par intersection finie. $\square$

IV-1.1.3 Applications mesurables

Définition IV-1.7 (Espace mesurable). Un espace mesurable est un couple $(E, \mathcal{E})$ où $\mathcal{E}$ étant une tribu sur E .

Pour une application $f : E_1 \rightarrow E_2$ et pour $A \in \mathcal{E}_2$, on note $f^{-1}(A)$ l'image réciproque par f de A défini par

$$f^{-1}(A) := \{x \in E_1 : f(x) \in A\}.$$

Définition IV-1.8 (Applications mesurables). Soient $(E_1, \mathcal{E}_1)$ et $(E_2, \mathcal{E}_2)$ deux espaces mesurables. Une application f de E_1 dans E_2 est dite mesurable si, pour tout $A \in \mathcal{E}_2$, $f^{-1}(A) \in \mathcal{E}_1$.

Proposition IV-1.9. Soit $f : (E_1, \mathcal{E}_1) \rightarrow (E_2, \mathcal{E}_2)$. Lorsque $\mathcal{E}_2$ est la tribu engendrée par la famille de parties $\mathcal{C} \subset \mathcal{P}(E)$, f est mesurable si et seulement si $f^{-1}(A) \in \mathcal{E}_1$ pour tout $A \in \mathcal{C}$.

Démonstration. L'implication directe est triviale puisque $\mathcal{C} \in \mathcal{E}_2$, et pour l'implication réciproque, il suffit de montrer que l'ensemble $\{A \in \mathcal{E}_2 : f^{-1}(A) \in \mathcal{E}_1\}$ des parties de E_2 est une tribu sur E_2 qui contient $\mathcal{C}$ et elle contient donc en partie ma tribu engendrée $\mathcal{E}_2$. $\square$

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 $(E, \mathcal{E})$ un espace mesurable. Si pour tout $x \in E$, 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

$$\liminf f_n(x) = \lim_n \uparrow \inf_{k \geq n} f_k(x) \quad \limsup f_n(x) = \lim_n \downarrow \sup_{k \geq n} f_k(x); \quad (\text{IV-1.1})$$

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 $f : E \rightarrow \overline{\mathbb{R}}$ soit mesurable, il suffit que, pour tout $a \in \overline{\mathbb{R}}$, $\{f > a\} := \{x : f(x) > a\} \in \mathcal{E}$. 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 $\mathcal{P}(E)$, $\mathbb{1}_{A_n}$ est une application mesurable. Les relations

$$\mathbb{1}_{\bigcap_{n=0}^{\infty} A_n} = \prod_{n=0}^{\infty} \mathbb{1}_{A_n} = \inf_{n \geq 0} \mathbb{1}_{A_n}, \quad \mathbb{1}_{\bigcup_{n=0}^{\infty} A_n} = \sup_{n \geq 0} \mathbb{1}_{A_n},$$

entraînent que

$$\liminf \mathbb{1}_{A_n} = \mathbb{1}_{\liminf A_n}, \quad \limsup \mathbb{1}_{A_n} = \mathbb{1}_{\limsup A_n}.$$

Soient f, g des fonctions numériques mesurables définies sur l'espace mesurable $(E, \mathcal{E})$. Alors $\phi : x \mapsto (f(x), g(x))$ est mesurable de $(E, \mathcal{E})$ dans $\mathbb{R}^2$ puisque pour tout pavé $A \times B$ de $\mathbb{R}^2$, $\phi^{-1}(A \times B) = f^{-1}(A) \cap g^{-1}(B) \in \mathcal{E}$. 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 IV-1.10. *Soit $(E, \mathcal{E})$ un espace mesurable et $f, g : E \rightarrow \overline{\mathbb{R}}$ 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 IV-1.1 ; 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 IV-1.11 (Fonction réelle étagée). *Soit $(E, \mathcal{E})$ un ensemble mesurable. Une application $f : E \rightarrow \mathbb{R}$ est dite étagée si elle s'écrit*

$$f = \sum_{k=1}^p a_k \mathbb{1}_{A_k}, \quad A_k \in \mathcal{E}.$$

Proposition IV-1.12. *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

$$f_n(x) := \sum_{k=0}^{n2^n-1} 2^{-n} k \mathbb{1}_{\{x : 2^{-n} k \leq f(x) < 2^{-n}(k+1)\}} + n \mathbb{1}_{\{x : f(x) \geq n\}}. \quad (\text{IV-1.2})$$

□

Dans la suite, nous noterons :

$[\mathcal{E}]$ l'ensemble des fonctions réelles mesurables,
 $b\mathcal{E}$ l'ensemble des fonctions réelles mesurables bornées,
 $\mathcal{E}^+$ l'ensemble des fonctions mesurables à valeurs $\overline{\mathbb{R}^+}$,
 $e\mathcal{E}^+$ l'ensemble des fonctions étagées positives.

IV-1.1.4 Mesurabilité pour une tribu engendrée par des applications

Théorème IV-1.13. Soient $\mathcal{H}$ un espace vectoriel de fonctions numériques bornées définies sur E vérifiant

$$1 \in \mathcal{H}, f_n \in \mathcal{H} \text{ et } 0 \leq f_n \uparrow f \text{ bornée} \Rightarrow f \in \mathcal{H}. \quad (\text{IV-1.3})$$

Soit $\mathcal{C}$ un ensemble de parties de E stable par intersection finie et tel que $A \in \mathcal{C} \Rightarrow \mathbb{1}_A \in \mathcal{H}$.
 Alors $\mathcal{H}$ contient toutes les fonctions $\sigma(\mathcal{C})$ -mesurables bornées.

Démonstration. Soit $\mathcal{M} := \{A \subset E : \mathbb{1}_A \in \mathcal{H}\}$. Alors $\mathcal{M}$ est une classe monotone sur E ; on a en effet $\mathbb{1}_E = 1 \in \mathcal{H}$; $\mathbb{1}_{B \setminus A} = \mathbb{1}_B - \mathbb{1}_A$ et $\mathcal{H}$ est un espace vectoriel; pour une suite croissante $\{A_n, n \in \mathbb{N}\}$ d'éléments de $\mathcal{M}$, $\mathbb{1}_{\bigcup_n A_n} = \mathbb{1}_{\lim A_n} = \lim \uparrow \mathbb{1}_{A_n} \in \mathcal{H}$.

Puisque $\mathcal{C} \subset \mathcal{M}$, on peut appliquer le Théorème IV-1.6 et l'on a $\sigma(\mathcal{C}) \subset \mathcal{M}$.

Soit $f : E \rightarrow \mathbb{R}$ une fonction $\sigma(\mathcal{C})$ -mesurable, étagée. Donc c'est une combinaison linéaire (finie) d'indicatrices d'éléments de $\sigma(\mathcal{C})$ et donc indicatrices d'éléments de $\mathcal{M}$. Par suite, $f \in \mathcal{H}$. Considérons maintenant le cas où f est une fonction positive bornée $\sigma(\mathcal{C})$ -mesurable; alors c'est la limite croissante de fonctions positives étagées $\sigma(\mathcal{C})$ -mesurable (et donc de fonctions dans $\mathcal{H}$) ce qui entraîne que $f \in \mathcal{H}$. Enfin, pour toute fonction f bornée $\sigma(\mathcal{C})$ -mesurable, on a $f \in \mathcal{H}$. $\square$

Définition IV-1.14 (Tribu engendrée par une application). Soit X une application de Ω dans un espace mesurable $(E, \mathcal{E})$. 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 $X_i : \Omega \rightarrow (E_i, \mathcal{E}_i)$, 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

$$\sigma(X) = \{X^{-1}(A), A \in \mathcal{E}\}.$$

Soit

$$\mathcal{C} := \left\{ A \subset \Omega, A = \bigcap_{k=1}^n X_{i_k}^{-1}(\Gamma_k), \Gamma_k \in \mathcal{E}_{i_k}, i_1, \dots, i_n \in I \right\}.$$

$\mathcal{C}$ est stable par intersection finie et l'on a $\sigma(\mathcal{C}) = \sigma(X_i, i \in I)$.

Proposition IV-1.15. Pour tout $i \in I$, soit $(E_i, \mathcal{E}_i)$ un espace mesurable et $X_i : \Omega \rightarrow E_i$ une application. Une application $\Phi : (A, \mathcal{A}) \rightarrow (\Omega, \sigma(X_i, i \in I))$ est mesurable si et seulement si pour tout $i \in I$, $X_i \circ \Phi$ est mesurable de $(A, \mathcal{A})$ dans $(E_i, \mathcal{E}_i)$.

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 IV-1.9) de vérifier que $\Phi^{-1}(C) \in \mathcal{A}$ pour tout C de la forme $X_i^{-1}(\Gamma_i)$ où $\Gamma_i \in \mathcal{E}_i$. 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 $\mathcal{E}_i$ par hypothèse. $\square$

Le Théorème IV-1.13 implique alors :

Corollaire IV-1.16. Soient $\mathcal{H}$ un espace vectoriel de fonctions numériques bornées définies sur Ω . Pour tout $i \in I$, soit $(E_i, \mathcal{E}_i)$ un espace mesurable et une application $X_i : \Omega \rightarrow E_i$. On suppose que $\mathcal{H}$ vérifie (IV-1.3) et que, pour tout $i_1, \dots, i_n \in I$ et tout $\Gamma_k \in \mathcal{E}_{i_k}$,

$$\prod_{k=1}^n \mathbb{1}_{\Gamma_k} \circ X_{i_k} \in \mathcal{H}.$$

Alors $\mathcal{H}$ contient toutes les fonctions $\sigma(X_i, i \in I)$ -mesurables bornées.

On suppose que, pour tout $i \in I$, $(E_i, \mathcal{E}_i) = (E, \mathcal{E})$. On note $F := E^{\mathbb{N}}$. Pour $x = \{x_n, n \in \mathbb{N}\} \in F$, on définit $\xi_n : F \rightarrow E$ par $\xi_n(x) := x_n$ et on pose $\mathcal{F} = \sigma(\xi_n, n \in \mathbb{N})$.

Corollaire IV-1.17. Soient, pour tout $i \in I$, $X_i : \Omega \rightarrow (E, \mathcal{E})$ et $Y : \Omega \rightarrow \mathbb{R}$ (resp. $\Omega \rightarrow \overline{\mathbb{R}^+}$). 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 $h : F \rightarrow \mathbb{R}$ (resp. $h : F \rightarrow \mathbb{R}^+$) $\mathcal{F}$ -mesurable telle que $Y = h(X_{i_1}, \dots, X_{i_n}, \dots)$.

Démonstration. Vu Proposition IV-1.15, si $h \in b\mathcal{F}$, $h(X_{i_1}, \dots, X_{i_n}, \dots) \in b\sigma(X_i, i \in I)$. Dans l'autre sens, soit

$$\mathcal{H} = \{Z : \Omega \rightarrow \mathbb{R}; Z = h(X_{i_1}, \dots, X_{i_n}, \dots), i_k \in I, h \in b\mathcal{F}\}.$$

On vérifie (assez) facilement que $\mathcal{H}$ est un espace vectoriel de fonctions numériques bornées vérifiant (IV-1.3) et contenant $\prod_{k=1}^n \mathbb{1}_{\Gamma_k}(X_{i_k})$. Appliquant le Corollaire IV-1.16, $\mathcal{H} \supset b\sigma(X_i, i \in I)$. $\square$

Corollaire IV-1.18. Soit $\mathcal{H}$ un espace vectoriel de fonctions numériques bornées définies sur $\mathbb{R}^d$. On suppose que $\mathcal{H}$ vérifie (IV-1.3) et contient toutes les fonctions continues à support compact. Alors $\mathcal{H} \supset b\mathcal{B}(\mathbb{R}^d)$.

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 $f_n(x) := 1 \wedge nd(x, U^c)$ où $d(x, U^c)$ désigne la distance de x au fermé U^c . Donc $\mathbb{1}_U \in \mathcal{H}$ et on applique le Théorème IV-1.13. $\square$

Enfin combinant les Corollaires IV-1.16 and IV-1.18, on obtient, notant $C_k(\mathbb{R}^d)$ l'espace des fonctions continues à support compact sur $\mathbb{R}^d$,

Corollaire IV-1.19. Soient $\mathcal{H}$ 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 $\mathcal{H}$ vérifie (IV-1.3) et que, pour tout $i_1, \dots, i_n \in I$ et toute fonction $f_j \in C_k(\mathbb{R}^d)$,

$$\prod_{j=1}^n f_j \circ X_{i_j} \in \mathcal{H}.$$

Alors $\mathcal{H}$ contient toutes les fonctions $\sigma(X_i, i \in I)$ -mesurables bornées.

IV-1.1.5 Espaces produits

Étant donné deux ensembles arbitraires E_1 et E_2 , on note

$$E_1 \times E_2 := \{(e_1, e_2) : e_1 \in E_1, e_2 \in E_2\}.$$

L'application $X_i : E_1 \times E_2 \rightarrow E_i$ ($i = 1, 2$) qui à (e_1, e_2) fait correspondre e_i s'appelle l'*application coordonnée i*.

Si A est un sous-ensemble arbitraire de $E_1 \times E_2$, on note

$$A_{e_1} := \{e_2 \in E_2 : (e_1, e_2) \in A\},$$

dite la *section* de A en e_1 . Pour tout $e_1 \in E_1$ (fixé), l'application $A \mapsto A_{e_1}$ de $\mathcal{P}(E_1 \times E_2) \rightarrow \mathcal{P}(E_2)$ est un homomorphisme pour les opérations de réunion, d'intersection et de complémentation : en effet, si $\{A^\alpha\}$ est une famille d'éléments de $\mathcal{P}(E_1 \times E_2)$, on a

$$\left(\bigcup_{\alpha} A^\alpha \right)_{e_1} = \bigcup_{\alpha} A_{e_1}^\alpha, \quad \left(\bigcap_{\alpha} A^\alpha \right)_{e_1} = \bigcap_{\alpha} A_{e_1}^\alpha, \quad (A^c)_{e_1} = (A_{e_1})^c.$$

Si f est une application arbitraire de $E_1 \times E_2$ 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 $f \mapsto f_{e_1}$ (e_1 fixé) préserve les opérations habituelles sur les fonctions (additions, multiplication, ratio), y compris la convergence simple.

Un *pavé* de $E_1 \times E_2$ est un sous-ensemble de la forme

$$A_1 \times A_2 := \{(e_1, e_2) \in E_1 \times E_2, e_1 \in A_1, e_2 \in A_2\};$$

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 IV-1.20 (Pavé mesurable, Tribu produit, Produit d'espaces mesurables). Soient $(E_1, \mathcal{E}_1)$ et $(E_2, \mathcal{E}_2)$ deux espaces mesurables.

Un pavé mesurable $A_1 \times A_2$ de $(E_1, \mathcal{A}_1) \times (E_2, \mathcal{A}_2)$ est un pavé de $E_1 \times E_2$ tel que $A_1 \in \mathcal{E}_1$ et $A_2 \in \mathcal{E}_2$.

On appelle *tribu produit* de $\mathcal{E}_1$ et de $\mathcal{E}_2$, la plus petite tribu sur $E_1 \times E_2$ qui contient les pavés mesurables de $(E_1, \mathcal{A}_1) \times (E_2, \mathcal{A}_2)$. Elle est notée $\mathcal{E}_1 \otimes \mathcal{E}_2$.

L'espace mesurable $(E_1 \times E_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ est appelé le *produit* des espaces mesurables $(E_1, \mathcal{A}_1)$ et $(E_2, \mathcal{A}_2)$.

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 $E_1 \times E_2$ 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 :

$$(A_1 \times A_2)^c = (A_1^c \times E_2) \cup (E_1 \times A_2^c) = (A_1^c \times A_2^c) \cup (A_1 \times A_2^c) \cup (A_1^c \times A_2).$$

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 IV-1.21. $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) \otimes \dots \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^d)$

Démonstration. On pose $\mathcal{E}^{\otimes d} := \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) \otimes \dots \otimes \mathcal{B}(\mathbb{R})$. Si U un ouvert de $\mathbb{R}^d$, $U = \bigcup_n P_n$ 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 $(n_1, \dots, n_d) \in \mathbb{N}^d$ et sur tous les centres $(r_1, \dots, r_d) \in \mathbb{Q}^d$ tels que le pavé est inclus dans U . Donc $U \in \mathcal{E}^{\otimes d}$ et $\mathcal{B}(\mathbb{R}^d) \subset \mathcal{E}^{\otimes d}$.

Réciproquement, soient $X_1, X_2, \dots, X_d$ les applications coordonnées de $\mathbb{R}^d$ sur $\mathbb{R}$. Les X_k sont des fonctions continues donc mesurables de $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ d'où $\mathcal{E}^{\otimes d} = \sigma(X_1, \dots, X_d) \subset \mathcal{B}(\mathbb{R}^d)$. $\square$

Nous terminons cette section par des propriétés sur la mesurabilité des sections qui seront utiles notamment pour justifier le théorème de Fubini (voir Théorème IV-1.49 et Théorème IV-1.49).

Proposition IV-1.22. Soit e_1 fixé. La section A_{e_1} de toute partie mesurable A de $(E_1 \times E_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$, est une partie mesurable de $(E_2, \mathcal{A}_2)$. De même, la section f_{e_1} d'une fonction numérique mesurable f sur $(E_1 \times E_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ est une fonction numérique mesurable sur $(E_2, \mathcal{A}_2)$.

Démonstration. Soit $\mathcal{C}_{e_1}$ la classe des parties A de $E_1 \times E_2$ telles que $A_{e_1} \in \mathcal{A}_2$. On voit facilement que tout pavé mesurable appartient à $\mathcal{C}_{e_1}$ et que $\mathcal{C}_{e_1}$ est stable pour les opérations de complémentation et d'intersection dénombrable. Il s'en suit que $\mathcal{C}_{e_1}$ contient $\mathcal{A}_1 \otimes \mathcal{A}_2$, ce qui démontre la première partie de la proposition. La seconde partie est alors une simple conséquence de l'identité : $(f_{e_1})^{-1}(B) = [f^{-1}(B)]_{e_1}$. $\square$

Corollaire IV-1.23. Pour que le pavé non vide $B_1 \times B_2$ de $E_1 \times E_2$ appartienne à $\mathcal{A}_1 \otimes \mathcal{A}_2$, il faut et il suffit que $B_1 \in \mathcal{A}_1$ et que $B_2 \in \mathcal{A}_2$.

Pour que la fonction réelle non identiquement nulle $f_1(e_1)f_2(e_2)$ définie sur $E_1 \times E_2$ soit $\mathcal{E}_1 \otimes \mathcal{E}_2$ mesurable, il faut et il suffit que f_1 (resp. f_2) soit $\mathcal{E}_1$ (resp. $\mathcal{E}_2$) 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 $\mathcal{A}_2$ d'après la proposition précédente. De même : $B_1 \in \mathcal{A}_1$. Un raisonnement analogue s'applique aux fonctions. $\square$

IV-1.2 Mesures

IV-1.2.1 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 IV-1.24 (Somme par paquets). Soit I un ensemble dénombrable, $\{P_j, j \in J\}$ 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

$$\sum_{i \in I} a_i = \sum_{j \in J} \left(\sum_{i \in P_j} a_i \right).$$

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 $S_n^\phi \leq S_m^\psi$; puis $S^\phi \leq S^\psi$. Permutant ϕ et ψ , on a $S^\psi \leq S^\phi$ 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 . $\square$

Définition IV-1.25 (Mesure, Espace mesuré). On appelle mesure sur l'espace mesurable $(E, \mathcal{E})$ toute application μ de $\mathcal{E}$ 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 $\mathcal{E}$ deux à deux disjoints,

$$\mu \left(\bigcup_{n=0}^{\infty} A_n \right) = \sum_{n=0}^{\infty} \mu(A_n).$$

Le triplet $(E, \mathcal{E}, \mu)$ s'appelle un espace mesuré.

Les propriétés suivantes découlent de façon élémentaire de la définition.

Proposition IV-1.26. Soit $(E, \mathcal{E}, \mu)$ un espace mesuré.

- (i) Si $A, B \in \mathcal{E}$ sont disjoints, alors $\mu(A \cup B) = \mu(A) + \mu(B)$.
- (ii) si $A, B \in \mathcal{E}$ et $A \subset B$, $\mu(A) \leq \mu(B)$,
- (iii) si $\{A_n, n \in \mathbb{N}\}$ est une suite d'éléments de $\mathcal{E}$, $\mu(\bigcup_{n=0}^{\infty} A_n) \leq \sum_{n=0}^{\infty} \mu(A_n)$,
- (iv) si $\{A_n, n \in \mathbb{N}\}$ est une suite d'éléments de $\mathcal{E}$ telle que, pour tout $n \in \mathbb{N}$, $A_n \subset A_{n+1}$ alors $\lim_{n \rightarrow \infty} \uparrow \mu(A_n) = \mu(A)$ où $A := \bigcup_{n=0}^{\infty} A_n$.
- (v) si $\{A_n, n \in \mathbb{N}\}$ est une suite d'éléments de $\mathcal{E}$ 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 $\lim_{n \rightarrow \infty} \downarrow \mu(A_n) = \mu(A)$ où $A := \bigcap_{n=0}^{\infty} A_n$.

Démonstration. Pour la première propriété, on écrit $A \cup B$ comme l'union dénombrable $A \cup B \cup \emptyset \cup \emptyset \cup \dots$ 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 $A \subset B$. $\square$

Définition IV-1.27 (Mesure σ -finie, Mesure bornée, Probabilité, Espace de probabilité). 1. Une mesure μ sur $(E, \mathcal{E})$ est dite σ -finie s'il existe une suite $\{E_n, n \in \mathbb{N}\}$ d'éléments de E telle que $E = \bigcup_{n=1}^{\infty} E_n$ et $\mu(E_n) < +\infty$ pour tout $n \in \mathbb{N}$.
 2. La mesure μ est bornée si $\mu(E) < +\infty$. Si $\mu(E) = 1$, la mesure μ est appelée une probabilité. Lorsque μ est une probabilité, le triplet $(E, \mathcal{E}, \mu)$ 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 $(E, \mathcal{E})$, elles sont égales si et seulement si elles coïncident sur une classe stable par intersection finie engendrant $\mathcal{E}$.

Proposition IV-1.28. Soient μ et ν deux mesures sur $(E, \mathcal{E})$ et $\mathcal{C} \subset \mathcal{E}$ une classe d'ensembles stable par intersection finie. On suppose que, pour tout $A \in \mathcal{C}$, $\mu(A) = \nu(A) < +\infty$ et que $E = \lim \uparrow E_n$ avec $E_n \in \mathcal{C}$. Alors $\mu(A) = \nu(A)$ pour tout $A \in \sigma(\mathcal{C})$.

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 $\mu(E) = \nu(E) < +\infty$. Soit $\mathcal{M} := \{A \in \mathcal{E} : \mu(A) = \nu(A)\}$. On vérifie facilement que $\mathcal{M}$ est une classe monotone qui contient $\mathcal{C}$. Par le Théorème IV-1.6, on a $\sigma(\mathcal{C}) \subset \mathcal{M}$. Le cas $\mu(E) = \nu(E) = +\infty$ se traite en appliquant ce résultat aux mesures $\mu_n(A) := \mu(A \cap E_n)$ et $\nu_n(A) := \nu(A \cap E_n)$, puis en utilisant la propriété $\mu(A) = \lim_n \uparrow \mu_n(A)$. $\square$

IV-1.2.2 Ensembles négligeables

Définition IV-1.29 (Ensemble négligeable, presque-partout, presque-sûrement). Soit $(E, \mathcal{E}, \mu)$ 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 $B \in \mathcal{E}$ 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 $\{x \in E, f(x) \neq g(x)\}$ est négligeable.

Définition IV-1.30 (Espace mesuré complet). L'espace mesuré $(E, \mathcal{E}, \mu)$ est dit complet si $\mathcal{E}$ contient la classe des ensembles négligeables $\mathcal{N}$.

On peut toujours "compléter" un espace mesuré. Pour ce faire, on définit $\overline{\mathcal{E}} = \sigma(\mathcal{E}, \mathcal{N})$. Alors $A \in \overline{\mathcal{E}}$ si et seulement si $A = B \cup N$ avec $B \in \mathcal{E}$ et $N \in \mathcal{N}$. On peut prolonger μ à $\overline{\mathcal{E}}$ en posant $\mu(A) = \mu(B)$ (il est facile de voir que ceci ne dépend pas de l'écriture de A). L'espace $(E, \overline{\mathcal{E}}, \mu)$ est alors complet et s'appelle le complété de $(E, \mathcal{E}, \mu)$. Enfin on vérifie aisément que $f : E \rightarrow \mathbb{R}$ est $\overline{\mathcal{E}}$ -mesurable si et seulement si il existe $g, h : E \rightarrow \mathbb{R}$ $\mathcal{E}$ -mesurables telles que $g \leq f \leq h$ et $g = h$ p.p.

IV-1.2.3 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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Observons d'abord que $\mathcal{C} := \{]a, b], -\infty < a \leq b < +\infty\}$ est une classe stable par intersection finie et que $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$ (voir Section IV-1.1). Il résulte alors de la Proposition IV-1.28 qu'une mesure μ sur $\mathcal{B}(\mathbb{R})$ 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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, on définit la fonction F par

$$\begin{aligned} F(0) &= 0, \\ F(x) &= \mu(]0, x]), \quad x > 0, \\ F(x) &= -\mu(]x, 0]), \quad x < 0. \end{aligned}$$

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 $\mathcal{B}(\mathbb{R})$ telle que $\mu(]a, b]) = F(b) - F(a)$? Il est facile de décrire l'algèbre $\mathcal{A}$ engendrée par $\mathcal{C}$. On a

$$\mathcal{A} = \{A = \cup_{k=1}^n]a_k, b_k], -\infty \leq a_1 < b_1 < a_2 < \dots < b_{n-1} < a_n < b_n \leq +\infty\}$$

en convenant que, si $b_n = +\infty$, $]a_n, b_n] =]a_n, +\infty[$. On définit μ sur $\mathcal{A}$ par $\mu(A) = \sum_{k=1}^n F(b_k) - F(a_k)$ où $F(+\infty) = \lim_{x \rightarrow +\infty} F(x)$, $F(-\infty) = \lim_{x \rightarrow -\infty} F(x)$. Il est facile de montrer que μ est additive sur $\mathcal{A}$, un peu plus délicat de montrer que μ est σ -additive sur $\mathcal{A}$ mais cela se fait. On a donc construit une mesure μ sur $\mathcal{A}$ telle que $\mu(]a, b]) = F(b) - F(a)$. Pour passer à $\mathcal{B}(\mathbb{R})$, on utilise le théorème de Carathéodory :

Théorème IV-1.31 (Théorème de Carathéodory). *Soit μ une mesure sur une algèbre $\mathcal{A}$, alors μ se prolonge en une mesure sur $\sigma(\mathcal{A})$. De plus, si μ est σ -finie, ce prolongement est unique.*

Puisque dans cette section $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$, nous avons :

Théorème IV-1.32. *Soit F une application de $\mathbb{R}$ dans $\mathbb{R}$ continue à droite et croissante. Il existe une et une seule mesure μ sur $\mathcal{B}(\mathbb{R})$ telle que, pour tout $a < b$, $\mu(]a, b]) = F(b) - F(a)$.*

Démonstration. Voir [?, Section 3]. □

► 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 $\mathcal{B}(\mathbb{R})$ vérifiant, pour tout intervalle I , $\lambda_{\text{Leb}}(I) = |I|$. C'est la *mesure de Lebesgue* sur $\mathbb{R}$. Si $\mathcal{N}$ est la classe des ensembles λ_{Leb} -négligeables, $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{C}, \mathcal{N})$ s'appelle la tribu des ensembles Lebesgue-mesurables (elle est beaucoup plus "grosse" que $\mathcal{B}(\mathbb{R})$) et λ_{Leb} se prolonge sans peine à $\mathcal{B}(\mathbb{R})$.

► Mesure de Lebesgue sur $\mathbb{R}^d$

Soit λ_{Leb} la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. On définit alors, sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $\lambda_{\text{Leb}}^{\otimes d} = \lambda_{\text{Leb}} \otimes \lambda_{\text{Leb}} \otimes \dots \otimes \lambda_{\text{Leb}}$. On peut appliquer la proposition IV-1.28 à

$$\mathcal{C} := \left\{ A, A = \prod_{i=1}^d]a_i, b_i[: -\infty < a_i < b_i < +\infty \right\}.$$

On obtient que $\lambda_{\text{Leb}}^{\otimes d}$ est l'unique mesure sur $\mathcal{B}(\mathbb{R}^d)$ 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 $\mathbb{R}^d$* .

IV-1.3 Intégration

Soit $(E, \mathcal{E}, \mu)$ un espace mesuré. L'objectif de cette section est de construire l'intégrale d'une fonction mesurable $f : E \rightarrow \mathbb{R}$ par rapport à la mesure μ . Cette intégrale sera notée $\mu(f)$ ou $\int f d\mu$.

Nous commençons par la définir pour des fonctions positives étagées, puis pour des fonctions positives et enfin des fonctions signées.

IV-1.3.1 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 $A_k \in \mathcal{E}$. On pose

$$\int f d\mu := \sum_{k=1}^p a_k \mu(A_k).$$

La première propriété à vérifier est que $\int f d\mu$ 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 IV-1.33. Soient $f, g \in e\mathcal{E}^+$.

- (i) Si $\mu\{f \neq g\} = 0$, alors $\int f d\mu = \int g d\mu$.
- (ii) Pour $a, b \in \mathbb{R}^+$, on a $\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$.
- (iii) Si $f \leq g$, alors $\int f d\mu \leq \int g d\mu$.

On a aussi le résultat plus technique suivant qui est la clé de la construction :

Lemme IV-1.34. Si $\{f_n, n \in \mathbb{N}\}, \{g_n, n \in \mathbb{N}\}$ sont des suites croissantes d'éléments de $e\mathcal{E}^+$ et si $\lim \uparrow f_n = \lim \uparrow g_n$, alors on a $\lim \uparrow \int f_n d\mu = \lim \uparrow \int g_n d\mu$.

Démonstration. Il suffit de montrer que

$$\lim \uparrow f_n \geq g = \sum_{k=1}^p \alpha_k \mathbb{1}_{A_k} \in e\mathcal{E}^+ \implies \lim \uparrow \int f_n d\mu \geq \int g d\mu.$$

Soient $c \in]0, 1[$ et $E_n := \{f_n \geq cg\}$. On a $E_n \in \mathcal{E}$, $E_n \uparrow E$ et $f_n \geq cg \mathbb{1}_{E_n}$ d'où

$$\int f_n d\mu \geq c \int g \mathbb{1}_{E_n} d\mu = c \sum_{k=1}^p \alpha_k \mu(A_k \cap E_n).$$

On obtient, lorsque $n \rightarrow +\infty$,

$$\lim \uparrow \int f_n d\mu \geq c \sum_{k=1}^p \alpha_k \mu(A_k) = c \int g d\mu.$$

Puisque $c < 1$ est arbitraire, cela entraîne $\lim \uparrow \int f_n d\mu \geq \int g d\mu$ et conclut la preuve. $\square$

On peut maintenant définir l'intégrale d'une fonction $f \in \mathcal{E}^+$. Par la Proposition IV-1.12, il existe une suite $\{f_n, n \in \mathbb{N}\}$ de fonctions de $\mathcal{E}^+$ telles que $f = \lim \uparrow f_n$. On a alors $\int f_n d\mu \uparrow$ et on pose

$$\int f d\mu := \lim \uparrow \int f_n d\mu.$$

Le point important est que, d'après le Lemme IV-1.34, cette limite ne dépend pas de la suite $\{f_n, n \in \mathbb{N}\}$ choisie. On dira que $f \in \mathcal{E}^+$ est *intégrable* si $\int f d\mu < +\infty$.

On a les propriétés suivantes

Proposition IV-1.35. Soient $f, g \in \mathcal{E}^+$.

- (i) Pour $a, b \in \mathbb{R}^+$, on a $\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$.
- (ii) Si $f \leq g$, alors $\int f d\mu \leq \int g d\mu$.
- (iii) Si $\int f d\mu < +\infty$, alors $f < +\infty$ p.p.
- (iv) Si $\int f d\mu = 0$, 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 $\mathcal{E}^+$; 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 $f = f \mathbb{1}_{f=0} + f \mathbb{1}_{f>0} \geq f \mathbb{1}_{f>0}$. $\square$

IV-1.3.2 Intégration des fonctions réelles ou complexes

On pose

$$L^1(E, \mathcal{E}, \mu) := \left\{ f \in [\mathcal{E}], \int |f| d\mu < +\infty \right\}. \quad (\text{IV-1.4})$$

Quand il n'y a pas d'ambiguïté sur l'espace sur lequel on intègre, on posera $L^1(\mu) := L^1(E, \mathcal{E}, \mu)$.

Si $f \in L^1(\mu)$, $f^+ := \max(f, 0)$ et $f^- := \max(-f, 0)$ sont intégrables et on pose

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu.$$

Il est facile de voir (vu que $|f + g| \leq |f| + |g|$) que $L^1(\mu)$ est un espace vectoriel et que $f \rightarrow \int f d\mu$ est une forme linéaire positive sur $L^1(\mu)$.

On a aussi les propriétés :

Proposition IV-1.36. Soient $f, g \in L^1(\mu)$.

- (i) Si $f \leq g$ p.p., alors $\int f d\mu \leq \int g d\mu$.
- (ii) $|\int f d\mu| \leq \int |f| d\mu$.
- (iii) Si $A \in \mathcal{E}$, alors $f \mathbb{1}_A \in L^1(\mu)$. On pose $\int_A f d\mu := \int f \mathbb{1}_A d\mu$.
- (iv) Si pour tout $A \in \mathcal{E}$, $\int_A f d\mu \geq 0$ alors $f \geq 0$ p.p.
- (v) Si, pour tout $A \in \mathcal{E}$, $\int_A f d\mu \leq \int_A g d\mu$, alors $f \leq g$ 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 :

$$\left| \int f \, d\mu \right| = \left| \int f^+ \, d\mu - \int f^- \, d\mu \right| \leq \int f^+ \, d\mu + \int f^- \, d\mu = \int (f^+ + f^-) \, d\mu = \int |f| \, d\mu.$$

La troisième propriété découle de la majoration $|f| \mathbb{1}_A \leq |f|$. 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 $A = \{g - f < 0\}$. $\square$

Si f est $\mathcal{E}$ -mesurable à valeurs $\mathbb{C}$, on pose ($|f|$ désignant le module),

$$L^1_{\mathbb{C}} = L^1_{\mathbb{C}}(E, \mathcal{E}, \mu) := \left\{ f \text{ } \mathcal{E}\text{-mesurable complexe, } \int |f| \, d\mu < +\infty \right\}. \quad (\text{IV-1.5})$$

On définit alors, pour $f \in L^1_{\mathbb{C}}$,

$$\int f \, d\mu := \int \operatorname{Re}(f) \, d\mu + i \int \operatorname{Im}(f) \, d\mu.$$

$L^1_{\mathbb{C}}$ est un espace vectoriel sur $\mathbb{C}$ et $f \rightarrow \int f \, d\mu$ une forme linéaire sur $L^1_{\mathbb{C}}$. On a aussi

Proposition IV-1.37. *Pour toute fonction $f \in L^1_{\mathbb{C}}$, $|\int f \, d\mu| \leq \int |f| \, d\mu$.*

Démonstration. On a $\int f \, d\mu = r e^{i\theta}$ et

$$\begin{aligned} \left| \int f \, d\mu \right| &= r = \operatorname{Re} \left(e^{-i\theta} \int f \, d\mu \right) = \operatorname{Re} \left(\int (e^{-i\theta} f) \, d\mu \right) = \int \operatorname{Re}(e^{-i\theta} f) \, d\mu \\ &\leq \int |e^{-i\theta} f| \, d\mu = \int |f| \, d\mu. \end{aligned}$$

L'inégalité est une conséquence de la positivité de l'intégrale pour des fonctions de $L^1(\mu)$ (voir Proposition IV-1.36). $\square$

IV-1.3.3 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 IV-1.38 (Convergence monotone ou Beppo-Levi). *Soit $(E, \mathcal{E}, \mu)$ un espace mesuré et $\{f_n, n \in \mathbb{N}\}$ une suite croissante de fonctions de $\mathcal{E}^+$. Alors*

$$\lim \uparrow \int f_n \, d\mu = \int \lim \uparrow f_n \, d\mu.$$

Démonstration. Pour tout n , il existe une famille $\{f_{n,k}, k \in \mathbb{N}\}$ de fonction de $\mathcal{E}^+$ telles que $f_n = \lim \uparrow_k f_{n,k}$. On pose $g_k := \max_{n \leq k} f_{n,k}$. On a $g_k \in \mathcal{E}^+$, la suite $\{g_k, k \in \mathbb{N}\}$ est croissante et pour $n \leq k$,

$$f_{n,k} \leq g_k \leq f_n, \quad \int f_{n,k} \, d\mu \leq \int g_k \, d\mu \leq \int f_n \, d\mu.$$

On pose $f := \lim \uparrow f_n$. On a, pour $k \rightarrow +\infty$,

$$f_n \leq \lim \uparrow g_k \leq f_n, \quad \int f_n d\mu \leq \lim \uparrow \int g_k d\mu = \int \lim \uparrow g_k d\mu \leq \int f_n d\mu,$$

puis pour $n \rightarrow +\infty$,

$$f \leq \lim \uparrow g_n \leq f, \quad \lim \uparrow \int f_n d\mu \leq \int \lim \uparrow g_n d\mu \leq \lim \uparrow \int f_n d\mu. \quad \square$$

On en déduit $f = \lim \uparrow g_n$ et $\int f d\mu = \lim \uparrow \int f_n d\mu$.

En appliquant le théorème précédent avec $f_n \leftarrow \sum_{k=0}^n g_k$, on obtient le corollaire suivant :

Corollaire IV-1.39. Soit $(E, \mathcal{E}, \mu)$ un espace mesuré et $\{g_n, n \in \mathbb{N}\}$ une suite d'éléments de $\mathcal{E}^+$. Nous avons

$$\sum_n \int g_n d\mu = \int \sum_n g_n d\mu.$$

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 IV-1.40. (Lemme de Fatou) Soit $\{f_n, n \in \mathbb{N}\}$ une suite d'éléments de $\mathcal{E}^+$. Alors

$$\int \liminf f_n d\mu \leq \liminf \int f_n d\mu.$$

Démonstration. On a $\liminf f_n = \lim_n \uparrow \inf_{k \geq n} f_k$, qui est la limite d'une suite croissante de fonctions de $\mathcal{E}^+$. Par application du Théorème IV-1.38 et en utilisant l'inégalité $\inf_{k \geq n} f_k \leq f_k$ pour tout $k \geq n$, on a

$$\int \liminf f_n d\mu = \lim_n \uparrow \int \inf_{k \geq n} f_k d\mu \leq \lim_n \uparrow \inf_{k \geq n} \int f_k d\mu = \liminf \int f_n d\mu. \quad \square$$

Théorème IV-1.41 (Théorème de convergence dominée ou de Lebesgue). Soit $(E, \mathcal{E}, \mu)$ un espace mesuré et $\{f_n, n \in \mathbb{N}\}$ une suite d'éléments de $L^1_{\mathbb{C}}$ telle que

- (i) il existe une fonction $f \in [\mathcal{E}]$ telle que $\lim_{n \rightarrow \infty} f_n = f$ p.p.
- (ii) il existe une fonction positive $g \in L^1(\mu)$ telle que $|f_n| \leq g$ p.p.

Alors

$$\lim \int f_n d\mu = \int f d\mu.$$

Démonstration. Remarquons tout d'abord qu'il suffit de considérer la cas réel ; ainsi, dans la suite, $f_n \in L^1(\mu)$ pour tout n . Appliquant la Proposition IV-1.40 aux fonctions positives $g + f_n$ et $g - f_n$, puis en retranchant $\int g d\mu$ qui est finie par hypothèse, on a

$$\int \liminf f_n d\mu \leq \liminf \int f_n d\mu \leq \limsup \int f_n d\mu \leq \int \limsup f_n d\mu.$$

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 IV-1.41 en se souvenant que $\lim_{t \rightarrow t_0} \int f_t d\mu = \int f d\mu$ si et seulement si pour toute suite $\{t_n, n \in \mathbb{N}\}$ tendant vers t_0 , $\lim_{t_n \rightarrow t_0} \int f_{t_n} d\mu = \int f d\mu$.

Corollaire IV-1.42. Soit $(E, \mathcal{E}, \mu)$ un espace mesuré, U un ouvert de $\mathbb{R}^d$ et $\{f_t, t \in U\}$ une famille d’éléments de $L^1_{\mathbb{C}}$. On suppose que $\lim_{t \rightarrow t_0} f_t = f$ p.p. et qu’il existe une fonction $g \in L^1(\mu)$ telle que pour tout $t \in U$, $|f_t| \leq g$ p.p. Alors $\lim_{t \rightarrow t_0} \int f_t d\mu = \int f d\mu$.

Donnons un exemple d’utilisation de ce corollaire.

Proposition IV-1.43. Soient $(E, \mathcal{E}, \mu)$ un espace mesuré, I un intervalle ouvert et $\{f(t, \cdot), t \in I\}$ une famille d’éléments de $L^1_{\mathbb{C}}(\mu)$. On pose, pour tout $t \in I$,

$$\phi(t) := \int f(t, x) \mu(dx).$$

On suppose qu’il existe $A \in \mathcal{E}$ et une fonction $g \in L^1(\mu)$ 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

$$\phi'(t) = \int \frac{\partial f}{\partial t}(t, x) \mu(dx).$$

Démonstration. Soit $t \in I$. Puisque $\mu(A^c) = 0$, on a pour tout $h \in \mathbb{R}$,

$$h^{-1} \{ \phi(t+h) - \phi(t) \} = \int_A h^{-1} (f(t+h, x) - f(t, x)) \mu(dx).$$

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

$$\lim_{h \rightarrow 0} h^{-1} (f(t+h, x) - f(t, x)) = \frac{\partial f}{\partial t}(t, x).$$

On peut appliquer le Corollaire IV-1.42 et encore une fois $\mu(A^c) = 0$ pour obtenir

$$\lim_{h \rightarrow 0} \int_A h^{-1} (f(t+h, x) - f(t, x)) \mu(dx) = \int_A \frac{\partial f}{\partial t}(t, x) \mu(dx) = \int \frac{\partial f}{\partial t}(t, x) \mu(dx). \quad \square$$

IV-1.3.4 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 $a \leq x \leq b$,

$$F(x) := \int_a^x f(t) dt, \quad G(x) := \int \mathbb{1}_{[a, x]} f d\mu;$$

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

$$\int_a^b f(t) dt = \int \mathbb{1}_{[a, b[} f d\mu.$$

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 $\int_{-\infty}^{+\infty} f(t) dt$ soit absolument convergente. Lorsque $a \downarrow -\infty$ et $b \uparrow +\infty$,

- d'une part, par définition, $\int_a^b |f(t)| dt \rightarrow \int_{-\infty}^{+\infty} |f(t)| dt < +\infty$ et $\int_a^b f(t) dt \rightarrow \int_{-\infty}^{+\infty} f(t) dt$;
- d'autre part, $\int \mathbb{1}_{[a, b[} |f| d\mu \rightarrow \int |f| d\mu$ (par application du théorème de convergence monotone) ce qui implique que $f \in L^1(\mu)$ puis $\int \mathbb{1}_{[a, b[} f d\mu \rightarrow \int f d\mu$ (théorème de Lebesgue) puisque $|\mathbb{1}_{[a, b[} f| \leq |f| \in L^1(\mu)$.

Donc

$$\int_{-\infty}^{+\infty} f(t) dt = \int f d\mu.$$

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 $f : \mathbb{R}^k \rightarrow \mathbb{R}$.

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 intégrable au sens de Riemann.

Par contre, si $\int_{-\infty}^{+\infty} f(t) dt$ est convergente mais pas absolument convergente (par exemple $f(x) = \sin x/x$), alors $f \notin L^1(\mu)$.

► Mesure de comptage

Soient E un ensemble dénombrable et $\mu : E \rightarrow \overline{\mathbb{R}^+}$. On pose pour tout $A \in E$,

$$\mu(A) := \sum_{x \in A} \mu(x).$$

Le Lemme IV-1.24 implique que μ est une mesure sur l'espace mesurable $(E, \mathcal{P}(E))$. On a alors

$$L^1(\mu) = \left\{ f, \sum_{x \in E} |f(x)| \mu(x) < +\infty \right\}$$

et, pour $f \in L^1(\mu)$,

$$\int f d\mu = \sum_{x \in E} f(x) \mu(x).$$

En particulier si on prend pour μ la mesure de comptage i.e. $\mu(x) = 1$ pour tout $x \in E$, on a

$$L^1(\mu) = \left\{ f, \sum_{x \in E} |f(x)| < +\infty \right\} \quad \text{et} \quad \int f d\mu = \sum_{x \in E} f(x).$$

Il est intéressant d'énoncer dans ce cadre les théorèmes de convergence/permutations limite-intégrale, vus précédemment. On a

- (i) (Beppo-Levi) Si $0 \leq f_n \uparrow f$, alors $\sum_x f_n(x) \uparrow \sum_x f(x)$.
- (ii) (Fatou) Si $0 \leq f_n$, alors $\sum_x \liminf_n f_n(x) \leq \liminf_n \sum_x f_n(x)$.
- (iii) (Convergence dominée) Si $f_n \rightarrow f$ et si $|f_n| \leq g$ 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 $(E, \mathcal{E})$, on peut associer une application I de $\mathcal{E}^+$ dans $\overline{\mathbb{R}^+}$ en posant

$$I(f) := \int f d\mu, \quad \forall f \in \mathcal{E}^+.$$

L'application I a alors les propriétés suivantes :

$$\begin{aligned} I(f + g) &= I(f) + I(g), \\ I(af) &= aI(f), \quad a \in \mathbb{R}^+, \\ I(f_n) \uparrow I(f) &\text{ si } f_n \uparrow f. \end{aligned}$$

Réciproquement on a,

Proposition IV-1.44. Soient $(E, \mathcal{E})$ un espace mesurable et I une application de $\mathcal{E}^+$ dans $\overline{\mathbb{R}^+}$ telle que

- (i) si $f, g \in \mathcal{E}^+$, $I(f + g) = I(f) + I(g)$;
- (ii) si $f \in \mathcal{E}^+$ et $a \in \mathbb{R}^+$, $I(af) = aI(f)$,
- (iii) si $f_n \in \mathcal{E}^+$ et si $f_n \uparrow f$, $I(f_n) \uparrow I(f)$.

Alors l'application $\mu : \mathcal{E} \rightarrow \overline{\mathbb{R}^+}$ définie par

$$\mu(A) := I(\mathbb{1}_A)$$

est une mesure sur $\mathcal{E}$ et on a, pour toute fonction $f \in \mathcal{E}^+$, $I(f) = \int f d\mu$.

Démonstration. Soient $\{A_n, n \in \mathbb{N}\}$ une suite d'éléments de $\mathcal{E}$ 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 $f \in \mathcal{E}^+$, $I(f) = \int f d\mu$. On conclut facilement en utilisant la Proposition IV-1.12. $\square$

Théorème IV-1.45. Soient h une application mesurable de $(E, \mathcal{E})$ dans $(Y, \mathcal{Y})$ et μ une mesure sur $(E, \mathcal{E})$. La formule

$$\nu(A) = \mu(h^{-1}(A)), \quad A \in \mathcal{Y},$$

définit une mesure sur $(Y, \mathcal{Y})$ appelée mesure image de μ par h et notée $\mu \circ h^{-1}$ ou μ^h . On a, pour toute $f \in \mathcal{Y}^+$,

$$\int f d\mu \circ h^{-1} = \int f \circ h d\mu. \quad (\text{IV-1.6})$$

De plus $f \in [\mathcal{Y}]$ est $\mu \circ h^{-1}$ -intégrable si et seulement si $f \circ h$ est μ -intégrable et dans ce cas, f vérifie (IV-1.6).

Démonstration. On considère la fonctionnelle $I(f) = \int f \circ h d\mu$, $f \in \mathcal{Y}^+$ et on applique la Proposition IV-1.44. La mesure associée à I est

$$\nu(A) = \int \mathbb{1}_A \circ h d\mu = \int \mathbb{1}_{h^{-1}(A)} d\mu = \mu(h^{-1}(A)).$$

On conclut facilement. $\square$

► Mesures à densité

Théorème IV-1.46. Soient $(E, \mathcal{E}, \mu)$ un espace mesuré et $h \in \mathcal{E}^+$. La formule

$$\nu(A) := \int_A h d\mu, \quad A \in \mathcal{E},$$

définit une mesure sur $\mathcal{E}$, notée $h \cdot \mu$. On a, pour toute $f \in \mathcal{E}^+$,

$$\int f d\nu = \int fh d\mu. \quad (\text{IV-1.7})$$

De plus $f \in [\mathcal{E}]$ est ν -intégrable si et seulement si fh est μ -intégrable et dans ce cas, f vérifie la relation (IV-1.7). Enfin, si $\mu(\{h \neq h'\}) = 0$ alors les mesures $h \cdot \mu$ et $h' \cdot \mu$ coïncident.

Démonstration. On considère la fonctionnelle $I(f) = \int fh d\mu$, $f \in \mathcal{E}^+$, et on applique la Proposition IV-1.44. La dernière assertion s'établit en écrivant $f = f^+ - f^-$. $\square$

Définition IV-1.47 (densité d'une mesure, absolue continuité, domination, mesure étrangère). Soient $(E, \mathcal{E})$ un espace mesurable et μ une mesure sur $\mathcal{E}$.

- (i) Pour $f \in \mathcal{E}^+$, la mesure $\nu : B \mapsto \int_B f d\mu$ sur $\mathcal{E}$ 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 $f = \frac{d\nu}{d\mu}$.
- (ii) Une mesure ν est dite absolument continue par rapport à μ si, pour tout $B \in \mathcal{E}$ 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 $N \in \mathcal{E}$ 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 $N(0, 1)$ 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 $\lambda_{\text{Leb}}(\mathbb{N}) = 0$ et $\mu(\mathbb{N}^c) = 0$.

Nous admettrons le résultat très important suivant ; on trouvera une démonstration dans [?, Théorème 32.2].

Théorème IV-1.48 (Radon-Nikodym). Soit μ une mesure σ -finie sur un espace mesurable $(E, \mathcal{E})$. 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.

IV-1.3.5 Intégration par rapport à une mesure produit

Théorème IV-1.49 (Théorème de Fubini). Soient $(E_1, \mathcal{E}_1, \mu_1)$ et $(E_2, \mathcal{E}_2, \mu_2)$ deux espaces mesurés avec μ_1 et μ_2 σ -finies.

Il existe une unique mesure sur $\mathcal{E}_1 \otimes \mathcal{E}_2$, notée $\mu_1 \otimes \mu_2$ et appelée mesure produit de μ_1 et μ_2 , telle que,

$$\text{pour tout } A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2, \mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1) \mu_2(A_2).$$

De plus, pour toute fonction $f \in (\mathcal{E}_1 \otimes \mathcal{E}_2)^+$,

$$\int f d\mu_1 \otimes \mu_2 = \int \left[\int f(x_1, x_2) d\mu_1(x_1) \right] d\mu_2(x_2) = \int \left[\int f(x_1, x_2) d\mu_2(x_2) \right] d\mu_1(x_1).$$

Soit $f \in L^1_{\mathbb{C}}(\mu_1 \otimes \mu_2)$. Alors,

- (i) $\int |f(x_1, x_2)| d\mu_2(x_2) < +\infty$ μ_1 p.p. et $x_1 \mapsto \int f(x_1, x_2) d\mu_2(x_2) \in L^1_{\mathbb{C}}(\mu_1)$,
- (ii) $\int |f(x_1, x_2)| d\mu_1(x_1) < +\infty$ μ_2 p.p. et $x_2 \mapsto \int f(x_1, x_2) d\mu_1(x_1) \in L^1_{\mathbb{C}}(\mu_2)$, et

$$\int f d\mu_1 \otimes \mu_2 = \int \left[\int f(x_1, x_2) d\mu_1(x_1) \right] d\mu_2(x_2) = \int \left[\int f(x_1, x_2) d\mu_2(x_2) \right] d\mu_1(x_1).$$

Démonstration. (i) On applique la Proposition IV-1.28 à

$$\mathcal{C} := \{A, A = A_1 \times A_2, A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2, \mu_1(A_1) < +\infty, \mu_2(A_2) < +\infty\}.$$

(ii) On applique la proposition IV-1.44 à $I_1(f) := \int [\int f(x_1, x_2) d\mu_1(x_1)] d\mu_2(x_2)$ ce qui donne l'existence. Mais on peut aussi appliquer la proposition IV-1.44 à

$$I_2(f) := \int \left[\int f(x_1, x_2) d\mu_2(x_2) \right] d\mu_1(x_1)$$

et, vu l'unicité, on a $I_1(f) = I_2(f)$.

- (iii) Si $f \in L^1_{\mathbb{C}}(\mu_1 \otimes \mu_2)$, on applique le résultat précédent à $[\operatorname{Re}(f)]^+$, $[\operatorname{Re}(f)]^-$, $[\operatorname{Im}(f)]^+$ et $[\operatorname{Im}(f)]^-$. $\square$

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 $\mathbb{R}^+$; 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 .

Chapitre IV-2

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-2.1 Convergence en probabilité

Définition IV-2.1 (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 $(\Omega, \mathcal{F}, \mathbb{P})$. Nous dirons que la suite $\{X_n, n \in \mathbb{N}\}$ converge en probabilité vers X et nous noterons

$$X_n \xrightarrow{\mathbb{P}\text{-prob}} X,$$

si pour tout $\varepsilon > 0$, nous avons :

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

Exemple IV-2.2. 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 $(X_1, \dots, X_n)$, 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

$$\text{Var}(T_n) = \sum_{i=1}^{n-1} \text{Var}(I_i) + 2 \sum_{i=1}^{n-2} \text{Cov}(I_i, I_{i+1}) = \frac{3(n-1)}{16} - \frac{2(n-2)}{16} = \frac{n+1}{16}.$$

En appliquant l'inégalité de Bienayme-Tchebychev (Lemme IV-3.2), nous avons, pour tout $\delta > 0$,

$$\mathbb{P}(|T_n - (n-1)/4| \geq \delta) \leq \frac{n+1}{16\delta^2}.$$

soit aussi

$$\mathbb{P}(|T_n/n - (1-1/n)/4| \geq \varepsilon) \leq \frac{n+1}{16\varepsilon^2 n^2} \rightarrow 0.$$

L'intuition est donc que $T_n/n \xrightarrow{\mathbb{P}\text{-prob}} 1/4$. Montrons-le rigoureusement. En écrivant que

$$\begin{aligned} \{|T_n/n - 1/4| \geq \varepsilon\} &\subset \{|T_n/n - (n-1)/(4n)| + 1/(4n) \geq \varepsilon\} \\ &\subset \{|T_n/n - (n-1)/(4n)| \geq \varepsilon/2\} \cup \{1/(4n) \geq \varepsilon/2\} \end{aligned}$$

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 $n^{-1}T_n \xrightarrow{\mathbb{P}\text{-prob}} 1/4$. ◇

Proposition IV-2.3. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. Supposons qu'il existe $p > 0$ tel que $\lim_{n \rightarrow \infty} \mathbb{E}[\|X_n - X\|^p] = 0$. Alors $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$.

Démonstration. Par l'inégalité de Markov (Lemme IV-3.1), pour tout $\varepsilon > 0$,

$$\mathbb{P}(\|X_n - X\| > \varepsilon) \leq \frac{\mathbb{E}[\|X_n - X\|^p]}{\varepsilon^p} \rightarrow 0.$$

□

Comme le montre la proposition suivante, la convergence en probabilité d'un vecteur aléatoire est équivalente à la convergence de ses coordonnées.

Proposition IV-2.4. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. $(X_n, Y_n) \xrightarrow{\mathbb{P}\text{-prob}} (X, Y)$ si et seulement si $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$ et $Y_n \xrightarrow{\mathbb{P}\text{-prob}} Y$.

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-2.5. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. Supposons que $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$. Alors, il existe une suite décroissante $\{\delta_n, n \in \mathbb{N}\}$ telle que $\lim_{n \rightarrow \infty} \delta_n = 0$ et vérifiant

$$\lim_{n \rightarrow \infty} \mathbb{P}(\|X_n - X\| \geq \delta_n) = 0.$$

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 $\lim_{n \rightarrow \infty} \delta_n = 0$ et $\lim_{n \rightarrow \infty} \mathbb{P}(\|X_n - X\| \geq \delta_n) = 0$. □

Théorème IV-2.6. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. 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 $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$, alors $g(X_n) \xrightarrow{\mathbb{P}\text{-prob}} g(X)$.

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 $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$. Nous avons $\mathbb{P}(X \in B_\delta \cap C^c) = 0$ et $\lim_{\delta \rightarrow 0} \mathbb{P}(X \in B_\delta \cap C) = 0$ par continuité de g . $\square$

Corollaire IV-2.7. Soit $\{X_n, n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ définis sur $(\Omega, \mathcal{F}, \mathbb{P})$. Soit $g : \mathbb{R}^d \mapsto \mathbb{R}^m$ une fonction continue au point $c \in \mathbb{R}^d$. Si $X_n \xrightarrow{\mathbb{P}\text{-prob}} c$, alors $g(X_n) \xrightarrow{\mathbb{P}\text{-prob}} g(c)$.

Démonstration. Il suffit d'appliquer le théorème IV-2.6 avec $X = c$. $\square$

IV-2.2 Convergence presque-sûre

Définition IV-2.8 (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 $(\Omega, \mathcal{F}, \mathbb{P})$. Nous dirons que la suite $\{X_n, n \in \mathbb{N}\}$ converge presque-sûrement vers X et nous noterons

$$X_n \xrightarrow{\mathbb{P}\text{-p.s.}} X,$$

si, pour tout $\varepsilon > 0$,

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

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et soit $\{A_n, n \in \mathbb{N}\}$ une suite d'éléments de $\mathcal{F}$. 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 IV-1.1)

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{p=n}^{\infty} A_p.$$

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

$$A_n = \limsup_n \{\|X_n - X\| > \varepsilon\} = \{\limsup_n \|X_n - X\| > \varepsilon\}. \quad (\text{IV-2.1})$$

Proposition IV-2.9. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. 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 $\Omega_0 \in \mathcal{F}$ de probabilité 1 tel que, pour tout $\omega \in \Omega_0$, $\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)$.

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-2.1),

$$\limsup_n \{\|X_n - X\| \geq \varepsilon\} = \{\limsup_n \|X_n - X\| \geq \varepsilon\} \subseteq \Omega_0^c.$$

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 . $\square$

Comme le montre le résultat ci-dessous, la convergence presque-sûre implique la convergence en probabilité.

Théorème IV-2.10. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. Si $X_n \xrightarrow{\mathbb{P}\text{-p.s.}} X$ alors $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$.

Démonstration. Soit $\varepsilon > 0$. Posons $A_n := \bigcup_{k \geq n} \{\|X_k - X\| \geq \varepsilon\}$. On a

$$\{\|X_n - X\| \geq \varepsilon\} \subseteq A_n, \quad \lim_n A_n = \bigcap_n A_n.$$

Par suite, et en utilisant le théorème de convergence monotone (voir Théorème IV-1.38), 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 $X_n \xrightarrow{\mathbb{P}\text{-p.s.}} X$, $\mathbb{P}(\bigcap_n A_n) = 0$, ce qui conclut la preuve. $\square$

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-2.11 (Lemme de Borel-Cantelli). Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $\{A_n, n \in \mathbb{N}\}$ une suite d'éléments de $\mathcal{F}$. Alors :

$$\sum_{n \geq 0} \mathbb{P}(A_n) < \infty \Rightarrow \mathbb{P}\left(\limsup_n A_n\right) = 0. \quad (\text{IV-2.2})$$

Démonstration. Remarquons que $\{\bigcup_{p \geq n} A_p, n \in \mathbb{N}\}$ est une suite décroissante d'événements dont la limite est $\limsup_n A_n$, et l'on a

$$\mathbb{P}\left(\limsup_n A_n\right) = \lim_n \mathbb{P}\left(\bigcup_{p \geq n} A_p\right).$$

Or $\mathbb{P}(\bigcup_{p \geq n} A_p) \leq \sum_{p \geq n} \mathbb{P}(A_p)$ qui tend vers zero quand $n \rightarrow \infty$ par hypothèse, ce qui montre (IV-2.2). $\square$

Le lemme de Borel-Cantelli permet d'établir la condition suffisante de convergence presque-sûre suivante.

Lemme IV-2.12. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. 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 $X_n \xrightarrow{\mathbb{P}\text{-p.s.}} X$.

Démonstration. D'après le lemme de Borel-Cantelli (voir Lemme IV-2.11), il existe un ensemble $N \in \mathcal{F}$ 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 $N(\omega)$, $\|X_n(\omega) - X(\omega)\| < \varepsilon_n$. $\square$

Théorème IV-2.13. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. 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 $X_n \xrightarrow{\mathbb{P}\text{-p.s.}} X$, alors $g(X_n) \xrightarrow{\mathbb{P}\text{-p.s.}} g(X)$.

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 entraîne $\mathbb{P}(\Omega_0 \cap \Omega_1) = 1$. Or, pour tout $\omega \in \Omega_0 \cap \Omega_1$, nous avons $\lim_{n \rightarrow \infty} g(X_n(\omega)) = g(X(\omega))$. Ce qui conclut la preuve d'après la Proposition IV-2.9. $\square$

IV-2.3 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 décorrélées : c'est l'objectif de la section IV-2.3.1. Dans la IV-2.3.2, nous montrons une version plus forte de la convergence en loi, qui permet d'établir la consistance forte de certains estimateurs.

IV-2.3.1 Loi faible des grands nombres

Définition IV-2.14 (Variables décorrélées). Soient $X_1, \dots, X_n$, des variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ et de carré intégrable, $\mathbb{E}[X_k^2] < \infty$ pour tout $k \in \{1, \dots, n\}$. Les variables aléatoires $X_1, \dots, X_n$ sont décorrélées si, pour tout $1 \leq k \neq \ell \leq n$,

$$\text{Cov}(X_k, X_\ell) := \mathbb{E}[(X_k - \mathbb{E}[X_k])(X_\ell - \mathbb{E}[X_\ell])] = 0.$$

Le principal intérêt des variables décorrélées est donné par le lemme suivant.

Lemme IV-2.15 (Lemme de Bienaymé). Soient $X_1, \dots, X_n$, des variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ 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,

$$\text{Var} \left(\sum_{k=1}^n X_k \right) = \sum_{k=1}^n \text{Var}(X_k) .$$

Démonstration. Un calcul immédiat montre que

$$\text{Var} \left(\sum_{k=1}^n X_k \right) = \sum_{k=1}^n \mathbb{E}[(X_k - \mathbb{E}[X_k])^2] + \sum_{1 \leq k \neq \ell \leq n} \mathbb{E}[(X_k - \mathbb{E}[X_k])(X_\ell - \mathbb{E}[X_\ell])] = \sum_{k=1}^n \text{Var}(X_k) . \quad \square$$

Ce lemme permet d'établir, par des arguments élémentaires, une loi des grands nombres, attribuée à Markov.

Théorème IV-2.16 (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

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) = 0 ,$$

et que les variables aléatoires $\{X_k, k \in \mathbb{N}\}$ sont décorrélées. Alors

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\left(\frac{1}{n} \sum_{k=1}^n \{X_k - \mathbb{E}[X_k]\} \right)^2 \right] = 0 , \quad (\text{IV-2.3})$$

et

$$\frac{1}{n} \sum_{k=1}^n \{X_k - \mathbb{E}[X_k]\} \xrightarrow{\mathbb{P}\text{-prob}} 0 . \quad (\text{IV-2.4})$$

Démonstration. En utilisant le lemme de Bienaymé (Lemme IV-2.15), nous avons

$$\text{Var} \left(\frac{1}{n} \sum_{k=1}^n \{X_k - \mathbb{E}[X_k]\} \right) = \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) ,$$

ce qui montre (IV-2.3). La relation (IV-2.4) est une conséquence de la proposition IV-2.3. $\square$

Corollaire IV-2.17 (Loi faible pour des variables i.i.d. de carré intégrable). Soit $\{Y_n, n \in \mathbb{N}\}$ une suite de variables définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, i.i.d. et de carré intégrable. Alors

$$\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{\mathbb{P}\text{-prob}} \mathbb{E}[Y_1] .$$

On peut affaiblir la condition de moment du corollaire IV-2.17 en tronquant les variables aléatoires. Le résultat est important, mais sa preuve est plus technique et peut être omise en première lecture.

Théorème IV-2.18 (Loi faible des grands nombres pour des variables i.i.d. intégrables). Soit $\{X_k, k \in \mathbb{N}\}$ des variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ i.i.d. et intégrables. Alors

$$\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{\mathbb{P}\text{-prob}} \mathbb{E}[X_1] .$$

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 $T_n := \sum_{k=1}^n X_k \mathbb{1}_{\{|X_k| \leq n\varepsilon^3\}}$. Notons tout d'abord que, sur l'événement $\bigcap_{k=1}^n \{|X_k| \leq n\varepsilon^3\}$, $R_n = T_n$. Par suite,

$$\begin{aligned} & \mathbb{P}(|R_n - \mathbb{E}[T_n]| \geq n\varepsilon) \\ &= \mathbb{P}\left(|R_n - \mathbb{E}[T_n]| \geq n\varepsilon, \bigcap_{k=1}^n \{|X_k| \leq n\varepsilon^3\}\right) + \mathbb{P}\left(|R_n - \mathbb{E}[T_n]| \geq n\varepsilon, \bigcup_{k=1}^n \{|X_k| > n\varepsilon^3\}\right) \\ &\leq \mathbb{P}(|T_n - \mathbb{E}[T_n]| \geq n\varepsilon) + \mathbb{P}\left(\bigcup_{k=1}^n \{|X_k| > n\varepsilon^3\}\right) . \end{aligned} \quad (\text{IV-2.5})$$

En utilisant l'inégalité de Bienayme-Tchebychev (Lemme IV-3.2), nous avons

$$\begin{aligned} \mathbb{P}(|T_n - \mathbb{E}[T_n]| \geq n\varepsilon) &\leq \frac{1}{n\varepsilon^2} \text{Var}(X_1 \mathbb{1}_{|X_1| \leq n\varepsilon^3}) \leq \frac{1}{n\varepsilon^2} \mathbb{E}[X_1^2 \mathbb{1}_{|X_1| \leq n\varepsilon^3}] \\ &\leq \frac{n\varepsilon^3}{n\varepsilon^2} \mathbb{E}[|X_1|] \leq \varepsilon \mathbb{E}[|X_1|] . \end{aligned} \quad (\text{IV-2.6})$$

D'autre part, nous avons

$$\mathbb{P}\left(\bigcup_{k=1}^n \{|X_k| > n\varepsilon^3\}\right) \leq \sum_{k=1}^n \mathbb{P}(|X_k| > n\varepsilon^3) = n\mathbb{P}(|X_1| > n\varepsilon^3) .$$

En utilisant l'inégalité $\mathbb{1}_{\{|X_1| > n\varepsilon^3\}} \leq (|X_1|/n\varepsilon^3) \mathbb{1}_{\{|X_1| > n\varepsilon^3\}}$, nous avons

$$n\mathbb{P}(|X_1| > n\varepsilon^3) \leq \frac{1}{\varepsilon^3} \mathbb{E}[|X_1| \mathbb{1}_{|X_1| > n\varepsilon^3}]$$

et le théorème de convergence dominée implique

$$\limsup_{n \rightarrow \infty} n\mathbb{P}(|X_1| > n\varepsilon^3) = 0 .$$

Par conséquent, en combinant ce résultat avec (IV-2.6), nous obtenons

$$\limsup_{n \rightarrow \infty} \mathbb{P}(|R_n - \mathbb{E}[T_n]| \geq n\varepsilon) \leq \varepsilon \mathbb{E}[|X_1|] . \quad (\text{IV-2.7})$$

Notons que

$$\frac{\mathbb{E}[T_n]}{n} = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[X_k \mathbb{1}_{\{|X_k| \leq n\varepsilon^3\}}] = \mathbb{E}[X_1 \mathbb{1}_{\{|X_1| \leq n\varepsilon^3\}}]$$

et le théorème de convergence dominée implique que

$$\lim_{n \rightarrow \infty} n^{-1} \mathbb{E}[T_n] = \mathbb{E}[X_1] = 0 . \quad (\text{IV-2.8})$$

En utilisant (IV-2.7) et (IV-2.8), nous avons

$$\begin{aligned} \limsup_n \mathbb{P}\left(\left|\frac{R_n}{n}\right| \geq \varepsilon\right) &\leq \limsup_n \mathbb{P}(|R_n - \mathbb{E}[T_n]| \geq n\varepsilon/2) + \limsup_n \mathbb{P}(|\mathbb{E}[T_n]| \geq n\varepsilon/2) \\ &\leq \varepsilon \mathbb{E}[|X_1|] , \end{aligned}$$

ce qui conclut la preuve car ε est arbitraire. $\square$

Nous établirons plus loin (voir Théorème IV-2.38) ce résultat en utilisant d'autres outils basés sur la fonction caractéristique.

Il est possible de généraliser le Théorème IV-2.18 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-2.19 (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 $(\Omega, \mathcal{F}, \mathbb{P})$, i.i.d. et telles que $\mathbb{E}[X_1^+] < \infty$ ou $\mathbb{E}[X_1^-] < \infty$. Alors

$$\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{\mathbb{P}\text{-prob}} \mathbb{E}[X_1] .$$

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-2.18. 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-2.18, nous obtenons donc que

$$\frac{1}{n} \sum_{k=1}^n X_k \wedge c \xrightarrow{\mathbb{P}\text{-prob}} \mathbb{E}[X_1 \wedge c] .$$

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

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(n^{-1} \sum_{k=1}^n X_k \geq M \right) \geq \lim_{n \rightarrow \infty} \mathbb{P} \left(n^{-1} \sum_{k=1}^n (X_k \wedge c) \geq M \right) = 1 .$$

Comme M est arbitraire, ceci montre que $n^{-1} \sum_{k=1}^n X_k \xrightarrow{\mathbb{P}\text{-prob}} +\infty$. $\square$

Exemple IV-2.20 (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-2.34) d'une loi de Cauchy de paramètre $\theta = (\alpha, \sigma)$ est donnée par

$$\phi_\theta(t) = \exp(i\alpha t - \sigma|t|) .$$

Cela entraîne 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

$$\exp(i\alpha_1 t - \sigma_1|t|) \exp(i\alpha_2 t - \sigma_2|t|) = \exp(i(\alpha_1 + \alpha_2)t - (\sigma_1 + \sigma_2)|t|)$$

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. $\diamond$

IV-2.3.2 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-2.21 (Loi forte des grands nombres). Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, i.i.d. et intégrables. Alors

$$\bar{X}_n := n^{-1} \sum_{i=1}^n X_i \xrightarrow{\mathbb{P}\text{-p.s.}} \mathbb{E}[X_1] .$$

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-2.22. Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, intégrables, identiquement distribuées et de variance finie et non-corrélées. On a alors :

$$n^{-1} \sum_{i=1}^n X_i \xrightarrow{\mathbb{P}\text{-p.s.}} \mathbb{E}[X_1] .$$

Démonstration. Nous allons utiliser l'inégalité de Markov (Lemme IV-3.1) et le lemme de Borel-Cantelli (Lemme IV-2.11). 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

$$Y_n := \sum_{i=1}^n X_i, \quad Z_m := \sup_{1 \leq k \leq 2m+1} (|X_{m^2+1} + \dots + X_{m^2+k}|) .$$

Admettons avoir les propriétés suivantes : lorsque m tend vers l'infini,

$$\frac{Y_{m^2}}{m^2} \xrightarrow{\mathbb{P}\text{-p.s.}} 0 \tag{IV-2.9}$$

et

$$\frac{Z_m}{m^2} \xrightarrow{\mathbb{P}\text{-p.s.}} 0 . \tag{IV-2.10}$$

On utilise la majoration suivante :

$$\left| \frac{Y_n}{n} \right| \leq \left| \frac{Y_{m(n)^2}}{m(n)^2} \right| + \left| \frac{Z_{m(n)}}{m(n)^2} \right| ,$$

où $m(n)$ est l'entier m tel que

$$m^2 < n \leq (m+1)^2 .$$

Comme $\lim_{n \rightarrow \infty} m(n) = +\infty$, on conclut en utilisant (IV-2.9) et (IV-2.10).

Prouvons les deux relations admises ; considérons (IV-2.9). Pour un $\varepsilon > 0$ quelconque, l'inégalité de Bienayme-Tchebychev donne

$$\mathbb{P}(|Y_{m^2}/m^2| \geq \varepsilon) \leq \frac{\text{Var}(Y_{m^2})}{m^4 \varepsilon^2} = \frac{m^2 \sigma^2}{m^4 \varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \frac{1}{m^2} .$$

On conclut en utilisant l'inégalité de Borel-Cantelli. Considérons maintenant (IV-2.10). Posons $|\xi_{m,k}| = |X_{m^2+1} + \dots + X_{m^2+k}|$. Notons tout d'abord que

$$\mathbb{P}\left(\left|\frac{Z_m}{m^2}\right| \geq \varepsilon\right) = \mathbb{P}(|Z_m| \geq m^2 \varepsilon) \leq \mathbb{P}\left(\bigcup_{k=1}^{2m+1} \{|\xi_{m,k}| \geq m^2 \varepsilon\}\right) \leq \sum_{k=1}^{2m+1} \mathbb{P}(|\xi_{m,k}| \geq m^2 \varepsilon).$$

L'inégalité de Bienayme-Tchebychev nous donne d'autre part

$$\mathbb{P}(|\xi_{m,k}| \geq m^2 \varepsilon) \leq \frac{\text{Var}(\xi_{m,k})}{m^4 \varepsilon^2} \leq \frac{(2m+1)\sigma^2}{m^4 \varepsilon^2}$$

puisque $\xi_{m,k}$ est la somme de moins de $2m+1$ variables aléatoires. En combinant les trois dernières inégalités, on voit que

$$\mathbb{P}\left(\left|\frac{Z_m}{m^2}\right| \geq \varepsilon\right) \leq \frac{\sigma^2 (2m+1)^2}{\varepsilon^2 m^4}.$$

On conclut en utilisant encore le Lemme de Borel-Cantelli. $\square$

Exemple IV-2.23 (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

$$\rho := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\text{Var}(X) \text{Var}(Y)}},$$

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 $a + bX + cY = 1$ $\mathbb{P}$ -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

$$R_n := \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}, \quad \text{où } \bar{X}_n := \frac{1}{n} \sum_{k=1}^n X_k, \quad \bar{Y}_n := \frac{1}{n} \sum_{k=1}^n Y_k.$$

Nous allons démontrer que $R_n \xrightarrow{\mathbb{P}\text{-p.s.}} \rho$. 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, $C_n := n^{-1} \sum_{i=1}^n X_i Y_i \xrightarrow{\mathbb{P}\text{-p.s.}} \mathbb{E}[XY]$, $\bar{X}_n \xrightarrow{\mathbb{P}\text{-p.s.}} \mathbb{E}[X]$ et $\bar{Y}_n \xrightarrow{\mathbb{P}\text{-p.s.}} \mathbb{E}[Y]$. De façon similaire, on montre que $V_n := n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \xrightarrow{\mathbb{P}\text{-p.s.}} \text{Var}(X)$ et $W_n := n^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \xrightarrow{\mathbb{P}\text{-p.s.}} \text{Var}(Y)$. 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-2.13) montre que $R_n \xrightarrow{\mathbb{P}\text{-p.s.}} \rho$. $\diamond$

IV-2.4 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-2.24 (Convergence étroite). Soient $\{\mu_n, n \in \mathbb{N}\}$ une suite de probabilités sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ et μ une probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Nous dirons que la suite $\{\mu_n, n \in \mathbb{N}\}$ converge étroitement vers μ , ce que nous notons $\mu_n \xrightarrow{w} \mu$, si, pour toute fonction continue bornée $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \int f d\mu_n = \int f d\mu .$$

Définition IV-2.25 (Convergence en loi). Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$:

$$\lim_{n \rightarrow \infty} \mathbb{E}_n[f(X_n)] = \mathbb{E}[f(X)] .$$

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-2.26 (Théorème de Portmanteau). Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Les assertions suivantes sont équivalentes.

- (i) $X_n \Longrightarrow X$,
- (ii) Pour toute fonction f bornée et lipschitzienne, $\lim_{n \rightarrow \infty} \mathbb{E}_n[f(X_n)] = \mathbb{E}[f(X)]$,
- (iii) $\liminf_{n \rightarrow \infty} \mathbb{P}_n(X_n \in G) \geq \mathbb{P}(X \in G)$ pour tout ensemble ouvert G ,
- (iv) $\limsup_{n \rightarrow \infty} \mathbb{P}_n(X_n \in F) \leq \mathbb{P}(X \in F)$ 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

$$\lim_{n \rightarrow \infty} \mathbb{P}_n(X_n \in A) = \mathbb{P}(X \in A)$$

- (vi) Pour toute fonction continue et positive

$$\liminf_{n \rightarrow \infty} \mathbb{E}_n[f(X_n)] \geq \mathbb{E}[f(X)] .$$

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

$$\lim_{n \rightarrow \infty} F_{X_n}(x_1, \dots, x_d) = F_X(x_1, \dots, x_d)$$

en tout point $(x_1, \dots, x_d)$ où F_X est continue.

Exemple IV-2.27 (un exemple élémentaire). Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ 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, $\lim_{n \rightarrow \infty} \mathbb{E}[f(X_n)] = \lim_{n \rightarrow \infty} f(1/n) = f(0)$. La loi limite est donc δ_0 .

Les fonctions de répartition des variables aléatoires $X = 0$ et X_n sont respectivement

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}, \quad F_{X_n}(x) = \begin{cases} 0 & \text{si } x < 1/n \\ 1 & \text{si } x \geq 1/n \end{cases}.$$

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 $\lim_n F_{X_n}(0) \neq F_0(0)$. $\diamond$

Exemple IV-2.28 (Sanity check !). Pour $n \geq 1$, soit X_n une variable aléatoire de loi uniforme et à valeurs dans $\{1/n, 2/n, \dots, (n-1)/n, 1\}$. 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 $x \in [0, 1]$, $\lim_{n \rightarrow \infty} F_n(x) = x$; 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-2.26 item (v) montre que $X_n \Rightarrow \text{Unif}([0, 1])$.

On peut aussi prouver cette convergence en utilisant une autre caractérisation : pour toute fonction f continue sur $[0, 1]$, nous avons

$$\mathbb{E}[f(X_n)] = \frac{1}{n} \sum_{k=1}^n f(k/n) \rightarrow_{n \rightarrow \infty} \int_0^1 f(x) dx,$$

et le terme de droite se relit comme $\mathbb{E}[f(X)]$ où $X \sim \text{Unif}([0, 1])$. $\diamond$

Le théorème suivant donne les liens entre la convergence en loi et la convergence en probabilité.

Théorème IV-2.29. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Soient $\{X_n, n \in \mathbb{N}\}$ une suite de vecteurs aléatoires et X un vecteur aléatoire définis sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $\mathbb{R}^d$. Alors

- (i) $X_n \xrightarrow{\mathbb{P}\text{-prob}} X$ implique $X_n \Rightarrow X$,
- (ii) $X_n \xrightarrow{\mathbb{P}\text{-prob}} c$, où c est une constante, si et seulement si $X_n \Rightarrow c$.

Démonstration. (i) Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ Lipschitzienne bornée. Notons

$$\|f\|_\infty := \sup_x |f(x)| \quad \text{and} \quad |f|_{\text{Lip}} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|}.$$

Pour tout $\varepsilon > 0$,

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \leq \varepsilon |f|_{\text{Lip}} + 2\|f\|_\infty \mathbb{P}(\|X_n - X\| \geq \varepsilon).$$

Le second terme du membre de droite tend vers 0 et le premier peut être rendu arbitrairement petit. Donc, $\lim_{n \rightarrow \infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ pour toute fonction f Lipschitzienne bornée. Nous concluons en appliquant le théorème IV-2.26.

(ii) Soit $\varepsilon > 0$ et soit $B(c, \varepsilon) := \{x \in \mathbb{R}^d : \|x - c\| < \varepsilon\}$ la boule ouverte de centre c et de rayon ε . Nous avons $\mathbb{P}(\|X_n - c\| \geq \varepsilon) = \mathbb{P}(X_n \in B(c, \varepsilon)^c)$. Si $X_n \Rightarrow c$, le théorème IV-2.26 montre que $\limsup_n \mathbb{P}(X_n \in B(c, \varepsilon)^c) \leq \mathbb{P}(X \in B(c, \varepsilon)^c) = 0$. $\square$

IV-2.4.1 Opérations sur les limites en loi

Théorème IV-2.30 (Transformation continue). Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. 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 $X_n \Rightarrow X$ alors $g(X_n) \Rightarrow g(X)$.

Démonstration. Par définition, $\{g(X_n) \in F\} = \{X_n \in g^{-1}(F)\}$. Pour tout fermé F , on a :

$$g^{-1}(F) \subset \overline{g^{-1}(F)} \subset (g^{-1}(F) \cup C^c); \quad (\text{IV-2.11})$$

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 $\lim_{n \rightarrow \infty} x_n = x$. Si $x \in C$, $g(x_n) \rightarrow g(x)$, car g est continue au point x , et comme $g(x_n) \in F$ et F est fermé, $g(x) \in F$, ce qui implique que $x \in g^{-1}(F)$.

On déduit de (IV-2.11) et du théorème de Portmanteau (Théorème IV-2.26), comme $X_n \Rightarrow X$ et $\mathbb{P}(X \in C^c) = 0$:

$$\begin{aligned} \limsup \mathbb{P}(g(X_n) \in F) &\leq \limsup \mathbb{P}(X_n \in \overline{g^{-1}(F)}) \leq \mathbb{P}(X \in \overline{g^{-1}(F)}) \\ &\leq \mathbb{P}(X \in g^{-1}(F) \cup C^c) = \mathbb{P}(X \in g^{-1}(F)) = \mathbb{P}(g(X) \in F), \end{aligned}$$

et, donc, en appliquant de nouveau le théorème de Portmanteau, $g(X_n) \Rightarrow g(X)$. $\square$

La proposition IV-2.4 montre que la convergence en probabilité d'une suite de vecteurs aléatoires $X_n = (X_{n,1}, \dots, X_{n,k})$ 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-2.31. 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 $X_n \Rightarrow N(0, 1)$; et puisque pour tout n , Y_n suit aussi une loi normale centrée réduite, nous avons aussi $Y_n \Rightarrow N(0, 1)$. 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-2.30, le vecteur (X_n, Y_n) ne peut pas converger en loi. $\diamond$

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-2.32. Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n, Y_n des vecteurs aléatoires, définis sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeur respectivement dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ et $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

- (i) (cas $d = p$) Si $X_n \Rightarrow X$ et $X_n - Y_n \xrightarrow{\mathbb{P}\text{-prob}} 0$, alors $Y_n \Rightarrow X$.
- (ii) Soit $c \in \mathbb{R}^p$. Si $X_n \Rightarrow X$ et $Y_n \xrightarrow{\mathbb{P}\text{-prob}} c$, alors $(X_n, Y_n) \Rightarrow (X, c)$.

Démonstration. (i) Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ Lipschitzienne bornée, de constante de Lipschitz notée $|f|_{\text{Lip}}$. Pour tout $\varepsilon > 0$,

$$\begin{aligned} |\mathbb{E}_n[f(X_n)] - \mathbb{E}_n[f(Y_n)]| &= \left| \mathbb{E}_n \left[(f(X_n) - f(Y_n)) \left(\mathbb{1}_{\|X_n - Y_n\| \leq \varepsilon} + \mathbb{1}_{\|X_n - Y_n\| \geq \varepsilon} \right) \right] \right| \\ &\leq |f|_{\text{Lip}} \varepsilon + 2|f|_{\infty} \mathbb{P}_n(\|X_n - Y_n\| \geq \varepsilon). \end{aligned}$$

Le second terme tend vers 0 puisque $X_n - Y_n \xrightarrow{\mathbb{P}\text{-prob}} 0$ (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-2.26, on a $\lim_n \mathbb{E}_n[f(X_n)] = \mathbb{E}[f(X)]$ et $Y_n \Rightarrow X$.

(ii) Remarquons que

$$\left| \begin{bmatrix} X_n \\ Y_n \end{bmatrix} - \begin{bmatrix} X_n \\ c \end{bmatrix} \right| = \left| \begin{bmatrix} 0 \\ Y_n - c \end{bmatrix} \right| \xrightarrow{\mathbb{P}\text{-prob}} 0.$$

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 $(X_n, c) \implies (X, c)$ Pour toute fonction continue bornée $f : (x, y) \rightarrow f(x, y)$, la fonction $f(\cdot, c) : x \rightarrow f(x, c)$ est continue et bornée et $|\mathbb{E}[f(X_n, c)] - \mathbb{E}[f(X, c)]| \rightarrow 0$, car $X_n \implies X$; par suite, $(X_n, c) \implies (X, c)$. $\square$

La propriété (ii) et le théorème de continuité (théorème IV-2.30) montrent que pour toute fonction $g : (x, y) \rightarrow g(x, y)$ continue sur un ensemble $C \times \{c\}$ où C est tel que $\mathbb{P}(X \in C) = 1$, on a $g(X_n, Y_n) \implies g(X, c)$. 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-2.33 (Lemme de Slutsky). *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n, Y_n des v.a. réelles, définies sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X une v.a. réelle définie sur $(\Omega, \mathcal{F}, \mathbb{P})$.*

Si $X_n \implies X$ et $Y_n \implies c$ où c est une constante, alors

- (i) $X_n + Y_n \implies X + c$;
- (ii) $Y_n X_n \implies cX$;
- (iii) Si $c \neq 0$, $Y_n^{-1} X_n \implies c^{-1} X$.

Le procédé de Cramér-Wold (Théorème IV-2.37) 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-2.4.2 Caractérisation par la fonction caractéristique

Le théorème de Portmanteau (Théorème IV-2.26) 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-2.34 (Fonction caractéristique). *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ défini sur $(\Omega, \mathcal{F}, \mathbb{P})$. La fonction caractéristique du vecteur aléatoire X est définie par*

$$t = (t_1, \dots, t_d) \rightarrow \mathbb{E}[e^{it^T X}] .$$

Exemple IV-2.35 (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 $N(\mu, \Sigma)$. Alors X a même loi que $\mu + \sqrt{\Sigma}Y$ où $Y \sim N(0, I)$, et $\sqrt{\Sigma}$ désigne une matrice $d \times d$ satisfaisant $\sqrt{\Sigma}\sqrt{\Sigma}^T = \Sigma$.

En dimension $d = 1$, la fonction caractéristique d'une loi $N(0, 1)$ est donnée par (voir Lemme ??) :

$$\mathbb{E}[e^{itX}] = \exp\left(-\frac{1}{2}t^2\right) . \quad (\text{IV-2.12})$$

Dans le cas $X \sim N(\mu, \Sigma)$, remarquons que

$$\phi_X(t) = \mathbb{E}[e^{it^T X}] = \mathbb{E}[e^{it^T (\mu + \sqrt{\Sigma}Y)}] = e^{it^T \mu} \mathbb{E}[e^{i\sqrt{\Sigma}^T t^T Y}] .$$

Or Y a ses composantes indépendantes de loi $N(0, 1)$ donc pour tout $t \in \mathbb{R}^d$

$$\mathbb{E}[e^{it^T Y}] = \prod_{k=1}^d \mathbb{E}[e^{it_k^T Y_k}] = \exp\left(-\frac{1}{2}|t|^2\right).$$

On en déduit que

$$\phi_X(t) = \exp\left(it^T \mu\right) \exp\left(-\frac{1}{2}t^T \Sigma t\right).$$

◇

Pour tout $t \in \mathbb{R}^d$, la fonction $x \rightarrow e^{it^T x}$ est continue et bornée. Par conséquent, si $X_n \Rightarrow X$, $\mathbb{E}_n[e^{it^T X_n}] \rightarrow \mathbb{E}[e^{it^T X}]$. Le théorème de Levy montre que la réciproque est vraie.

Théorème IV-2.36 (Lévy). Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Alors

1. $X_n \Rightarrow X$ si et seulement si $\mathbb{E}_n[\exp(it^T X_n)] \rightarrow \mathbb{E}[\exp(it^T X)]$ 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 $\{\mathbb{E}_n[\exp(it^T X_n)], n \in \mathbb{N}\}$ converge vers $\phi(t)$, alors ϕ est la fonction caractéristique d'un vecteur aléatoire X à valeurs dans $\mathbb{R}^d$ et $X_n \Rightarrow X$.

Démonstration. Voir par exemple [?, Théorème 26.3 et son Corollaire 1].

□

La fonction caractéristique d'un vecteur aléatoire $X = (X_1, \dots, X_d)$ 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 : $\psi(t) = \mathbb{E}[e^{it^T X}] = \phi(1)$ où $\phi : u \in \mathbb{R} \rightarrow \mathbb{E}[e^{iuY}]$.

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-2.37 (Cramér-Wold). Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. $X_n \Rightarrow X$ si et seulement si, pour tout $t \in \mathbb{R}^d$, $t^T X_n \Rightarrow t^T X$.

Démonstration. Supposons que $X_n \Rightarrow X$. Alors, pour tout $t \in \mathbb{R}^d$ et tout $u \in \mathbb{R}$, $\mathbb{E}_n[e^{iu t^T X_n}] \rightarrow \mathbb{E}[e^{iu t^T X}]$, et donc $t^T X_n \Rightarrow t^T X$ par application du théorème de Levy (théorème IV-2.36).

Réciproquement, supposons que, pour tout $t \in \mathbb{R}^d$, $t^T X_n \Rightarrow t^T X$. Alors $\mathbb{E}_n[e^{i(t^T X_n)}] \rightarrow \mathbb{E}[e^{i(t^T X)}]$ et donc $X_n \Rightarrow X$, encore par application du théorème de Levy. □

Le Théorème IV-2.36 permet d'obtenir une autre preuve de la loi faible des grands nombres que celle donnée dans la démonstration du théorème IV-2.18.

Théorème IV-2.38 (Loi faible des grands nombres). Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, i.i.d. et intégrables. Alors

$$n^{-1} \sum_{i=1}^n X_i \xrightarrow{\mathbb{P}\text{-prob}} \mathbb{E}[X_1].$$

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

$$\psi_n(t) = \mathbb{E} \left[\exp \left(itn^{-1} \sum_{k=1}^n X_k \right) \right] = \prod_{k=1}^n \mathbb{E}[\exp(itn^{-1} X_k)] = (\phi(n^{-1}t))^n.$$

Comme $\mathbb{E}[|X_1|] < \infty$ existe, la fonction ϕ est dérivable en 0 et

$$[\phi(n^{-1}t)]^n = \left(1 + \frac{it}{n} \mu + o(n^{-1}) \right)^n \xrightarrow{n \rightarrow \infty} e^{it\mu}; \quad (\text{IV-2.13})$$

(la dérivabilité de la fonction ϕ s'établit par application du théorème de convergence dominée, voir Théorème IV-1.41 et Proposition IV-1.43 ; pour ceux qui ne sont pas familiers du passage à la limite (IV-2.13), on peut l'établir en utilisant les mêmes arguments que ceux de la démonstration du Théorème IV-2.39).

Dans (IV-2.13), le membre de droite est la fonction caractéristique de la variable aléatoire qui vaut μ avec probabilité 1. Le théorème de Levy IV-2.36 montre que $\bar{X}_n \Rightarrow \mu$ et donc $\bar{X}_n \xrightarrow{\mathbb{P}\text{-prob}} \mu$ par le Théorème IV-2.29. $\square$

IV-2.5 Théorème de la limite centrale

IV-2.5.1 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-2.39 (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 $(\Omega, \mathcal{F}, \mathbb{P})$, i.i.d. et possédant un moment d'ordre 2. On note μ l'espérance et Σ la matrice de covariance. Alors :

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \Rightarrow N(0, \Sigma),$$

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 $\sqrt{n}(\bar{X}_n - \mu)$ où $\bar{X}_n := n^{-1} \sum_{k=1}^n X_k$. On remarque que

$$\begin{aligned} \mathbb{P}(\mu - a/\sqrt{n} < \bar{X}_n \leq \mu + a/\sqrt{n}) &= \mathbb{P}(-a < \sqrt{n}(\bar{X}_n - \mu) \leq a) \\ &= F_n(a) - F_n(-a) \\ &\rightarrow \Phi(a/\sigma) - \Phi(-a/\sigma), \end{aligned}$$

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-2.39) D'après le procédé de Cramér-Wold (Théorème IV-2.37), il suffit de montrer que pour tout $t \in \mathbb{R}^d$,

$$\sqrt{nt}^T Z_n \Rightarrow t^T W, \quad \text{où} \quad W \sim N(0, \Sigma), \quad Z_n := n^{-1} \sum_{k=1}^n (X_k - \mu). \quad (\text{IV-2.14})$$

Notons que Z_n est un vecteur aléatoire centré et de matrice de covariance Σ , de sorte que $\sqrt{nt}^T Z_n$ 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 $n \geq 1$ $\sqrt{nt}^T Z_n$ est une v.a. réelle constante, égale à son espérance donc nulle. Par suite, $\sqrt{nt}^T Z_n \Rightarrow 0$ ce qui est bien équivalent à (IV-2.14) 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 $i = 1, \dots, n$,

$$Y_i := \sigma^{-1}(X_i - \mu), \quad \bar{Y}_n := n^{-1} \sum_{i=1}^n Y_i.$$

Les variables aléatoires $\{Y_n, n \in \mathbb{N}\}$ sont de moyenne nulle et de variance unité : $\mathbb{E}[Y_i] = 0$ et $\text{Var}(Y_i) = 1$. 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 :

$$\lim_{n \rightarrow \infty} \phi_n(t) = \exp(-t^2/2);$$

nous conclurons en utilisant le théorème de Lévy et le résultat (IV-2.12).

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 $t \in \mathbb{R}$,

$$[\psi(n^{-1/2}t)]^n \rightarrow_{n \rightarrow \infty} e^{-t^2/2}.$$

Dans la suite, $t \in \mathbb{R}$ est fixé. Remarquons tout d'abord que

$$\begin{aligned} \left| [\psi(n^{-1/2}t)]^n - e^{-t^2/2} \right| &= \left| [\psi(n^{-1/2}t)]^n - [e^{-t^2/(2n)}]^n \right| \\ &\leq n \left| \psi(n^{-1/2}t) - e^{-t^2/(2n)} \right|, \end{aligned}$$

car $|\psi(n^{-1/2}t)| \leq 1$ et $e^{-t^2/n} \leq 1$. Nous avons, en utilisant l'inégalité triangulaire

$$n|\psi(t/\sqrt{n}) - e^{-t^2/2n}| \leq n|\psi(t/\sqrt{n}) - (1 - t^2/2n)| + n|(1 - t^2/2n) - e^{-t^2/2n}|.$$

Nous allons utiliser deux inégalités élémentaires

$$|e^{iu} - 1 - iu| \leq |u|^2/2 \quad (\text{IV-2.15})$$

$$|e^{iu} - 1 - iu - (iu)^2/2| \leq |u|^3/6. \quad (\text{IV-2.16})$$

En posant $u = it^2/2n \geq 0$, nous obtenons en utilisant (IV-2.15)

$$n|(1 - t^2/2n) - e^{-t^2/2n}| \leq n(t^2/2n)^2/2 = t^4/8n \rightarrow_{n \rightarrow \infty} 0.$$

Comme $\mathbb{E}[Y] = 0$ et $\text{Var}(Y) = \mathbb{E}[Y^2] = 1$, nous avons

$$\begin{aligned} n|\psi(t/\sqrt{n}) - (1 - t^2/2n)| &= n|\mathbb{E}[e^{itY/\sqrt{n}} - (1 + itY/\sqrt{n} + i^2 t^2 Y^2/2n)]| \\ &\leq n\mathbb{E}[|e^{itY/\sqrt{n}} - (1 + itY/\sqrt{n} + i^2 t^2 Y^2/2n)|]. \end{aligned}$$

En utilisant (IV-2.15) avec $u = tY/\sqrt{n}$, nous avons

$$\begin{aligned} |e^{itY/\sqrt{n}} - (1 + itY/\sqrt{n} + i^2 t^2 Y^2/2n)| &\leq |e^{itY/\sqrt{n}} - (1 + itY/\sqrt{n})| + t^2 Y^2/2n \\ &\leq t^2 Y^2/2n + t^2 Y^2/2n \\ &= t^2 Y^2/n. \end{aligned}$$

D'autre part, en utilisant cette fois l'inégalité (IV-2.16), nous obtenons

$$|e^{itY/\sqrt{n}} - (1 + itY/\sqrt{n} + i^2 t^2 Y^2/2n)| \leq |t|^3 |Y|^3 / 6n^{3/2}.$$

Par suite, pour tout ensemble A , on a

$$|e^{itY/\sqrt{n}} - (1 + itY/\sqrt{n} + i^2 t^2 Y^2/2n)| \leq (t^2 Y^2/n) \mathbb{1}_A + (|t|^3 |Y|^3/6n^{3/2}) \mathbb{1}_{A^c}.$$

Pour tout $\delta > 0$ et $n \in \mathbb{N}$, posons $A = A(\delta, n) := \{|Y| > \delta n^{1/6}\}$. Il vient pour tout $n \geq 1$,

$$\begin{aligned} n\mathbb{E}[|e^{itY/\sqrt{n}} - (1 + itY/\sqrt{n} + i^2 t^2 Y^2/2n)|] \\ \leq n\mathbb{E}[(t^2 Y^2/n) \mathbb{1}_A] + n\mathbb{E}[|t|^3 |Y|^3/6n^{3/2}) \mathbb{1}_{A^c}] \\ \leq t^2 \mathbb{E}[Y^2 \mathbb{1}_{\{|Y| > \delta n^{1/6}\}}] + |t|^3 \delta^3/6. \end{aligned}$$

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

$$t^2 \mathbb{E}[Y^2 \mathbb{1}_{\{|Y| > \delta n^{1/6}\}}] \leq \varepsilon/2. \quad \square$$

Exemple IV-2.40 (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

$$\bar{Y}_n := n^{-1} \sum_{k=1}^n Y_k, \quad S_n^2 := n^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$$

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-2.18) et du théorème de continuité (Théorème IV-2.6), nous avons :

$$S_n^2 = n^{-1} \sum_{i=1}^n Y_i^2 - (\bar{Y}_n)^2 \xrightarrow{\mathbb{P}\text{-prob}} (\mathbb{E}Y_1^2 - (\mathbb{E}Y_1)^2) = \text{Var}(Y_1).$$

Le théorème de continuité implique aussi que $S_n = \sqrt{S_n^2}$ converge en probabilité vers $\sqrt{\text{Var}(Y_1)} = \sigma$. Le théorème de la limite centrale (théorème IV-2.39) montre que $\sqrt{n}\bar{Y}_n \Rightarrow N(0, \sigma^2)$ et le lemme de Slutsky (Lemme IV-2.33) implique que $T_n \Rightarrow N(0, 1)$. $\diamond$

IV-2.5.2 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-2.41 (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 $(\Omega, \mathcal{F}, \mathbb{P})$, indépendants, centrés et tels que $\mathbb{E}[\|Y_{n,i}\|^2] < \infty$ pour $i \in \{1, \dots, k_n\}$. Supposons les conditions de Lindeberg–Feller vérifiées :

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{k_n} \mathbb{E}[\|Y_{n,i}\|^2 \mathbb{1}_{\{\|Y_{n,i}\| > \varepsilon\}}] = 0, \quad \text{pour tout } \varepsilon > 0, \quad (\text{IV-2.17})$$

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{k_n} \text{Var}(Y_{n,i}) = \Sigma. \quad (\text{IV-2.18})$$

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-3.1), il est aisé de démontrer qu'une condition suffisante pour (IV-2.17) est

$$\exists \delta > 0, \quad \lim_{n \rightarrow \infty} \sum_{i=1}^{k_n} \mathbb{E}[|Y_{n,i}|^{2+\delta}] = 0. \quad (\text{IV-2.19})$$

Démonstration. Par le procédé de Cramér-Wold (Théorème IV-2.37), 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 $i \in \{1, \dots, k_n\}$, $\text{Var}(Y_{n,i}) = \text{Var}(X_{n,i})$,
- 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.*,

$$|f''|_{\text{Lip}} := \sup_{(x,y) \in \mathbb{R} \times \mathbb{R}, x \neq y} \frac{|f''(x) - f''(y)|}{|x - y|} < \infty, \quad (\text{IV-2.20})$$

où f'' est la dérivée seconde de f . On a la décomposition

$$\mathbb{E}[f(R_n)] - \mathbb{E}[f(T_n)] = \sum_{k=1}^{k_n} (\mathbb{E}[f(R_{n,k} + X_{n,k})] - \mathbb{E}[f(R_{n,k} + Y_{n,k})]),$$

où $R_{n,k} := (\sum_{j < k} X_{n,j}) + (\sum_{j > k} Y_{n,j})$. Développons $f(R_{n,k} + X_{n,k})$ au voisinage de $R_{n,k}$:

$$f(R_{n,k} + X_{n,k}) = f(R_{n,k}) + X_{n,k} f'(R_{n,k}) + \frac{X_{n,k}^2}{2} f''(R_{n,k}) + \frac{X_{n,k}^2}{2} [f''(R_{n,k} + \theta_{n,k} X_{n,k}) - f''(R_{n,k})],$$

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 $k \in \{1, \dots, n\}$,

$$\mathbb{E}[f'(R_{n,k})(X_{n,k} - Y_{n,k})] = \mathbb{E}[f'(R_{n,k})] (\mathbb{E}[X_{n,k}] - \mathbb{E}[Y_{n,k}]) = 0,$$

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

$$\mathbb{E}[f''(R_{n,k})(X_{n,k}^2 - Y_{n,k}^2)] = \mathbb{E}[f''(R_{n,k})] (\mathbb{E}[X_{n,k}^2] - \mathbb{E}[Y_{n,k}^2]) = 0.$$

Comme f'' est une fonction Lipschitzienne, pour tout $\varepsilon > 0$, nous avons, pour tout $\theta \in [0, 1]$,

$$\begin{aligned} |f''(R_{n,k} + \theta X_{n,k}) - f''(R_{n,k})| &\leq |f''|_{\text{Lip}} |X_{n,k}|, \\ |f''(R_{n,k} + \theta Y_{n,k}) - f''(R_{n,k})| &\leq |f''|_{\text{Lip}} |Y_{n,k}| \mathbb{1}_{\{|Y_{n,k}| \leq \varepsilon\}} + 2|f''|_{\infty} \mathbb{1}_{\{|Y_{n,k}| > \varepsilon\}}, \end{aligned}$$

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 $\varepsilon > 0$,

$$\begin{aligned} &|\mathbb{E}[f(R_{n,k} + X_{n,k})] - \mathbb{E}[f(R_{n,k} + Y_{n,k})]| \\ &\leq \frac{1}{2} |f''|_{\text{Lip}} (\mathbb{E}[|X_{n,k}|^3] + \mathbb{E}[|Y_{n,k}| \mathbb{1}_{\{|Y_{n,k}| \leq \varepsilon\}}]) + |f''|_{\infty} \mathbb{E}[|Y_{n,k}| \mathbb{1}_{\{|Y_{n,k}| > \varepsilon\}}]. \end{aligned}$$

Remarquons que $\mathbb{E}[|X_{n,k}|^3] = \sigma_{n,k}^3 m_3$ où $m_3 := \mathbb{E}[|Z|^3]$ avec $Z \sim N(0, 1)$ et $\sigma_{n,k}^2 := \mathbb{E}[X_{n,k}^2] = \mathbb{E}[Y_{n,k}^2]$. Remarquons aussi que

$$\mathbb{E}[|Y_{n,k}|^3 \mathbb{1}_{\{|Y_{n,k}| \leq \varepsilon\}}] \leq \varepsilon \mathbb{E}[|Y_{n,k}|^2] = \varepsilon \sigma_{n,k}^2.$$

Ces inégalités conduisent à la majoration :

$$|\mathbb{E}f(R_n) - \mathbb{E}f(T_n)| \leq \frac{1}{2} |f''|_{\text{Lip}} \left(m_3 \sum_{k=1}^{k_n} \sigma_{n,k}^3 + \varepsilon \sum_{k=1}^{k_n} \sigma_{n,k}^2 \right) + |f''|_{\infty} \sum_{k=1}^{k_n} \mathbb{E}[Y_{n,k}^2 \mathbb{1}_{\{|Y_{n,k}| > \varepsilon\}}].$$

On a d'autre part

$$\sum_{k=1}^{k_n} \sigma_{n,k}^3 \leq \sum_{k=1}^{k_n} \sigma_{n,k}^2 \max_{k \in \{1, \dots, k_n\}} (\sigma_{n,k}).$$

Or (IV-2.18) implique que pour tout $1 \leq k \leq k_n$, et tout $\varepsilon > 0$, nous avons

$$\sigma_{n,k}^2 \leq \varepsilon^2 + \mathbb{E}[Y_{n,k}^2 \mathbb{1}_{\{|Y_{n,k}| > \varepsilon\}}] \leq \varepsilon^2 + \sum_{j=1}^{k_n} \mathbb{E}[Y_{n,j}^2 \mathbb{1}_{\{|Y_{n,j}| > \varepsilon\}}].$$

et donc

$$\limsup_{n \rightarrow \infty} \max_{k \in \{1, \dots, k_n\}} \sigma_{n,k}^2 = 0.$$

D'où, pour tout $\varepsilon > 0$,

$$\begin{aligned} |\mathbb{E}f(R_n) - \mathbb{E}f(T_n)| &\leq \frac{1}{2} |f''|_{\text{Lip}} \left(m_3 \sum_{k=1}^{k_n} \sigma_{n,k}^2 \left(\varepsilon^2 + \sum_{j=1}^{k_n} \mathbb{E}[Y_{n,j}^2 \mathbb{1}_{\{|Y_{n,j}| > \varepsilon\}}] \right)^{1/2} + \varepsilon \sum_{k=1}^{k_n} \sigma_{n,k}^2 \right) \\ &\quad + |f''|_{\infty} \sum_{k=1}^{k_n} \mathbb{E}[Y_{n,k}^2 \mathbb{1}_{\{|Y_{n,k}| > \varepsilon\}}]. \end{aligned}$$

Par hypothèse, quand n tend vers l'infini, la série $\sum_k \mathbb{E}[Y_{n,k}^2 \mathbb{1}_{\{|Y_{n,k}| > \varepsilon\}}]$ tend vers zéro et $\sum_k \sigma_{n,k}^2$ tend vers une constante σ^2 . D'où, pour tout $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} |\mathbb{E}f(R_n) - \mathbb{E}f(T_n)| \leq \frac{1}{2} |f''|_{\text{Lip}} (m_3 + 1) \sigma^2 \varepsilon.$$

En faisant tendre, dans un deuxième temps, ε vers zéro, on obtient donc une limite nulle. Comme les $(X_{n,j}, j \in \{1, \dots, k_n\})$ 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 $\sum_{i=1}^{k_n} \text{Var}(X_{n,i})$. Puisque la fonction caractéristique de R_n vaut

$$t \mapsto \exp \left(-\frac{1}{2} \left(\sum_{i=1}^{k_n} \text{Var}(X_{n,i}) \right) t^2 \right) = \exp \left(-\frac{1}{2} \left(\sum_{i=1}^{k_n} \text{Var}(Y_{n,i}) \right) t^2 \right)$$

elle converge vers $\exp(-\sigma^2 t^2 / 2)$ 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-2.36). $\square$

Exemple IV-2.42. 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

$$s_n^2 := \sum_{i=1}^n \text{Var}(Y_i) = \sum_{i=1}^n p_i(1 - p_i).$$

Par construction, $\lim_{n \rightarrow \infty} s_n = \infty$. Posons $Y_{n,i} = s_n^{-1}(Y_i - p_i)$. Clairement, $\sum_{i=1}^n \text{Var}(Y_{n,i}) = 1$. D'autre part, nous avons

$$\mathbb{E}[|Y_{n,i}|^3] = s_n^{-3} \left\{ p_i^3(1-p_i) + (1-p_i)^3 p_i \right\} \leq 2s_n^{-3} p_i(1-p_i),$$

ce qui implique

$$\sum_{i=1}^n \mathbb{E}[|Y_{n,i}|^3] \leq \frac{2}{s_n^3} \sum_{i=1}^n p_i(1-p_i) = \frac{2}{s_n} \rightarrow 0,$$

car $\lim_{n \rightarrow \infty} s_n = \infty$. Par conséquent la condition (IV-2.17) est satisfaite avec $\varepsilon = 1$. Le Théorème IV-2.41 montre que

$$s_n^{-1} \sum_{i=1}^n (Y_i - p_i) \Rightarrow N(0, 1).$$

Exemple IV-2.43 (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 $n \geq 0$, $\sum_{i=1}^n c_{n,i}^2 > 0$ et

$$\delta_n := \max_{1 \leq i \leq n} \frac{c_{n,i}^2}{\sum_{j=1}^n c_{n,j}^2} \rightarrow 0. \quad (\text{IV-2.21})$$

Alors, nous avons

$$\frac{\sum_{i=1}^n c_{n,i}(X_i - \mu)}{\sigma \sqrt{\sum_{i=1}^n c_{n,i}^2}} \Rightarrow N(0, 1). \quad (\text{IV-2.22})$$

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, $\sum_{i=1}^n \text{Var}(Y_{n,i}) = 1$. Nous avons de plus

$$\begin{aligned} \mathbb{E}[Y_{n,i}^2 \mathbb{1}_{\{|Y_{n,i}| > \varepsilon\}}] &= s_n^{-2} c_{n,i}^2 \mathbb{E}[(X_1 - \mu)^2 \mathbb{1}_{\{c_{n,i}|X_1 - \mu| > \varepsilon s_n\}}] \\ &\leq c_{n,i}^2 \mathbb{E}[(X_1 - \mu)^2 \mathbb{1}_{\{\sqrt{\delta_n}|X_1 - \mu| > \varepsilon \sigma\}}]. \end{aligned}$$

Par conséquent, par le théorème de convergence dominée (Théorème IV-1.41), nous avons

$$\frac{1}{s_n^2} \sum_{i=1}^n c_{n,i}^2 \mathbb{E}[(X_1 - \mu)^2 \mathbb{1}_{\{\sqrt{\delta_n}|X_1 - \mu| > \varepsilon \sigma\}}] = \frac{\mathbb{E}[(X_1 - \mu)^2 \mathbb{1}_{\{\sqrt{\delta_n}|X_1 - \mu| > \varepsilon \sigma\}}]}{\sigma^2} \rightarrow 0.$$

De façon intuitive, la condition (IV-2.21) montre que, lorsque $n \rightarrow \infty$, la contribution de chaque terme dans la somme

$$\sum_{i=1}^n \frac{c_{n,i}(X_i - \mu)}{\sigma \sqrt{\sum_{j=1}^n c_{n,j}^2}}, \quad (\text{IV-2.23})$$

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-2.23) est toujours égale à $(X_1 - \mu)/\sigma$ et il n'y a bien évidemment pas de passage à la limite. $\diamond$

IV-2.5.3 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

$$\bar{X}_n := n^{-1} \sum_{k=1}^n X_k, \quad Y_n := \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}.$$

Le théorème de la limite centrale montre que $Y_n \Rightarrow N(0, 1)$ ce qui implique en particulier que pour tout $x \in \mathbb{R}$, $\lim_{n \rightarrow \infty} F_n(x) = \Phi(x)$ (voir Théorème IV-2.26) où F_n (resp. Φ) désigne la fonction de répartition de Y_n (resp. d'une loi $N(0, 1)$). Comme la fonction $x \mapsto \Phi(x)$ est continue, le théorème de Polya (Théorème IV-2.44) 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-2.44 (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 $X_n \Rightarrow X$, alors

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0.$$

Démonstration. La fonction F étant continue, il existe des points $-\infty = x_0 < x_1 < \dots < x_k = \infty$ 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 $x_{i-1} \leq x \leq x_i$:

$$\begin{aligned} F_n(x) - F(x) &\leq F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + 1/k \\ F_n(x) - F(x) &\geq F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - 1/k. \end{aligned}$$

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,

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \lim_{n \rightarrow \infty} \sup_{i \in \{0, \dots, k\}} |F_n(x_i) - F(x_i)| + 1/k = 1/k,$$

ce qui permet de conclure, en choisissant k arbitrairement grand. $\square$

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-2.45 (Berry-Esseen). Soient $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires réelles définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, 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 ,

$$\left| \mathbb{P} \left(\sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \leq x \right) - \Phi(x) \right| \leq \frac{C}{\sqrt{n}} \frac{\mathbb{E}[|X_1 - \mu|^3]}{\sigma^3}, \quad (\text{IV-2.24})$$

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

Démonstration. Voir [?, Chapitre XVI, Section V, Théorème 1] dans le cas où $C = 3$. $\square$

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 $C < 0.4097$ [?].

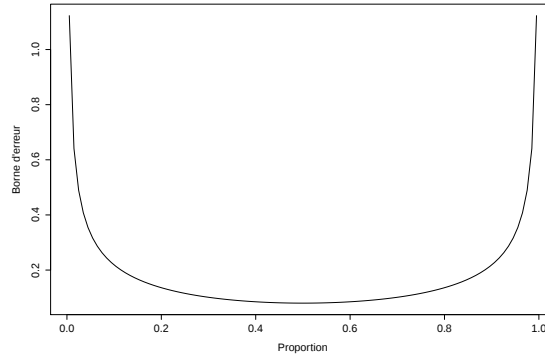
Si $\mathbb{E}[|X_1 - \mu|^3]/\sigma^3 < \infty$, le terme de droite de (IV-2.24) tend vers 0 et donc le terme de gauche tend uniformément vers 0 en x . Si la loi de la variable aléatoire X_1 appartient à une famille de distributions telle que

$$\frac{\mathbb{E}[|X_1 - \mu|^3]}{\sigma^3} \leq B,$$

pour $B < \infty$, alors la convergence est aussi uniforme par rapport à la loi de X_1 !

Exemple IV-2.46 (Erreur d'approximation pour une loi de Bernoulli). Soit $(X_1, \dots, X_n)$ 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-2.1 – 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 $\theta \in \Theta$,

$$|\mathbb{P}_\theta(\sqrt{n}(\bar{X}_n - \theta)/\sigma_\theta \leq x) - \Phi(x)| \leq \frac{0.8}{\sqrt{n}} \frac{\sigma_\theta(1 - 2\sigma_\theta)}{\sigma_\theta^{3/2}}.$$

Nous avons représenté cette borne dans la Figure IV-2.1 pour $n = 100$. Nous voyons que l'erreur d'approximation est minimale quand $\theta = 1/2$ mais que l'erreur est grande quand $\theta \rightarrow 0$ ou $\theta \rightarrow 1$, 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. $\diamond$

IV-2.5.4 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-2.6). 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 $\sqrt{n}\{T_n - \theta\} \Rightarrow T$, alors $\sqrt{n}\{g(T_n) - g(\theta)\} \Rightarrow g'(\theta)T$. En particulier, si $\sqrt{n}(T_n - \theta) \Rightarrow N(0, \sigma^2(\theta))$, alors $\sqrt{n}\{g(T_n) - g(\theta)\} \Rightarrow N(0, [g'(\theta)]^2 \sigma^2(\theta))$. La même question se pose lorsque $\mathbf{T}_n = (T_{n,1}, \dots, T_{n,p})$ est un vecteur aléatoire et $g = (g_1, \dots, g_m)$ 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 $J_g(\theta)$ et définie par

$$J_g(\theta) := \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_1}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial \theta_1} & \cdots & \frac{\partial g_m}{\partial \theta_p} \end{pmatrix}. \quad (\text{IV-2.25})$$

Théorème IV-2.47 (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 $\lim_{n \rightarrow \infty} r_n = \infty$. Alors

$$r_n \{g(\mathbf{T}_n) - g(\theta)\} - r_n \mathbf{J}_g(\theta)(\mathbf{T}_n - \theta) \xrightarrow{\mathbb{P}\text{-prob}} 0.$$

et donc

$$r_n \{g(\mathbf{T}_n) - g(\theta)\} \Longrightarrow \mathbf{J}_g(\theta) \mathbf{G}.$$

Démonstration. Comme $\{r_n(\mathbf{T}_n - \theta), n \in \mathbb{N}\}$ converge en loi, par le lemme IV-2.33 nous avons

$$\mathbf{T}_n - \theta = r_n^{-1} r_n \{\mathbf{T}_n - \theta\} \xrightarrow{\mathbb{P}\text{-prob}} 0.$$

La différentiabilité de la fonction g au point $\theta \in D_g$ implique que

$$g(\theta + h) = g(\theta) + \mathbf{J}_g(\theta)h + R(h)$$

où $\lim_{\|h\| \rightarrow 0} \|R(h)\| / \|h\| = 0$ et $R(0) = 0$. Il existe donc une fonction $\tilde{R}(h)$, vérifiant $R(h) = \|h\| \tilde{R}(h)$, continue en zéro et telle que $\tilde{R}(0) = 0$. On a donc

$$r_n (g(\mathbf{T}_n) - g(\theta)) - \mathbf{J}_g(\theta) \{r_n(\mathbf{T}_n - \theta)\} = r_n \|\mathbf{T}_n - \theta\| \tilde{R}(\mathbf{T}_n - \theta).$$

Puisque $\mathbf{T}_n - \theta \xrightarrow{\mathbb{P}\text{-prob}} 0$, $\tilde{R}(\mathbf{T}_n - \theta) \xrightarrow{\mathbb{P}\text{-prob}} 0$; de plus, par le Théorème IV-2.30, $r_n \|\mathbf{T}_n - \theta\| \Longrightarrow \|G\|$. Nous concluons en appliquant le lemme de Slutsky (Lemme IV-2.33). $\square$

Exemple IV-2.48. Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires telles que

$$r_n(X_n - \mu) \Longrightarrow N(0, \sigma^2)$$

où $\{r_n, n \in \mathbb{N}\}$ est une suite à termes positifs et croissante telle que $\lim_{n \rightarrow \infty} r_n = \infty$. 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 $X_n \xrightarrow{\mathbb{P}\text{-prob}} \mu$ et le théorème de continuité (théorème IV-2.6) implique que :

$$X_n^2 \xrightarrow{\mathbb{P}\text{-prob}} \mu^2, \quad \exp(X_n) \xrightarrow{\mathbb{P}\text{-prob}} \exp(\mu),$$

et si $\mu \neq 0$

$$\frac{1}{X_n} \xrightarrow{\mathbb{P}\text{-prob}} \frac{1}{\mu}, \quad \log |X_n| \xrightarrow{\mathbb{P}\text{-prob}} \log |\mu|.$$

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-2.47 montre que, pour $\mu \neq 0$,

$$r_n(X_n^2 - \mu^2) \Longrightarrow N(0, 4\mu^2\sigma^2).$$

Pour $\mu = 0$, $g'(\mu) = 0$ et l'application de théorème IV-2.47 montre que

$$r_n X_n^2 \xrightarrow{\mathbb{P}\text{-prob}} 0.$$

On peut néanmoins obtenir un résultat plus précis, en exhibant une vitesse de convergence, en appliquant le théorème de continuité (théorème IV-2.30). En effet,

$$r_n (r_n X_n^2) = (r_n X_n)^2 \Longrightarrow \sigma^2 X^2$$

où X est une variable aléatoire gaussienne $N(0, 1)$. Par conséquent, X^2 suit une loi du χ^2 centrée à un degré de liberté.

- Pour tout μ , le théorème IV-2.47 appliqué à la fonction $g(x) = e^x$ montre que $r_n(\exp(X_n) - \exp(\mu)) \Rightarrow N(0, \exp(2\mu)\sigma^2)$.
- 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-2.47 à la fonction $g(x) = 1/x$ qui est différentiable au point $\mu \neq 0$ montre que

$$r_n(1/X_n - 1/\mu) \Rightarrow N(0, \sigma^2/\mu^4).$$

Le cas $\mu = 0$ n'est pas couvert par le théorème IV-2.47, mais on peut dans ce contexte encore utiliser le théorème de continuité (Théorème IV-2.30) qui montre que

$$1/(r_n X_n) \Rightarrow 1/(\sigma X),$$

où X est une variable aléatoire gaussienne $N(0, 1)$.

- Nous supposons ici encore que pour tout $n \in \mathbb{N}$, $\mathbb{P}(X_n = 0) = 1$. Pour $\mu > 0$, le théorème IV-2.47 appliqué avec $g(x) = \log(|x|)$ ($g'(\mu) = 1/\mu$) montre que

$$r_n\{\log(|X_n|) - \log(|\mu|)\} \Rightarrow N(0, \sigma^2/\mu^2). \quad (\text{IV-2.26})$$

Pour $\mu < 0$, nous pouvons appliqué encore le théorème IV-2.47. Dans ce cas, $g(\mu) = -1/\mu$ et donc nous avons encore (IV-2.26).

Le cas $\mu = 0$ n'est pas couvert par le théorème IV-2.47, mais le théorème de continuité montre que $\log |r_n X_n| \Rightarrow \log |N(0, \sigma^2)|$. $\diamond$

Exemple IV-2.49 (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 $k = 1, \dots, 4$. Considérons l'estimateur de la variance empirique

$$S_n^2 := n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X}_n)^2, \quad \text{où } \bar{X}_n := n^{-1} \sum_{i=1}^n X_i.$$

La version multidimensionnelle du théorème de la limite centrale montre que :

$$\sqrt{n} \left(\begin{bmatrix} n^{-1} \sum_{i=1}^n X_i \\ n^{-1} \sum_{i=1}^n X_i^2 \end{bmatrix} - \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} \right) \Rightarrow N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Gamma \right), \quad \text{où} \quad \Gamma := \begin{bmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{bmatrix}.$$

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,

$$\sqrt{n}(S_n^2 - (m_2 - m_1^2)) \Rightarrow -2m_1 T_1 + T_2,$$

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,

$$\sqrt{n}(S_n^2 - \mu_2) \Rightarrow N(0, \mu_4 - \mu_2^2),$$

où μ_k sont les moments des variables aléatoires recentrées, $\mu_k := \mathbb{E}[(X_1 - m_1)^k]$, $k = 1, \dots, 4$. $\diamond$

Exemple IV-2.50 (Coefficient de corrélation (voir Exemple IV-2.23)). 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

$$\rho := \frac{\text{Cov}(X_1, Y_1)}{\sqrt{\text{Var}(X_1) \text{Var}(Y_1)}}.$$

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

$$\bar{X}_n := n^{-1} \sum_{k=1}^n X_k, \quad \bar{Y}_n := n^{-1} \sum_{k=1}^n Y_k.$$

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

$$\mathbf{v} := (\mathbb{E}[X_1], \mathbb{E}[Y_1], \mathbb{E}[X_1^2], \mathbb{E}[Y_1^2], \mathbb{E}[X_1 Y_1]),$$

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-2.39) montre que

$$\sqrt{n}(\mathbf{T}_n - \mathbf{v}) \Rightarrow N(0, \Sigma).$$

◇

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-2.47), nous avons

$$\sqrt{n}(\hat{\rho}_n - \rho) \xRightarrow{\mathbb{P}} N(0, \gamma^2),$$

où $\gamma^2 := J_g(\mathbf{v}) \Sigma J_g(\mathbf{v})^T$. 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

$$\sqrt{n}(\hat{\rho}_n - \rho) \xRightarrow{\mathbb{P}} N(0, (1 - \rho^2)^2).$$

IV-2.6 Convergence des moments

Par définition, $X_n \Rightarrow X$ implique que pour toute fonction continue bornée f , $\mathbb{E}_n[f(X_n)] \rightarrow \mathbb{E}[f(X)]$. 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 $X_n \Rightarrow X$ et pour lesquelles nous n'avons pas $\mathbb{E}_n[f(X_n)] \rightarrow \mathbb{E}[f(X)]$ 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-2.51. *Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Si $X_n \Rightarrow X$, alors*

$$\mathbb{E}[\|X\|] \leq \liminf_{n \rightarrow \infty} \mathbb{E}_n[\|X_n\|].$$

Démonstration. En utilisant le théorème IV-2.30, nous avons $\|X_n\| \Rightarrow \|X\|$, ce qui implique que $\mathbb{P}_n(\|X_n\| > t) \rightarrow \mathbb{P}(\|X\| > t)$ 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 IV-1.40), nous avons donc

$$\mathbb{E}[\|X\|] = \int_0^\infty \mathbb{P}(\|X\| > t) dt \leq \liminf_{n \rightarrow \infty} \int_0^\infty \mathbb{P}_n(\|X_n\| > t) dt = \liminf_{n \rightarrow \infty} \mathbb{E}_n[\|X_n\|].$$

□

Corollaire IV-2.52. Pour tout $n \in \mathbb{N}$, soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ un espace de probabilité et X_n un vecteur aléatoire, défini sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

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 $X_n \Rightarrow X$ alors

$$\mathbb{E}[\|f(X)\|] \leq \liminf_{n \rightarrow \infty} \mathbb{E}_n[\|f(X_n)\|] .$$

Définition IV-2.53 (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 $(\Omega, \mathcal{F}, \mathbb{P})$. La suite $\{X_n, n \in \mathbb{N}\}$ est dite uniformément intégrable si

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M\}}] = 0 .$$

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-2.54. (i) Soit $\{X_n, n \in \mathbb{N}\}$ une suite de vecteurs aléatoires définis sur $(\Omega, \mathcal{F}, \mathbb{P})$, 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-3.1), pour tout $M > 0$,

$$\limsup_{n \rightarrow \infty} \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M\}}] \leq M^{-\delta} \limsup_{n \rightarrow \infty} \mathbb{E}[\|X_n\|^{1+\delta}] .$$

(ii) La condition $\|X_n\| \leq X$ implique $\{\|X_n\| \geq M\} \subset \{X \geq M\}$. Par conséquent, pour tout $M \geq 0$,

$$\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M\}} \leq X \mathbb{1}_{\{X \geq M\}},$$

ce qui implique que, en appliquant le théorème de convergence dominée (voir Théorème IV-1.41),

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M\}}] \leq \lim_{M \rightarrow \infty} \mathbb{E}[X \mathbb{1}_{\{X \geq M\}}] = 0.$$

(iii) Pour tout $n \in \mathbb{N}$, nous avons

$$\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M\}} \leq Y_n \mathbb{1}_{\{Y_n \geq M\}},$$

et donc

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M\}}] \leq \lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}[Y_n \mathbb{1}_{\{Y_n \geq M\}}] = 0.$$

(iv) Nous avons $\mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M\}}] = \mathbb{E}[\|X_1\| \mathbb{1}_{\{\|X_1\| \geq M\}}]$ et on conclut en appliquant le théorème de convergence dominée.

(v) Remarquons tout d'abord que

$$\begin{aligned} \|X_n + Y_n\| \mathbb{1}_{\{\|X_n + Y_n\| \geq M\}} &\leq (\|X_n\| + \|Y_n\|) \mathbb{1}_{\{\|X_n\| + \|Y_n\| \geq M\}} \\ &\leq 2 \|X_n\| \mathbb{1}_{\{\|X_n\| + \|Y_n\| \geq M\}} \mathbb{1}_{\{\|X_n\| \geq \|Y_n\|\}} + 2 \|Y_n\| \mathbb{1}_{\{\|X_n\| + \|Y_n\| \geq M\}} \mathbb{1}_{\{\|Y_n\| > \|X_n\|\}} \\ &\leq 2 \|X_n\| \mathbb{1}_{\{2\|X_n\| \geq M\}} + 2 \|Y_n\| \mathbb{1}_{\{2\|Y_n\| \geq M\}}. \end{aligned}$$

La preuve en découle immédiatement.

(vi) Par hypothèse, il existe une constante $C > 0$ telle que $\mathbb{P}(\sup_n \|Y_n\| \leq C) = 1$. Par suite

$$\mathbb{E}[\|X_n Y_n\| \mathbb{1}_{\{\|X_n Y_n\| \geq M\}}] \leq C \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\| \geq M/C\}}]$$

ce qui conclut la preuve. $\square$

Comme le montre le théorème ci-dessous, l'uniforme intégrabilité permet de relier la convergence en loi et la convergence des moments.

Théorème IV-2.55. Soit $f : \mathbb{R}^k \mapsto \mathbb{R}$ une fonction borélienne continue en tout point de $C \in \mathcal{B}(\mathbb{R}^k)$. Supposons que $X_n \Rightarrow X$ et $\mathbb{P}(X \in C) = 1$. Alors, $\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]$ 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 $\mathbb{E}[Y_n] \rightarrow \mathbb{E}[Y]$, 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 $Y_n \Rightarrow Y$. Nous notons $a \wedge b := \inf(a, b)$. L'inégalité triangulaire donne d'une part

$$\mathbb{E}[Y \wedge M] \leq |\mathbb{E}[Y_n \wedge M] - \mathbb{E}[Y \wedge M]| + \mathbb{E}[Y_n \wedge M] \quad (\text{IV-2.27})$$

et d'autre part

$$|\mathbb{E}[Y_n] - \mathbb{E}[Y]| \leq |\mathbb{E}[Y_n] - \mathbb{E}[Y_n \wedge M]| + |\mathbb{E}[Y_n \wedge M] - \mathbb{E}[Y \wedge M]| + |\mathbb{E}[Y \wedge M] - \mathbb{E}[Y]|. \quad (\text{IV-2.28})$$

Comme la fonction $y \mapsto y \wedge M$ est continue et bornée, $|\mathbb{E}[Y_n \wedge M] - \mathbb{E}[Y \wedge M]| \rightarrow 0$ quand $n \rightarrow \infty$. Le second terme de la partie droite de (IV-2.27) 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-2.28) 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. $\square$

Exemple IV-2.56. Soit $\{X_n, n \in \mathbb{N}\}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ telle que $X_n \Rightarrow X$ et $\limsup \mathbb{E}[\|X_n\|^p] < \infty$ pour $p > 0$. Alors, pour tout $0 \leq q < p$, nous avons $\mathbb{E}[\|X_n\|^q] \rightarrow \mathbb{E}[\|X\|^q]$. 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 :

$$\begin{aligned} \mathbb{E}[\|X_n\|^q \mathbb{1}_{\{\|X_n\|^q \geq M\}}] &= \mathbb{E}[\|X_n\|^q \mathbb{1}_{\{\|X_n\|^q / M \geq 1\}}] \\ &\leq \mathbb{E}[\|X_n\|^q (\|X_n\|^q / M)^{(p-q)/q}] = M^{1-p/q} \mathbb{E}[\|X_n\|^p], \end{aligned}$$

ce qui implique bien l'uniforme intégrabilité car $\sup_{n \geq 0} \mathbb{E}[\|X_n\|^p] < \infty$. $\diamond$

IV-2.7 Symboles o et O stochastiques

Définition IV-2.57 (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 $(\Omega, \mathcal{F}, \mathbb{P})$. La suite $\{X_n, n \in \mathbb{N}\}$ est dite bornée en probabilité si

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P}(\|X_n\| \geq M) = 0.$$

Si, pour $p > 0$, nous avons $\sup_{n \in \mathbb{N}} \mathbb{E}[\|X_n\|^p] < \infty$, l'inégalité de Markov (Lemme IV-3.1) montre que $\{X_n, n \in \mathbb{N}\}$ est bornée en probabilité, car

$$\sup_{n \geq 0} \mathbb{P}(\|X_n\| \geq M) \leq M^{-p} \sup_{n \geq 0} \mathbb{E}[\|X_n\|^p].$$

On montre aisément que

Lemme IV-2.58. Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$, définies sur l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Soit X un vecteur aléatoire à valeur $\mathbb{R}^d$. Si $X_n \Rightarrow X$, 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

$$X_n = o_P(1) \quad \text{signifie} \quad X_n \xrightarrow{\mathbb{P}\text{-prob}} 0, \quad (\text{IV-2.29})$$

$$X_n = O_P(1) \quad \text{si la suite } \{X_n, n \in \mathbb{N}\} \text{ est bornée en probabilité} \quad (\text{IV-2.30})$$

Plus généralement, pour $\{R_n, n \in \mathbb{N}\}$ une suite de variables aléatoires réelles, nous posons

$$X_n = o_P(R_n) \quad \text{signifie} \quad X_n = Y_n R_n \quad \text{avec} \quad Y_n = o_P(1), \quad (\text{IV-2.31})$$

$$X_n = O_P(R_n) \quad \text{signifie} \quad X_n = Y_n R_n \quad \text{avec} \quad Y_n = O_P(1). \quad (\text{IV-2.32})$$

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,

$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1), \\ o_P(1) + O_P(1) &= O_P(1), \\ O_P(1) o_P(1) &= o_P(1), \\ (1 + o_P(1))^{-1} &= 1 + o_P(1), \\ o_P(R_n) &= R_n o_P(1), \quad O_P(R_n) = R_n O_P(1). \end{aligned}$$

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 $X_n \xrightarrow{\mathbb{P}\text{-prob}} 0$ et $Y_n \xrightarrow{\mathbb{P}\text{-prob}} 0$, le théorème IV-2.32 implique que $(X_n, Y_n) \xrightarrow{\mathbb{P}\text{-prob}} (0, 0)$, ce qui équivaut à $(X_n, Y_n) \Rightarrow (0, 0)$. Le théorème de continuité (appliqué à $f : (x, y) \mapsto x + y$) implique $X_n + Y_n \Rightarrow 0$, qui équivaut à $X_n + Y_n \xrightarrow{\mathbb{P}\text{-prob}} 0$. La troisième règle est une façon concise d'écrire : si $\{X_n, n \in \mathbb{N}\}$ est bornée en probabilité et $Y_n \xrightarrow{\mathbb{P}\text{-prob}} 0$, alors $X_n Y_n \xrightarrow{\mathbb{P}\text{-prob}} 0$. Si $X_n \Rightarrow X$, alors ce résultat découle du lemme de Slutsky (car $X_n \Rightarrow X$ et $Y_n \xrightarrow{\mathbb{P}\text{-prob}} c$ implique que $Y_n X_n \Rightarrow cX$, donc si $c = 0$, $Y_n X_n \Rightarrow 0$ qui équivaut à $Y_n X_n \xrightarrow{\mathbb{P}\text{-prob}} 0$). 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-2.59. 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 $X_n \xrightarrow{\mathbb{P}\text{-prob}} 0$. Alors, pour tout $p > 0$,

- (i) Si $R(h) = o(\|h\|^p)$ quand $h \rightarrow 0$, alors $R(X_n) = o_P(\|X_n\|^p)$,
- (ii) Si $R(h) = O(\|h\|^p)$ quand $h \rightarrow 0$, 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-2.6) montre que $g(X_n) \xrightarrow{\mathbb{P}\text{-prob}} g(0) = 0$. La deuxième assertion se démontre de façon similaire. $\square$

Exemple IV-2.60 (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 :

$$S_n^2 = n^{-1} \sum_{i=1}^n (Y_i - \mu)^2 - (\bar{Y}_n - \mu)^2$$

$$\sqrt{n}(S_n^2 - \sigma^2) = n^{-1/2} \sum_{i=1}^n \left((Y_i - \mu)^2 - \sigma^2 \right) - \sqrt{n}(\bar{Y}_n - \mu)^2.$$

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-2.39).

$$n^{-1/2} \sum_{i=1}^n Z_i \implies N(0, \gamma), \quad \text{où } \gamma := \mathbb{E}(Y_1 - \mu)^4 - \sigma^4.$$

Le T.L.C. montre aussi que $\sqrt{n}(\bar{Y}_n - \mu) \implies N(0, \sigma^2)$; 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-2.58). Le lemme IV-2.59 montre que $n^{1/2}(\bar{Y}_n - \mu)^2 = O_P(n^{-1/2}) = o_P(1)$. Le lemme de Slutsky (Lemme IV-2.33) montre donc :

$$\sqrt{n}(S_n^2 - \sigma^2) \implies N(0, \gamma). \quad \diamond$$

Chapitre IV-3

Inégalités de déviations

IV-3.1 Inégalités de Markov et de Bienayme-Tchebychev

Lemme IV-3.1 (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-3.2 (Inégalité de Bienayme-Tchebychev). Soit Z une variable aléatoire réelle vérifiant $\mathbb{E}[Z^2] < \infty$. Notons $\mu = \mathbb{E}[Z]$. Alors, pour tout $\delta > 0$:

$$\mathbb{P}(|Z - \mu| \geq \delta) \leq \text{Var}(Z)/\delta^2. \quad (\text{IV-3.1})$$

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 $\delta < \sqrt{\text{Var}(Z)}$). Il est possible, à moindres frais, de proposer un énoncé de cette inégalité ne souffrant pas de ce problème :

Lemme IV-3.3 (Inégalité de Tchebychev-Cantelli). Soit Z une variable aléatoire réelle vérifiant $\mathbb{E}[Z^2] < \infty$. Alors, pour tout $\delta > 0$,

$$\mathbb{P}(Z - \mathbb{E}(Z) \geq \delta) \leq \frac{\text{Var}(Z)}{\text{Var}(Z) + \delta^2}.$$

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 :

$$\delta^2 \leq \mathbb{E}[(\delta - Z)^2] \mathbb{P}(Z \leq \delta) = (\delta^2 + \text{Var}(Z)) \mathbb{P}(Z \leq \delta),$$

ce qui donne le résultat désiré. □

IV-3.2 Inégalité de Jensen

Lemme IV-3.4 (Inégalité de Jensen). Soit I un intervalle ouvert de $\mathbb{R}$, f une fonction convexe sur I , Y une variable aléatoire réelle telle que $\mathbb{P}(Y \in I) = 1$ et $\mathbb{E}[|Y|] < \infty$. Alors

$$f(\mathbb{E}[Y]) \leq \mathbb{E}[f(Y)] . \quad (\text{IV-3.2})$$

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 $t \in I$, il existe une constante c_t telle que

$$f(t) + c_t(y - t) \leq f(y) , \quad \text{pour tout } y \in I. \quad (\text{IV-3.3})$$

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

$$f(\mathbb{E}[Y]) + c(y - \mathbb{E}[Y]) \leq f(y) , \quad \text{pour tout } y \in I$$

et par conséquent (IV-3.2) 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-3.3) est stricte sur l'évènement $\{Y \neq \mathbb{E}[Y]\}$ qui est de probabilité strictement positive. $\square$

IV-3.3 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-3.2). Supposons que $X_1, \dots, X_n$ 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

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| \geq \varepsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sum_{i=1}^n \text{Var}(X_i)}{n^2 \varepsilon^2} .$$

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

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - p\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2} . \quad (\text{IV-3.4})$$

Cette borne est très imprécise, comme nous allons maintenant le voir. Soit $\Phi(y) = \int_{-\infty}^y e^{-t^2/2} / \sqrt{2\pi} dt$ la fonction de répartition de la loi normale centrée réduite. Le théorème de la limite centrale montre que

$$n^{-1/2} \sum_{i=1}^n \{X_i - p\} \xrightarrow{\mathbb{P}} N(0, \text{Var}(X_1)) . \quad (\text{IV-3.5})$$

Comme les variables aléatoires $\{X_i, i \in \mathbb{N}\}$ sont binômiales, nous avons $\text{Var}(X_1) = p(1-p)$. Nous avons par conséquent,

$$\sqrt{\frac{n}{p(1-p)}} \left(\frac{1}{n} \sum_{i=1}^n X_i - p \right) \xrightarrow{\mathbb{P}} N(0, 1) ,$$

ce qui implique que, pour tout $y > 0$,

$$\begin{aligned} \mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{i=1}^n X_i - p\right) \geq y\right) &\rightarrow \int_y^\infty \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt \\ &\leq \frac{1}{y} \int_y^\infty \frac{t}{\sqrt{2\pi}} \exp(-t^2/2) dt \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-y^2/2}}{y}, \end{aligned}$$

Cette propriété suggère une inégalité de la forme, pour $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - p \geq \varepsilon\right) \approx n^{-1/2} e^{-n\varepsilon^2/(2p(1-p))}. \quad (\text{IV-3.6})$$

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-3.4)) 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 $t > 0$,

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}.$$

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-3.5 (Chernoff). Soit X une variable aléatoire réelle vérifiant, pour tout $\lambda \in \mathbb{R}$:

$$\phi_X(\lambda) = \mathbb{E}[e^{\lambda X}] < \infty. \quad (\text{IV-3.7})$$

On a, pour tout $a \geq 0$,

$$\mathbb{P}(X \geq a) \leq e^{-h_X(a)} \quad \text{pour tout } a \geq 0 \quad (\text{IV-3.8})$$

et

$$\mathbb{P}(X \leq -a) \leq e^{-h_X(-a)} \quad \text{pour tout } a \geq 0 \quad (\text{IV-3.9})$$

où la fonction h_X est la transformée de Crámer de X , donnée par :

$$h_X(a) = \begin{cases} \sup_{\lambda \geq 0} \{a\lambda - \log \phi_X(\lambda)\} & \text{si } a \geq 0 \\ \sup_{\lambda \leq 0} \{-a\lambda - \log \phi_X(\lambda)\} & \text{si } a \leq 0. \end{cases} \quad (\text{IV-3.10})$$

Démonstration. Pour $\lambda \geq 0$ et $a \geq 0$, la majoration $e^{\lambda(X-a)} \geq \mathbb{1}_{\{X \geq a\}}$ implique

$$\mathbb{P}(X \geq a) \leq \mathbb{E}[e^{\lambda(X-a)}] \leq \mathbb{E}[e^{\lambda X}] e^{-a\lambda} = e^{-(a\lambda - \psi_X(\lambda))}.$$

où $\psi_X(\lambda) = \log \phi_X(\lambda)$. L'Eq (IV-3.8) découle de l'optimisation de cette inégalité en $\lambda > 0$.

$$\mathbb{P}(X \geq a) \leq e^{-\sup_{\lambda \geq 0} (a\lambda - \psi_X(\lambda))} = e^{-h_X(a)}.$$

La preuve de (IV-3.9) est analogue. Pour $a \geq 0$ et $\lambda \leq 0$, nous avons

$$\mathbb{P}(X \leq -a) = \mathbb{E}[\mathbb{1}_{\{X \leq -a\}}] \leq \mathbb{E}[e^{\lambda(X+a)}] = e^{a\lambda + \psi_X(\lambda)} = e^{-(-a\lambda - \psi_X(\lambda))},$$

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-3.6. 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 $a \in \mathbb{R}$:

$$h_{\bar{X}_n}(a) = nh_X(a)$$

Démonstration. On note $\phi_{\bar{X}_n}(\lambda)$ la fonction génératrice de $\bar{X}_n$, alors en utilisant l'hypothèse d'indépendance

$$\phi_{\bar{X}_n}(\lambda) = \mathbb{E}\left[e^{\lambda n^{-1} \sum_{i=1}^n X_i}\right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i/n}] = \{\phi_X(\lambda/n)\}^n. \quad (\text{IV-3.11})$$

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 $a \geq 0$

$$h_{\bar{X}_n}(a) = \sup_{\lambda \geq 0} (a\lambda - n\psi_X(\lambda/n)) = n \sup_{\lambda \geq 0} (a\lambda/n - \psi_X(\lambda/n)) = nh_X(a).$$

La preuve pour $a \leq 0$ est identique. □

Corollaire IV-3.7. Sous les conditions du théorème IV-3.6, nous avons pour tout $a > 0$.

$$\begin{aligned} \mathbb{P}(\bar{X}_n \geq a) &\leq e^{-nh_X(a)}, \\ \mathbb{P}(\bar{X}_n \leq -a) &\leq e^{-nh_X(-a)}. \end{aligned}$$

IV-3.4 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 . [Hoeffding, 1963, Hoeffding, 1967] :

Lemme IV-3.8. 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

$$\psi_Y(\lambda) = \log \int \exp(\lambda y) \mathbb{Q}(dy) \quad \lambda > 0.$$

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

$$\begin{aligned}\mathbb{E}[e^{\lambda Y}] &\leq \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} \\ &= (1-p + pe^{\lambda(b-a)})e^{-p\lambda(b-a)} := e^{\rho(u)},\end{aligned}$$

où $u = \lambda(b-a)$ et $\rho(u) = -pu + \log(1-p + pe^u)$. Un calcul élémentaire montre que la dérivée de ρ est donnée par

$$\rho'(u) = -p + \frac{p}{p + (1-p)e^{-u}}$$

et donc que $\rho(0) = \rho'(0) = 0$. En notant que, pour tous réels a et b , $4ab \leq (a+b)^2$ (remarquons en effet que $2ab \leq a^2 + b^2$),

$$\rho''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \leq \frac{1}{4}.$$

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 $\lambda \rightarrow \psi_Y(\lambda)$ est deux fois dérivable et,

$$\psi_Y''(\lambda) = e^{-\psi_Y(\lambda)} \int y^2 \exp(\lambda y) \mathbb{Q}(dy) - e^{-2\psi_Y(\lambda)} \left(\int y \exp(\lambda y) \mathbb{Q}(dy) \right)^2. \quad (\text{IV-3.12})$$

Posons,

$$\mathbb{Q}_\lambda = e^{-\psi_Y(\lambda)} e^{\lambda y} \cdot \mathbb{Q}.$$

Par construction, $\mathbb{Q}_\lambda$ est une mesure de probabilité sur $\mathcal{B}(\mathbb{R})$. A l'aide de cette probabilité $\mathbb{Q}_\lambda$, on peut interpréter (IV-3.12) de la manière suivante :

$$\psi_Y''(\lambda) = \int \left(z - \int z \mathbb{Q}_\lambda(dz) \right)^2 \mathbb{Q}_\lambda(dz) = \text{Var}_\lambda(Z),$$

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

$$\text{Var}_\lambda(Z) = \text{Var}_\lambda(Z - (b+a)/2) \leq \mathbb{E}_\lambda \left[(Z - (b+a)/2)^2 \right] \leq \frac{(b-a)^2}{4},$$

ce qui implique

$$\psi_Y''(\lambda) \leq (b-a)^2/4. \quad (\text{IV-3.13})$$

En intégrant (IV-3.13) et en utilisant $\psi_Y(0) = \psi_Y'(0) = 0$, on déduit

$$\psi_Y(\lambda) \leq \lambda^2 \frac{(b-a)^2}{8}. \quad (\text{IV-3.14})$$

□

Nous pouvons maintenant utiliser directement ce résultat dans la borne obtenue par la méthode de Chernoff (voir Théorème IV-3.6) 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-3.9 (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 $a_i \leq X_i \leq b_i$. Pour tout $t > 0$, nous avons

$$\begin{aligned}\mathbb{P}(\bar{X}_n \geq t) &\leq e^{-2n^2 t^2 / \sum_{i=1}^n (b_i - a_i)^2}, \\ \mathbb{P}(\bar{X}_n \leq -t) &\leq e^{-2n^2 t^2 / \sum_{i=1}^n (b_i - a_i)^2}.\end{aligned}$$

Démonstration. En utilisant le théorème IV-3.6, nous avons pour tout $t > 0$,

$$\mathbb{P}(\bar{X}_n \geq t) \leq e^{-h_{\bar{X}_n}(t)} \quad \text{où } h_{\bar{X}_n}(t) = \sup_{\lambda \geq 0} \left\{ t\lambda - \sum_{i=1}^n \psi_{X_i}(\lambda/n) \right\}.$$

Le Lemme IV-3.8 implique que

$$\begin{aligned}h_{\bar{X}_n}(t) &= \sup_{\lambda \geq 0} \left\{ t\lambda - \left(\frac{\lambda}{n} \right)^2 \frac{1}{8} \sum_{i=1}^n (b_i - a_i)^2 \right\} \\ &= -2n^2 t^2 / \sum_{i=1}^n (b_i - a_i)^2\end{aligned}$$

□

Corollaire IV-3.10. 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 $t > 0$,

$$\mathbb{P}(|\bar{X}_n - p| \geq t) \leq 2 \exp(-2nt^2).$$

Démonstration. Appliquons l'inégalité de Hoeffding à $Y_i = X_i - p$. Les conditions du théorème IV-3.9 sont vérifiées avec $b_i - a_i = 1$. On conclut en écrivant

$$\mathbb{P}(|\bar{X}_n - p| \geq t) = \mathbb{P}(\bar{X}_n - p \geq t) + \mathbb{P}(\bar{X}_n - p \leq -t).$$

IV-3.5 Inégalité de Pisier

Dans la preuve de Théorème IV-3.9, 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-3.11 (Loi sous-gaussienne, variable sous-gaussienne). Soit $\mathbb{Q}$ une probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ vérifiant $\int_{-\infty}^{\infty} x \mathbb{Q}(dx) = 0$. La probabilité $\mathbb{Q}$ est dite sous-gaussienne de facteur de variance σ^2 si pour tout $\lambda \in \mathbb{R}$,

$$\psi(\lambda) = \log \int e^{\lambda x} \mathbb{Q}(dx) \leq \lambda^2 \sigma^2 / 2.$$

Nous notons $\mathcal{SG}(\sigma^2)$ 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 $\mathcal{SG}(\sigma^2)$.

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-3.12. 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 $\int x^2 \mu(dx) \leq \sigma^2$. $\diamond$

Théorème IV-3.13. Soit $\sigma > 0$, $n \geq 2$, et $\{Y_i\}_{i=1}^n$ n variables aléatoires réelles sous-gaussiennes de facteur de variance σ^2 . Alors,

$$\mathbb{E} \left[\max_{1 \leq i \leq n} Y_i \right] \leq \sigma \sqrt{2 \ln n}. \quad (\text{IV-3.15})$$

Si, de plus, $\mathbb{E}[e^{s(-Y_i)}] \leq e^{s^2 \sigma^2 / 2}$ pour tout $s > 0$ et $i \in \{1, \dots, n\}$, alors pour tout $n \geq 1$,

$$\mathbb{E} \left[\max_{1 \leq i \leq n} |Y_i| \right] \leq \sigma \sqrt{2 \ln(2n)}. \quad (\text{IV-3.16})$$

Démonstration. En utilisant l'inégalité de Jensen, pour tout $s > 0$, nous avons

$$e^{s \mathbb{E}[\max_{1 \leq i \leq n} Y_i]} \leq \mathbb{E}[e^{s \max_{1 \leq i \leq n} Y_i}] = \mathbb{E} \left[\max_{1 \leq i \leq n} e^{s Y_i} \right] \leq \sum_{i=1}^n \mathbb{E}[e^{s Y_i}] \leq n e^{s^2 \sigma^2 / 2}.$$

Par conséquent, en utilisant de nouveau l'inégalité de Jensen,

$$\mathbb{E} \left[\max_{1 \leq i \leq n} Y_i \right] \leq \frac{\ln n}{s} + \frac{s \sigma^2}{2},$$

et nous obtenons (IV-3.15) en posant $s = \sqrt{2 \ln n} / \sigma$.

Notons que

$$\max_{i \leq n} |Y_i| = \max(Y_1, -Y_1, \dots, Y_n, -Y_n).$$

En appliquant (IV-3.15) nous obtenons (IV-3.16). $\square$

Chapitre IV-4

Fonction de répartition, quantiles et statistiques d'ordre

IV-4.1 Fonction de répartition

Définition IV-4.1 (Fonction de répartition). Soit μ une mesure de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. La fonction de répartition de la mesure de probabilité μ est la fonction $F : x \mapsto \mu(]-\infty, x])$.

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ 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-4.2. Soit $F : \mathbb{R} \rightarrow [0, 1]$ 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) $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow \infty} F(x) = 1$.

Démonstration. (i) Par définition d'une probabilité, $0 \leq \mu(A) \leq 1$ pour tout $A \in \mathcal{B}(\mathbb{R})$.

(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

$$\mu \left(\bigcap_n]-\infty, a + h_n] \right) = \lim \downarrow \mu(]-\infty, a + h_n])$$

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 $\lim_{x \nearrow a} F(x)$ existe et vaut $\mu(] -\infty, a[)$.

(v) Soit $\{x_n, n \in \mathbb{N}\}$ une suite décroissante de réels telle que $\lim_{n \rightarrow \infty} x_n = -\infty$ 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 IV-1.26-(v) montre que

$$\lim_{x \rightarrow -\infty} F(x) = \lim_{n \rightarrow \infty} F(x_n) = \lim \downarrow \mu(] -\infty, x_n]) = \mu(\emptyset) = 0.$$

On prouve de même que $\lim_{x \rightarrow +\infty} F(x) = 1$. □

Définition IV-4.3 (Fonction de répartition empirique). La fonction de répartition empirique de l'échantillon $(X_1, \dots, X_n)$, notée $\hat{F}_n$, est définie par

$$x \mapsto \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{]-\infty, x]}(X_k).$$

La fonction de répartition empirique $\hat{F}_n$ est la fonction de répartition de la mesure empirique

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

La proposition suivante montre que $\hat{F}_n$ 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-4.4. Soit $\{X_k, k \in \mathbb{N}^*\}$ une suite de variables aléatoires réelles i.i.d. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ et de fonction de répartition F . Soit $\hat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1, \dots, X_n)$.

Pour tout $x_0 \in \mathbb{R}$ et $n \in \mathbb{N}^*$ on a

$$\mathbb{E}[\hat{F}_n(x_0)] = F(x_0), \quad (\text{IV-4.1})$$

$$\text{Var}(\hat{F}_n(x_0)) = n^{-1} F(x_0)(1 - F(x_0)), \quad (\text{IV-4.2})$$

$$\mathbb{P}\left(\left|\hat{F}_n(x_0) - F(x_0)\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}, \quad \text{pour tout } \varepsilon > 0. \quad (\text{IV-4.3})$$

De plus, nous avons pour tout $x_0 \in \mathbb{R}$,

$$\hat{F}_n(x_0) \xrightarrow{\mathbb{P}\text{-p.s.}} F(x_0), \quad (\text{IV-4.4})$$

et

$$\sqrt{n}(\hat{F}_n(x_0) - F(x_0)) \Longrightarrow N(0, F(x_0)(1 - F(x_0))). \quad (\text{IV-4.5})$$

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

$$\mathbb{E}[\hat{F}_n(x_0)] = n^{-1} \sum_{k=1}^n F(x_0) = F(x_0)$$

$$\text{Var}(\hat{F}_n(x_0)) = n^{-2} \sum_{k=1}^n F(x_0)(1 - F(x_0)) = n^{-1} F(x_0)(1 - F(x_0)).$$

La preuve de (IV-4.3) découle de l'inégalité d'Hoeffding (Théorème IV-3.9). □

Un raffinement de la preuve montre que la convergence (IV-4.4) de la fonction de répartition empirique est en fait uniforme en $x \in \mathbb{R}$.

Théorème IV-4.5 (Glivenko-Cantelli). Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires réelles i.i.d. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, de fonction de répartition F . Soit $\hat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1, \dots, X_n)$. Alors,

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\mathbb{P}\text{-p.s.}} 0. \quad (\text{IV-4.6})$$

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\left\{ \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \right\}^2 \right] = 0. \quad (\text{IV-4.7})$$

Démonstration. Nous allons nous ramener à la loi forte des grands nombres, en discrétisant le problème. Pour $m \in \mathbb{N}^*$ construisons $a_0, a_1, \dots, a_m, a_{m+1}$ de telle sorte que

$$[a_i, a_{i+1}[= \left\{ x \in \mathbb{R} : \frac{i}{m} \leq F(x) < \frac{i+1}{m} \right\}, \quad \text{pour } i = 1, 2, \dots, m-1,$$

et $a_0 = -\infty, a_{m+1} = +\infty$. Pour $i \in \{0, \dots, m\}$ et $x \in [a_i, a_{i+1}[$, nous avons

$$\begin{aligned} \hat{F}_n(x) - F(x) &\leq \hat{F}_n(a_{i+1}-) - F(a_i) \leq \hat{F}_n(a_{i+1}-) - F(a_{i+1}-) + \frac{1}{m}, \\ F(x) - \hat{F}_n(x) &\leq F(a_i) - \hat{F}_n(a_i) + \frac{1}{m}. \end{aligned}$$

Par conséquent, comme $\mathbb{R} =]a_0, a_1[\cup (\cup_{i=0}^m [a_i, a_{i+1}[)$ et $\lim_{x \rightarrow a_0} \hat{F}_n(x) = \lim_{x \rightarrow a_0} F(x) = 0$ et $\lim_{x \rightarrow a_{m+1}} \hat{F}_n(x) = \lim_{x \rightarrow a_{m+1}} F(x) = 1$, nous avons

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \max_{1 \leq i \leq m} \left\{ |F(a_i) - \hat{F}_n(a_i)| + |F(a_i-) - \hat{F}_n(a_i-)| + \frac{1}{m} \right\}, \quad (\text{IV-4.8})$$

Pour tout m , lorsque $n \rightarrow \infty$, (IV-4.4) entraîne

$$\limsup_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \leq \frac{1}{m}.$$

Comme m peut être choisi arbitrairement grand, ceci montre (IV-4.6). La preuve de (IV-4.7) découle du théorème de convergence dominé (Théorème IV-1.41) en remarquant que $\sup_{x \in \mathbb{R}} |F(x) - \hat{F}_n(x)| \leq 2$. $\square$

Il est possible d'obtenir un résultat beaucoup plus précis que celui donné par l'équation (IV-4.3) sur la déviation de la fonction de répartition empirique.

Théorème IV-4.6. [Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW).] Soit $X_1, \dots, X_n$ des variables i.i.d. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ et de fonction de répartition F . Soit $\hat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1, \dots, X_n)$. Pour tout $\varepsilon > 0$,

$$\mathbb{P} \left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2}. \quad (\text{IV-4.9})$$

On remarque que la borne est la même que celle qui apparaît dans la borne exponentielle (IV-4.3). Ce résultat a été établi par [?], qui a de plus montré que cette borne était optimale. La preuve de l'inégalité (IV-4.9) est délicate. En utilisant le fait que

$$\mathbb{E}[U] = \int_0^\infty \mathbb{P}(U > t) dt$$

pour une variable aléatoire positive U , on déduit du Théorème IV-4.6 que

$$\begin{aligned} \left(\mathbb{E} \left[\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \right] \right)^2 &\leq \mathbb{E} \left[\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|^2 \right] \\ &\leq \int_0^\infty \mathbb{P} \left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|^2 > t \right) dt \leq \frac{1}{n} \end{aligned}$$

ce qui complète le résultat donné par le Théorème IV-4.5.

Démonstration. (du théorème IV-4.6) Nous allons montrer un résultat plus faible : pour tout $\varepsilon > 0$,

$$\mathbb{P} \left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon \right) \leq 4 \left(\frac{2}{\varepsilon} + 1 \right) e^{-n \frac{\varepsilon^2}{2}} .$$

Soit ε et N_ε la partie entière supérieure de $2/\varepsilon$. Pour tout $t \in \mathbb{R}$, on définit

$$\hat{F}_n(t^-) := \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{]-\infty, t[}(X_k) .$$

Pour tout $k = 0, \dots, N_\varepsilon$, 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 $\hat{F}_n$ et F sont croissantes,

$$\hat{F}_n(a_k) - F(a_{k+1}^-) \leq \hat{F}_n(t) - F(t) \leq \hat{F}_n(a_{k+1}^-) - F(a_k) ,$$

et donc

$$\hat{F}_n(a_k) - F(a_k) - \frac{\varepsilon}{2} \leq \hat{F}_n(t) - F(t) \leq \hat{F}_n(a_{k+1}^-) - F(a_{k+1}^-) + \frac{\varepsilon}{2} .$$

Ainsi,

$$\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \leq \max_{k=1, \dots, N_\varepsilon} \left(\left| \hat{F}_n(a_k) - F(a_k) \right| \vee \left| \hat{F}_n(a_k^-) - F(a_k^-) \right| \right) + \frac{\varepsilon}{2} .$$

Or, par le théorème IV-4.4, pour tout $k = 1, \dots, N_\varepsilon$, on a,

$$\mathbb{P} \left(\left| \hat{F}_n(a_k) - F(a_k) \right| > \frac{\varepsilon}{2} \right) \leq 2e^{-n \frac{\varepsilon^2}{2}} , \quad \mathbb{P} \left(\left| \hat{F}_n(a_k^-) - F(a_k^-) \right| > \frac{\varepsilon}{2} \right) \leq 2e^{-n \frac{\varepsilon^2}{2}} .$$

On en déduit donc

$$\begin{aligned} \mathbb{P} \left(\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| > \varepsilon \right) &\leq \mathbb{P} \left(\bigcup_{k=1}^{N_\varepsilon} \left\{ \left| \hat{F}_n(a_k) - F(a_k) \right| > \frac{\varepsilon}{2} \right\} \cup \left\{ \left| \hat{F}_n(a_k^-) - F(a_k^-) \right| > \frac{\varepsilon}{2} \right\} \right) \\ &\leq \sum_{k=1}^{N_\varepsilon} \left(\mathbb{P} \left(\left| \hat{F}_n(a_k) - F(a_k) \right| > \frac{\varepsilon}{2} \right) + \mathbb{P} \left(\left| \hat{F}_n(a_k^-) - F(a_k^-) \right| > \frac{\varepsilon}{2} \right) \right) \\ &\leq 4N_\varepsilon e^{-n \frac{\varepsilon^2}{2}} \leq 4 \left(\frac{2}{\varepsilon} + 1 \right) e^{-n \frac{\varepsilon^2}{2}} . \end{aligned}$$

□

IV-4.2 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

$$F^{-1}(p) := \inf \{x \in \mathbb{R} : F(x) \geq p\}, \quad (\text{IV-4.10})$$

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-4.7. *Soit F la fonction de répartition d'une mesure de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. La fonction F^{-1} définie sur $]0, 1[$ par (IV-4.10) 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-4.8 (Quantiles). *Soit F la fonction de répartition d'une mesure de probabilité μ sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Pour $0 \leq p \leq 1$, le p -ième quantile de F est défini par*

$$Q(p) := F^{-1}(p) = \inf \{x \in \mathbb{R} : F(x) \geq p\}, \quad (\text{IV-4.11})$$

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

$$F \circ F^{-1}(p) = F(Q(p)) = p. \quad (\text{IV-4.12})$$

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-4.9. *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-4.12), donc

$$\mathbb{P}(F(X) \leq u) = \mathbb{P}(F(X) \leq F(F^{-1}(u))) = \mathbb{P}(X \leq F^{-1}(u)) = F \circ F^{-1}(u) = u. \quad \square$$

Définition IV-4.10 (Quantile empirique). Soient $(X_1, \dots, X_n)$ n variables aléatoires réelles définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. Le p -ième quantile empirique est le p -ième quantile de la fonction de répartition empirique $\hat{F}_n : x \mapsto \hat{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbb{1}_{]-\infty, x]}(X_i)$, i.e.

$$\hat{Q}_n(p) := \hat{F}_n^{-1}(p) = \inf \left\{ x \in \mathbb{R} : \hat{F}_n(x) \geq p \right\}.$$

Théorème IV-4.11. 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 $\hat{Q}_n$ 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 $\varepsilon > 0$,

$$\mathbb{P}(|\hat{Q}_n(p) - Q(p)| > \varepsilon) \leq 2e^{-2n\delta_\varepsilon^2},$$

où

$$\delta_\varepsilon := \min \{F(Q(p) + \varepsilon) - p, p - F(Q(p) - \varepsilon)\}. \quad (\text{IV-4.13})$$

Démonstration. Soit $\varepsilon > 0$. Nous avons

$$\mathbb{P}(|\hat{Q}_n(p) - Q(p)| > \varepsilon) = \mathbb{P}(\hat{Q}_n(p) > Q(p) + \varepsilon) + \mathbb{P}(\hat{Q}_n(p) < Q(p) - \varepsilon).$$

Nous allons borner ces deux termes séparément. Nous avons tout d'abord, en utilisant que $\hat{Q}_n(p)$ est par définition, l'infimum des réels vérifiant $\hat{F}_n(x) \geq p$,

$$\begin{aligned} \mathbb{P}(\hat{Q}_n(p) > Q(p) + \varepsilon) &= \mathbb{P}\left(p \geq \hat{F}_n(Q(p) + \varepsilon)\right) = \mathbb{P}\left(np \geq \sum_{i=1}^n \mathbb{1}_{\{X_i \leq Q(p) + \varepsilon\}}\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{\{X_i > Q(p) + \varepsilon\}} \geq n(1-p)\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n V_i - n\mathbb{E}[V_1] \geq n(1-p) - n\{1 - F(Q(p) + \varepsilon)\}\right), \end{aligned}$$

où l'on a posé $V_i := \mathbb{1}_{\{X_i > Q(p) + \varepsilon\}}$. Les variables aléatoires $V_1, \dots, V_n$ 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-3.9,

$$\mathbb{P}(\hat{Q}_n(p) > Q(p) + \varepsilon) = \mathbb{P}_F\left(\sum_{i=1}^n \{V_i - \mathbb{E}[V_i]\} > n\delta_1\right) \leq e^{-2n\delta_1^2}. \quad (\text{IV-4.14})$$

En procédant de même, nous avons

$$\mathbb{P}(\hat{Q}_n(p) < Q(p) - \varepsilon) = \mathbb{P}\left(\sum_{i=1}^n \{W_i - \mathbb{E}[W_i]\} \geq n\delta_2\right) \leq e^{-2n\delta_2^2}, \quad (\text{IV-4.15})$$

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-4.12. Soit $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires réelles i.i.d. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, de fonction de répartition F et de fonction quantile associée Q . Soit $\hat{Q}_n$ la fonction quantile empirique associée à la fonction de répartition empirique de l'échantillon $(X_1, \dots, X_n)$. Soit $p \in]0, 1[$. Supposons que pour tout $\varepsilon > 0$, $F(Q(p) + \varepsilon) > p$. Alors,

$$\hat{Q}_n(p) \xrightarrow{\mathbb{P}\text{-p.s.}} Q(p) .$$

Démonstration. D'après le théorème IV-4.11, pour tout $\varepsilon > 0$, il existe $\delta_\varepsilon > 0$ tel que

$$\sum_{n=1}^{\infty} \mathbb{P}(|\hat{Q}_n(p) - Q(p)| > \varepsilon) \leq 2 \sum_{n=1}^{\infty} e^{-2n\delta_\varepsilon^2} < \infty .$$

Par conséquent, le Lemme de Borel-Cantelli (Lemme IV-2.11) montre que $\hat{Q}_n(p) - Q(p) \xrightarrow{\mathbb{P}\text{-p.s.}} 0$ lorsque $n \rightarrow \infty$. $\square$

Proposition IV-4.13. Soit $(X_1, \dots, X_n)$ des variables aléatoires réelles i.i.d. définies sur le même espace $(\Omega, \mathcal{F}, \mathbb{P})$, et de fonction de répartition F . Soit $\hat{F}_n$ la fonction de répartition empirique associée à l'échantillon $(X_1, \dots, X_n)$. Pour tout $\varepsilon > 0$,

$$\mathbb{P} \left(\sup_{p \in]0, 1[} \left| F(\hat{F}_n^{-1}(p)) - p \right| > \varepsilon + \frac{1}{n} \right) \leq 2e^{-2n\varepsilon^2} .$$

Démonstration. Soit $p \in]0, 1[$. On écrit

$$F(\hat{F}_n^{-1}(p)) - p = \hat{F}_n(\hat{F}_n^{-1}(p)) - p + F(\hat{F}_n^{-1}(p)) - \hat{F}_n(\hat{F}_n^{-1}(p)) ,$$

dont on déduit la majoration

$$\left| F(\hat{F}_n^{-1}(p)) - p \right| \leq \left| \hat{F}_n(\hat{F}_n^{-1}(p)) - p \right| + \sup_{x \in \mathbb{R}} \left| F(x) - \hat{F}_n(x) \right| .$$

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 $\hat{F}_n(\hat{F}_n^{-1}(p)) = K_n/n$, donc $p \leq \hat{F}_n(\hat{F}_n^{-1}(p)) \leq p + 1/n$. Par suite,

$$\left| \hat{F}_n(\hat{F}_n^{-1}(p)) - p \right| \leq \frac{1}{n} .$$

On a finalement

$$\sup_{p \in]0, 1[} \left| F(\hat{F}_n^{-1}(p)) - p \right| \leq \frac{1}{n} + \sup_{x \in \mathbb{R}} \left| F(x) - \hat{F}_n(x) \right| .$$

On conclut avec l'inégalité de Dvoretzky-Kiefer-Wolfowitz (Théorème IV-4.6). $\square$

IV-4.3 Statistiques d'ordre

Définition IV-4.14 (Statistiques d'ordre). Soient $X_1, X_2, \dots, X_n$ n variables aléatoires réelles. Les valeurs ordonnées $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ de $X_1, X_2, \dots, X_n$ sont appelées les statistiques d'ordre.

Le minimum $X_{1:n}$ de $X_1, X_2, \dots, X_n$ est la 1-ère statistique d'ordre, le maximum $X_{n:n}$ est la n -ème statistique d'ordre. Notons que, pour tout $i \in \{1, \dots, n\}$,

$$X_{i:n} = \hat{F}_n^{-1}(i/n). \quad (\text{IV-4.16})$$

On appelle *médiane* de l'échantillon $(X_1, \dots, X_n)$, le quantile empirique d'ordre $1/2$; on le note M_n . On a

$$M_n := \hat{F}_n^{-1}(1/2) = \begin{cases} X_{m+1:n} & \text{si } n = 2m + 1 \\ X_{m:n} & \text{si } n = 2m \end{cases}.$$

Dans le cas où $n = 2m$, on trouve aussi parfois la définition symétrique suivante

$$M_{2m} = \frac{1}{2} (X_{m:2m} + X_{m+1:2m}).$$

Théorème IV-4.15 (Loi jointe des statistiques d'ordre). Soient $X_1, X_2, \dots, X_n$ des v.a. réelles i.i.d. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ et de loi admettant une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$. La loi jointe des statistiques d'ordre $(X_{1:n}, X_{2:n}, \dots, X_{n:n})$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, donnée par

$$(y_1, y_2, \dots, y_n) \mapsto n! f(y_1) f(y_2) \cdots f(y_n) \mathbb{1}_{\{y_1 < y_2 < \dots < y_n\}}.$$

Démonstration. Comme les variables $(X_1, \dots, X_n)$ 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,

$$\mathbb{P}(X_{1:n} < X_{2:n} < \dots < X_{n:n}) = 1.$$

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 $B \in \mathcal{B}(\mathbb{R}^n)$ un ensemble négligeable sous la mesure de Lebesgue. Nous avons

$$\begin{aligned} \mathbb{P}((X_{1:n}, \dots, X_{n:n}) \in B) &= \mathbb{P}((X_{1:n}, \dots, X_{n:n}) \in B, X_{1:n} < X_{2:n} < \dots < X_{n:n}) \\ &= \sum_{\pi} \mathbb{P}((X_{\pi(1)}, \dots, X_{\pi(n)}) \in B, X_{\pi(1)} < X_{\pi(2)} < \dots < X_{\pi(n)}) \end{aligned}$$

où la somme porte sur l'ensemble des $n!$ permutations π de l'ensemble $\{1, \dots, n\}$. Comme pour tout π , $\mathbb{P}((X_{\pi(1)}, \dots, X_{\pi(n)}) \in B, X_{\pi(1)} < X_{\pi(2)} < \dots < X_{\pi(n)}) = 0$ puisque B est négligeable et que $(X_{\pi(1)}, \dots, X_{\pi(n)})$ possède une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, nous avons $\mathbb{P}((X_{1:n}, \dots, X_{n:n}) \in B) = 0$. Ainsi, la loi jointe des statistiques d'ordre est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ (voir définition IV-1.47) et par le théorème de Radon-Nikodym (voir théorème IV-1.48), elle admet une densité.

Soit $-\infty \leq x_1, \dots, x_n \leq \infty$. L'événement $A := \{X_{1:n} \leq x_1, X_{2:n} \leq x_2, \dots, X_{n:n} \leq x_n\}$ est l'union des $n!$ événements disjoints

$$A_{\pi} := \{X_{\pi(1)} \leq x_1, X_{\pi(2)} \leq x_2, \dots, X_{\pi(n)} \leq x_n\} \cap \{X_{\pi(1)} < X_{\pi(2)} < \dots < X_{\pi(n)}\}$$

où π parcourt l'ensemble des permutations de $\{1, \dots, n\}$. Notons de plus que

$$\mathbb{P}(A_{\pi}) = \int \cdots \int_{y_1 < \dots < y_n} \prod_{i=1}^n \mathbb{1}_{\{y_i \leq x_i\}} f(y_i) dy_i,$$

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. $\square$

On obtient facilement la loi marginale des statistiques d'ordre.

Corollaire IV-4.16. Soit $X_1, \dots, X_n$ des v.a. réelles i.i.d. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, 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

$$x \mapsto \frac{n!}{(r-1)!(n-r)!} F^{r-1}(x) f(x) \{1 - F(x)\}^{n-r}. \quad (\text{IV-4.17})$$

Démonstration. Soit $1 \leq r \leq n$. Calculons la r -ième marginale de la loi jointe des statistiques d'ordre. D'après le théorème IV-4.15, nous avons

$$\begin{aligned} \mathbb{P}(X_{r:n} \leq x_*) &= n! \int_{-\infty}^{x_*} \left(\int \mathbb{1}_{y_1 < \dots < y_r} \prod_{i=1}^{r-1} f(y_i) dy_i \right) \left(\int \mathbb{1}_{y_r < \dots < y_n} \prod_{i=r+1}^n f(y_i) dy_i \right) f(y_r) dy_r, \\ &= n! \int_{-\infty}^{x_*} I_r(y_r) J_r(y_r) f(y_r) dy_r \end{aligned}$$

en ayant posé

$$\begin{aligned} I_r(x) &:= \int \mathbb{1}_{\{y_1 < \dots < y_{r-1} < x\}} \left(\prod_{i=1}^{r-1} f(y_i) \right) dy_1 \dots dy_{r-1}, \\ J_r(x) &:= \int \mathbb{1}_{\{x < y_{r+1} < \dots < y_n\}} \left(\prod_{i=r+1}^n f(y_i) \right) dy_{r+1} \dots dy_n, \end{aligned}$$

avec la convention $I_1(x) = J_n(x) = 1$. Notons que pour $r \geq 2$ et $\ell \leq n-1$, nous avons

$$I_r(x) = \int_{-\infty}^x f(x_{r-1}) I_{r-1}(x_{r-1}) dx_{r-1}, \quad J_\ell(x) = \int_x^{+\infty} J_{\ell+1}(y_{\ell+1}) f(y_{\ell+1}) dy_{\ell+1}.$$

On montre par une récurrence élémentaire que pour $r \geq 1$ et $\ell \leq n$

$$I_r(x) = \frac{F^{r-1}(x)}{(r-1)!}, \quad J_\ell(x) = \frac{(1-F(x))^{n-\ell}}{(n-\ell)!},$$

ce qui conclut la preuve. $\square$

Nous terminons en établissant la distribution asymptotique des quantiles centraux.

Théorème IV-4.17. Soit $\{X_n, n \in \mathbb{N}\}$ une suite de v.a. réelles i.i.d. définies sur le même espace $(\Omega, \mathcal{F}, \mathbb{P})$, 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

$$\sqrt{n}(k_n/n - p) \rightarrow 0. \quad (\text{IV-4.18})$$

Si F est dérivable au point $Q(p)$ et $F'(Q(p)) > 0$, alors

$$\sqrt{n}(X_{k_n:n} - Q(p)) \Rightarrow N\left(0, \frac{p(1-p)}{[F'(Q(p))]^2}\right).$$

Démonstration. Soit $p \in]0, 1[$. Par hypothèse, la fonction F est continue au point $Q(p)$ et donc

$$F(Q(p)) = p. \quad (\text{IV-4.19})$$

Nous étudions la limite quand $n \rightarrow \infty$ de la quantité

$$\mathbb{P}(\sqrt{n}(X_{k_n:n} - Q(p)) \leq a) = \mathbb{P}(X_{k_n:n} \leq Q(p) + an^{-1/2}) ,$$

et ce pour tout $a \in \mathbb{R}$. Définissons

$$K_n := \sum_{i=1}^n \mathbb{1}\{X_i \leq Q(p) + an^{-1/2}\} .$$

La clef de la preuve est d'observer que

$$\{X_{k_n:n} \leq Q(p) + an^{-1/2}\} = \{K_n \geq k_n\} . \quad (\text{IV-4.20})$$

Posons, pour $i \in \{1, \dots, n\}$,

$$Y_{n,i}(a) := n^{-1/2} \left\{ \mathbb{1}\{X_i \leq Q(p) + an^{-1/2}\} - F(Q(p) + an^{-1/2}) \right\} .$$

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 $a \in \mathbb{R}$,

$$Z_n(a) := \sum_{i=1}^n Y_{n,i}(a) \implies N(0, p(1-p)) , \quad (\text{IV-4.21})$$

par application du théorème de Lindeberg-Feller (Théorème IV-2.41). Notons tout d'abord que, pour tout $\varepsilon > 0$,

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}[Y_{n,i}^2(a) \mathbb{1}\{|Y_{n,i}(a)| \geq \varepsilon\}] &= n \mathbb{E}[Y_{n,1}^2(a) \mathbb{1}\{|Y_{n,1}(a)| \geq \varepsilon\}] \leq \mathbb{P}(|Y_{n,1}(a)| \geq \varepsilon) \\ &\leq \mathbb{P}(|\mathbb{1}\{X_1 \leq Q(p) + an^{-1/2}\} - F(Q(p) + an^{-1/2})| \geq \varepsilon\sqrt{n}) . \end{aligned}$$

On en déduit que

$$\lim_n \sum_{i=1}^n \mathbb{E}[Y_{n,i}^2(a) \mathbb{1}\{|Y_{n,i}(a)| \geq \varepsilon\}] = 0 ,$$

ce qui montre la condition (IV-2.17) du théorème IV-2.41. Remarquons d'autre part, que

$$\sum_{i=1}^n \mathbb{E}[Y_{n,i}^2(a)] = \text{Var} \left(\mathbb{1}\{X_1 \leq Q(p) + an^{-1/2}\} \right) = F(Q(p) + an^{-1/2})(1 - F(Q(p) + an^{-1/2}))$$

et comme par (IV-4.19), $F(Q(p)) = p$, 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-2.18) du théorème IV-2.41. La convergence (IV-4.21) découle directement du théorème de Lindeberg-Feller (théorème IV-2.41). Le théorème de Polya (Théorème IV-2.44) montre que la convergence (IV-4.21) est uniforme, i.e.

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(Z_n < x) - \Phi(\sigma^{-1}(p)x)| = 0 , \quad (\text{IV-4.22})$$

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 $\mathbb{P}(Z_n \geq x) = 1 - \mathbb{P}(Z_n < x)$ et $1 - \Phi(u) = \Phi(-u)$ pour tout $u \in \mathbb{R}$, (IV-4.22) implique

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(Z_n \geq x) - \Phi(-\sigma^{-1}(p)x)| = 0 . \quad (\text{IV-4.23})$$

Il découle maintenant de (IV-4.23) que

$$\begin{aligned}
 \mathbb{P}(\sqrt{n}(X_{k_n:n} - Q(p)) \leq a) &= \mathbb{P}(K_n \geq k_n) \\
 &= \mathbb{P}\left(\frac{K_n - nF(Q(p) + an^{-1/2})}{\sqrt{n}} \geq \frac{k_n - nF(Q(p) + an^{-1/2})}{\sqrt{n}}\right) \\
 &= \mathbb{P}\left(Z_n(a) \geq \frac{k_n - nF(Q(p) + an^{-1/2})}{\sqrt{n}}\right) \\
 &= \Phi(\sigma^{-1}(p)\{nF(Q(p) + an^{-1/2}) - k_n\}/\sqrt{n}) + o(1) .
 \end{aligned}$$

Comme la fonction F est dérivable au point $Q(p)$, nous avons en utilisant (IV-4.18),

$$\begin{aligned}
 \frac{nF(Q(p) + an^{-1/2}) - k_n}{\sqrt{n}} &= a \frac{F(Q(p) + an^{-1/2}) - p}{an^{-1/2}} + \frac{np - k_n}{\sqrt{n}} \\
 &= a \frac{F(Q(p) + an^{-1/2}) - p}{an^{-1/2}} - \sqrt{n} \left(\frac{k_n}{n} - p \right) \rightarrow aF'(Q(p)) .
 \end{aligned}$$

Nous avons donc établi que, pour tout $a \in \mathbb{R}$,

$$\mathbb{P}(\sqrt{n}(X_{k_n:n} - Q(p)) \leq a) \rightarrow \Phi(\sigma^{-1}(p)aF'(Q(p)))$$

ou de façon équivalente

$$\sqrt{n}(X_{k_n:n} - Q(p)) \Rightarrow N\left(0, \frac{\sigma^2(p)}{[F'(Q(p))]^2}\right) . \quad (\text{IV-4.24})$$

□

Chapitre IV-5

Famille de distributions

IV-5.1 Loi gaussienne

Définition IV-5.1 (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 $\mathbb{R}$:

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

La fonction caractéristique de la loi gaussienne réduite a pour expression :

$$\varphi(t) = \mathbb{E}[\exp(itX)] = \exp(-t^2/2)$$

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

$$\mu_{2n} = \mathbb{E}[X^{2n}] = \frac{(2n)!}{n! 2^n} = 1 \times 3 \times 5 \dots \times (2n-1)$$

Définition IV-5.2 (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 $X \sim N(\mu, \sigma^2)$. μ est l'espérance de X et σ^2 sa variance. Lorsque $\sigma^2 > 0$, la loi $N(\mu, \sigma^2)$ a une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par :

$$g_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

La fonction caractéristique d'une variable gaussienne de moyenne μ et de variance σ^2 est donnée par

$$\varphi_{\mu, \sigma^2}(t) = \exp\left(i\mu t - \frac{\sigma^2}{2}t^2\right). \quad (\text{IV-5.1})$$

IV-5.2 Loi gaussienne multivariée

Définition IV-5.3 (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 :

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{i=1}^n t_i \mathbb{E}[X_i] = \mathbf{t}^T \mathbb{E}[\mathbf{X}] \\ \mathbb{E}[(Y - \mathbb{E}[Y])^2] &= \sum_{i,j=1}^n t_i t_j \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbf{t}^T \Gamma \mathbf{t}\end{aligned}$$

où $\Gamma = (\text{Cov}(X_i, X_j))_{1 \leq i, j \leq n}$ est la matrice de covariance du vecteur $\mathbf{X}$.

Définition IV-5.4 (loi $N(\mu, \Gamma)$). Soit $\mu \in \mathbb{R}^n$ 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 μ et de covariance Γ et nous écrirons $\mathbf{X} \sim N(\mu, \Gamma)$, si, pour tout $\mathbf{t} = (t_1, \dots, t_n) \in \mathbb{R}^n$, nous avons $\mathbf{t}^T \mathbf{X} \sim N(\mathbf{t}^T \mu, \mathbf{t}^T \Gamma \mathbf{t})$.

Cette définition implique de façon immédiate :

Proposition IV-5.5. Soit $\mathbf{A}$ une matrice $m \times n$, $\mathbf{b} \in \mathbb{R}^m$ et soit $\mathbf{X} \sim N(\mu, \Gamma)$. Alors, $\mathbf{AX} + \mathbf{b} \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}^T \Gamma \mathbf{A})$.

Démonstration. Posons $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ et notons que, pour tout $\mathbf{s} \in \mathbb{R}^m$, nous avons :

$$\mathbf{s}^T \mathbf{Y} = (\mathbf{A}^T \mathbf{s})^T \mathbf{X} + \mathbf{s}^T \mathbf{b} \sim N(\mathbf{s}^T \mathbf{A}\mu + \mathbf{s}^T \mathbf{b}, \mathbf{s}^T \mathbf{A} \Gamma \mathbf{A}^T \mathbf{s}). \quad \square$$

Proposition IV-5.6. Soit $\Gamma \in \text{Mat}_n(\mathbb{R})$ une matrice semi-définie positive, $\text{Rang}(\Gamma) = k \leq n$ et soit $\mu \in \mathbb{R}^n$. $\mathbf{X} \sim N(\mu, \Gamma)$ si et seulement si, pour tout $\mathbf{A} \in \text{Mat}_{n,k}(\mathbb{R})$ tel que $\mathbf{AA}^T = \Gamma$, il existe $\mathbf{Z} \sim N(0, \mathbf{I}_k)$ tel que $\mathbf{X} = \mathbf{AZ} + \mu$.

Démonstration. Si $\mathbf{Z} \sim N(0, \mathbf{I}_k)$, nous déduisons de la proposition IV-5.5 $\mathbf{AZ} + \mu \sim N(\mu, \Gamma)$.

Réciproquement, soit $\mathbf{X} \sim N(\mu, \Gamma)$. Comme $\mathbf{A}$ est de rang k , la matrice $\mathbf{A}^T \mathbf{A} \in \text{Mat}_k(\mathbb{R})$ 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 : $\mathbf{A}^\# \mathbf{A} = \mathbf{I}_k$ et $\mathbf{AA}^\#$ est le projecteur orthogonal sur l'image de $\mathbf{A}$ (par construction, $\text{Im}(\mathbf{A}) = \text{Im}(\Gamma)$). Soit $\mathbf{Z} = \mathbf{A}^\# (\mathbf{X} - \mu)$. La proposition IV-5.5 implique que $\mathbf{Z} \sim N(0, \mathbf{I}_k)$. $\square$

On pourrait choisir cette caractérisation comme définition de la loi gaussienne $N(\mu, \Gamma)$.

La fonction caractéristique de $\mathbf{X} \sim N(\mu, \Gamma)$ se déduit directement de (IV-5.1). Comme la fonction caractéristique caractérise la loi, nous avons :

Proposition IV-5.7. $\mathbf{X} = (X_1, \dots, X_n) \sim N(\mu, \Gamma)$ si et seulement si sa fonction caractéristique $\varphi_{\mu, \Gamma}(\mathbf{t}) := \mathbb{E}[e^{i\mathbf{t}^T \mathbf{X}}]$ est donnée par :

$$\varphi_{\mu, \Gamma}(\mathbf{t}) = \exp\left(i\mathbf{t}^T \mu - \frac{1}{2} \mathbf{t}^T \Gamma \mathbf{t}\right) \quad (\text{IV-5.2})$$

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-5.8. Soit $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T \sim \mathcal{N}(\mu, \Gamma)$. $\mathbf{X}_1$ 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 $\Gamma_{12} = \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$. Réciproquement, supposons que $\Gamma_{12} = 0$. Alors $\mathbf{t}^T \Gamma \mathbf{t} = \mathbf{t}_1^T \Gamma_{11} \mathbf{t}_1 + \mathbf{t}_2^T \Gamma_{22} \mathbf{t}_2$, donc, d'après la proposition IV-5.6

$$\mathbb{E} \left[e^{i\mathbf{t}^T (\mathbf{x}_1^T, \mathbf{x}_2^T)^T} \right] = \varphi_{\mu, \Gamma}(\mathbf{t}) = \exp \left(i\mathbf{t}^T \mu - \frac{1}{2} \mathbf{t}^T \Gamma \mathbf{t} \right) = \mathbb{E} \left[e^{i\mathbf{t}_1^T \mathbf{x}_1} \right] \mathbb{E} \left[e^{i\mathbf{t}_2^T \mathbf{x}_2} \right].$$

Par suite, les variables aléatoires $\mathbf{X}_1$ et $\mathbf{X}_2$ sont indépendantes, ce qui conclut la preuve. $\square$

Corollaire IV-5.9. Soient $\mathbf{A}_1 \in \text{Mat}_{n, n_1}(\mathbb{R})$ et $\mathbf{A}_2 \in \text{Mat}_{n, n_2}(\mathbb{R})$ deux matrices telles que $\mathbf{A}_1^T \mathbf{A}_2 = \mathbf{0}_{n_1 \times n_2}$ et soit $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. 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-5.10. La *décorrélacion* 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 $\mathcal{N}(0, 1)$; $Y = \varepsilon X$, 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 $Y \sim \mathcal{N}(0, 1)$. De plus,

$$\mathbb{E}[XY] = \mathbb{E}[\varepsilon X^2] = \mathbb{E}[\varepsilon] \mathbb{E}[X^2] = 0,$$

et donc $\text{Cov}(X, Y) = 0$, ces variables aléatoires sont décorréliées. 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. $\diamond$

Proposition IV-5.11. Soit Γ une matrice définie positive $n \times n$. Une loi gaussienne de moyenne μ et de covariance Γ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ de la forme :

$$g_{\mu, \Gamma}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{\det(\Gamma)}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Gamma^{-1} (\mathbf{x} - \mu) \right), \quad \mathbf{x} \in \mathbb{R}^n. \quad (\text{IV-5.3})$$

Démonstration. Si $\Gamma = \mathbf{I}_n$ la proposition IV-5.8 montre que les variables aléatoires $X_1, \dots, X_n$ 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-5.6 : il existe $\mathbf{A}$ inversible et $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_n)$ tel que $\mathbf{X} = \mathbf{A} \mathbf{Z} + \mu$ et l'expression IV-5.3 découle de la formule du changement de variable. $\square$

La quantité $[(\mathbf{x} - \mu)^T \Gamma^{-1} (\mathbf{x} - \mu)]^{1/2}$ est souvent appelée la *distance de Mahalanobis* de $\mathbf{x}$ à μ . Les lignes de niveaux de la densité $f_{\mu, \Gamma}$, i.e. les ensembles $\{\mathbf{x} \in \mathbb{R}^n, f_{\mu, \Gamma}(\mathbf{x}) = c\}$ correspondent au lieu des points dont la distance de Mahalanobis à μ est constante. En écrivant $\mathbf{y} = H^T \mathbf{x}$ où H est une matrice unitaire qui diagonalise la matrice Γ , $H^T \Gamma H = D$, $D = \text{diag}(d_1^2, \dots, d_n^2)$, les lignes de niveaux sont donc les ellipsoïdes :

$$\sum_{i=1}^n (y_i - v_i)^2 / d_i^2 = c$$

centrées en $\mathbf{v} = H^T \mu$ et dont les axes principaux sont portés par les vecteurs propres $H = [\mathbf{h}_1, \dots, \mathbf{h}_n]$.

Loi !Gamma
Fonction
Gamma
Loi !du
\$\chi^2\$

IV-5.3 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 $\text{Re}(z) > 0$ par :

$$\Gamma(z) = \int_0^\infty \exp(-t) t^{z-1} dt = 2 \int_0^\infty \exp(-t^2) t^{2z-1} dt. \quad (\text{IV-5.4})$$

En intégrant par partie pour $x > 0$ réel positif l'expression précédente, nous avons :

$$\Gamma(x) = [-t^{x-1} e^{-t}]_0^\infty + (x-1) \int_0^\infty t^{x-2} e^{-t} dt = (x-1) \Gamma(x-1).$$

Donc, pour tout un entier naturel $n \geq 1$, $\Gamma(n) = (n-1)\Gamma(n-1) = \dots = (n-1)(n-2)\dots 1 = (n-1)!$.

Définition IV-5.12. (i) Pour tout réel $p > 0$, on appelle loi Gamma réduite à p degrés de liberté (et l'on note $\text{Gamma}(p)$) 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 $\text{Gamma}(p, \lambda)$, 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 $\text{Gamma}(p, \lambda)$ 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 $x \mapsto \lambda e^{-\lambda x}$ qui coïncide avec la loi $\Gamma(1, \lambda)$.

Si Z est une loi $\text{Gamma}(p)$, (IV-5.4) implique que, pour tout $r > -p$, nous avons pour tout $r > -p$,

$$\mathbb{E}[Z^r] = \frac{\Gamma(p+r)}{\Gamma(p)}. \quad (\text{IV-5.5})$$

La fonction caractéristique de la loi $\text{Gamma}(p, \lambda)$ est donnée par :

$$\phi_{\lambda,p}(t) = \int_0^\infty \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{i t x} e^{-\lambda x} dx = (1 - i t / \lambda)^{-p}. \quad (\text{IV-5.6})$$

Cette expression particulière de la fonction caractéristique a pour conséquence immédiate le lemme de convolution suivant pour les lois Gammas.

Lemme IV-5.13. Soit $(X_1, \dots, X_n)$ n variables aléatoires indépendantes distribuées suivant des lois $\text{Gamma}(p_i, \theta)$ avec $\theta > 0$ et $p_i > 0$, $i \in \{1, \dots, n\}$. Alors, $\sum_{i=1}^n X_i$ est distribuée suivant une loi $\text{Gamma}(\sum_{i=1}^n p_i, \theta)$.

IV-5.4 La loi du χ^2 à k degrés de liberté

Définition IV-5.14 (Loi du χ^2 -centrée ou Khi-deux). Soient $(X_1, \dots, X_v)$, v variables aléatoires gaussiennes de moyenne nulle et de variance unité indépendantes. La variable aléatoire $U = \sum_{i=1}^v X_i^2$ suit une loi du χ^2 centrée à v degrés de liberté, notée χ_v^2 .

Lemme IV-5.15. (i) Soit X une variable aléatoire gaussienne centrée réduite. X^2 suit une loi Gamma($1/2, 1/2$).
(ii) Soit $v \in \mathbb{N}$ et $X_1, X_2, \dots, X_v$ v variables aléatoires gaussiennes centrées réduites. $\sum_{i=1}^v X_i^2$ suit une loi Gamma($v/2, 1/2$).

Démonstration. (i) $\mathbb{P}(X^2 < z) = 0$ si $z < 0$. Pour $z > 0$, nous avons :

$$\begin{aligned}\mathbb{P}(X^2 < z) &= \mathbb{P}(-\sqrt{z} < X < \sqrt{z}) = \int_{-\sqrt{z}}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\ &= 2 \int_0^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \frac{2}{\sqrt{2\pi}} \int_0^z \exp\left(-\frac{u}{2}\right) \frac{1}{2\sqrt{u}} du\end{aligned}$$

Ceci conduit au résultat, en utilisant le résultat élémentaire $\Gamma(1/2) = \sqrt{\pi}$.

(ii) Le résultat est une conséquence élémentaire de Lemme IV-5.13. □

Proposition IV-5.16. La densité de la loi du χ^2 centrée à v degrés de liberté est :

$$f_v(x) = \frac{1}{2^{\frac{v}{2}} \Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-\frac{x}{2}} \mathbb{1}_{\{x>0\}} \quad (\text{IV-5.7})$$

Démonstration. La preuve découle de Lemme IV-5.15 et de l'expression de la densité d'une loi Gamma($v/2, 1/2$). □

Proposition IV-5.17. La moyenne de la loi du χ_v^2 centrée vaut v , sa variance $2v$.

Démonstration. Repartons de la définition de la loi χ_v^2 . Comme $(Z_1, \dots, Z_v)$ sont v variables gaussiennes centrées réduites, on a

$$\begin{aligned}\mathbb{E}[Z_i^2] &= \text{Var}(Z_i) = 1 \quad \text{donc} \quad \mathbb{E}\left[\sum_{i=1}^v Z_i^2\right] = v, \\ \text{Var}(Z_i^2) &= \mathbb{E}[Z_i^4] - (\mathbb{E}[Z_i^2])^2 = 2 \quad \text{donc} \quad \text{Var}\left(\sum_{i=1}^v Z_i^2\right) = 2v. \quad \square\end{aligned}$$

Proposition IV-5.18. Soient $(X_1, \dots, X_k)$, k variables aléatoires gaussiennes de moyenne $(\mu_1, \dots, \mu_k)$ 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 :

$$\left(\frac{1}{\sqrt{1-2it}}\right)^k \exp\left(\frac{it\gamma}{1-2it}\right)$$

(ii) La loi de la variable aléatoire U admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée :

$$f_{k,\gamma}(x) = \sum_{i=0}^{\infty} \frac{e^{-\gamma/2} (\gamma/2)^i}{i!} f_{k+2i}(x)$$

où f_v est la densité d'une loi du χ^2 à q degrés de liberté.

Démonstration. (i) Par définition, si $Z_1, \dots, Z_{k-1}, X$ sont des variables aléatoires indépendantes, $Z_i \sim N(0, 1)$ et $X \sim N(\mu, 1)$ alors la variable aléatoire $U = \sum_{i=1}^{k-1} Z_i^2 + X^2 \sim \chi_k^2(\gamma)$ avec $\gamma = \mu^2/2$. Notons que $\sum_{i=1}^{k-1} Z_i^2$ et X^2 sont indépendantes et que $\sum_{i=1}^{k-1} Z_i^2 \sim \chi_{k-1}^2$. Par conséquent,

$$\mathbb{E}[e^{itU}] = (1 - 2it)^{-(k-1)/2} \mathbb{E}[e^{itX^2}].$$

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-5.19 (Loi du χ^2 non centrée). Soient $(X_1, \dots, X_k)$, k variables aléatoires gaussiennes de moyenne $(\mu_1, \dots, \mu_k)$ 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-5.18. 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-5.20. Soit $\mathbf{Z} \sim N_n(\mu, \sigma^2 \mathbf{I})$ 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 $\|\Pi \mu\|^2$.

Démonstration. Soit $H = [\mathbf{h}_1, \dots, \mathbf{h}_k]$ une base orthonormale de l'image de Π . Nous avons donc $\Pi = HH^T$ et $H^T H = \mathbf{I}_k$ où $\mathbf{I}_k$ 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-5.5 implique que $\mathbf{E} \sim N_k(H^T \mu, \mathbf{I}_k)$. Par conséquent, $\|\mathbf{E}\|^2 \sim \chi_k^2(\delta)$ avec $\delta = \|H^T \mu\|^2 = \mu^T \Pi \mu$, ce qui conclut la preuve. □

IV-5.5 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 :

$$p(\theta, x) = p(\mu, \sigma, x) = \frac{1}{\pi \sigma \left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]} = \frac{1}{\pi} \left[\frac{\sigma}{(x-\mu)^2 + \sigma^2} \right] \quad (\text{IV-5.8})$$

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

$$\Phi_{\theta}(t) = \exp(i\mu t - \sigma|t|).$$

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

$$\exp(i\mu_1 t - \sigma_1|t|) \exp(i\mu_2 t - \sigma_2|t|) = \exp(i(\mu_1 + \mu_2)t - (\sigma_1 + \sigma_2)|t|)$$

et est donc distribuée suivant une loi de Cauchy de paramètre $(\mu_1 + \mu_2, \sigma_1 + \sigma_2)$. En particulier, si $X_1, \dots, X_n$ 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-2.18 car les variables de Cauchy ne sont pas intégrables.

IV-5.6 Loi de Student

Définition IV-5.21. 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-5.22. 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

$$f_{XY}(x, y) \propto e^{-x^2/2} y^{(r/2)-1} e^{-y/2}, \quad x \in \mathbb{R}, y > 0.$$

En appliquant la transformation $\phi : \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R} \times \mathbb{R}^+$, $(x, y) \mapsto (x(y/r)^{-1/2}, y)$, 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 ,

$$f_T(t) \propto \int_0^\infty e^{-y(1+t^2/r)/2} y^{((r+1)/2)-1} dy.$$

et on obtient la formule désirée en faisant le changement de variable $u = y(1 + t^2/r)/2$. □

Loi de
Cauchy

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-5.5).

Proposition IV-5.23. *La suite de loi de Student $\{t(r), r \in \mathbb{N}\}$ 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 $r \rightarrow \infty$, on démontre aisément que pour tout $t \in \mathbb{R}$,

$$\lim_{r \rightarrow \infty} f_r(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2).$$

En effet, pour tout $t \in \mathbb{R}$ nous avons

$$\lim_{r \rightarrow \infty} \left(1 + \frac{t^2}{r}\right)^{-(r+1)/2} = e^{-t^2/2}.$$

En utilisant le développement de Stirling de la fonction Γ

$$\Gamma(z) = \sqrt{2\pi} z^{z-\frac{1}{2}} e^{-z} \left[1 + \frac{1}{12z} + O\left(\frac{1}{z^2}\right)\right].$$

valide pour tout $z \in \mathbb{C} \setminus \mathbb{Z}_-$, on démontre que

$$\lim_{r \rightarrow \infty} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \frac{1}{(r)^{1/2}} = \frac{1}{\sqrt{2}}.$$

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

$$\lim_{r \rightarrow \infty} \int h(t) f_r(t) dt = \int h(t) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt,$$

ce qui montre que la suite de lois $t(r)$ 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 savons (voir Définition IV-5.14) 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-2.18), $r^{-1} \sum_{i=1}^r Z_i^2 \xrightarrow{\mathbb{P}\text{-prob}} 1$ et par le lemme de Slutsky (Lemme IV-2.33) nous avons donc que $T_r \Rightarrow N(0, 1)$. $\square$

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-5.24. Soient $(X_1, \dots, X_n)$ 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 $N(\mu, \sigma^2/n)$.

(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 $\mathbf{X} = (X_1, \dots, X_n) \sim N(\mu \mathbf{1}_n, \sigma^2 \mathbf{I}_n)$ où $\mathbf{1}_n = [1, \dots, 1]^T$. Notons que $\bar{X}_n = n^{-1} \mathbf{1}_n^T \mathbf{X}$ et donc que :

$$(n-1)S_n^2 = \|\mathbf{X} - n^{-1} \mathbf{1}_n \mathbf{1}_n^T \mathbf{X}\|^2 = \|(\mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^T) \mathbf{X}\|^2.$$

Remarquons que

$$\Pi := \mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^T$$

est un projecteur orthogonal de rang $(n-1)$ et $\Pi \mathbf{1}_n = 0$. La proposition IV-5.20 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-5.9 montre que $\bar{X}_n = n^{-1} \mathbf{1}_n^T \mathbf{X}$ et $\Pi \mathbf{X}$ sont indépendants et le résultat découle de : $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0, 1)$. $\square$

Remarque IV-5.25. 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. $\diamond$

IV-5.7 Loi de Fisher

Définition IV-5.26 (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 $F(q, r)$.

Proposition IV-5.27. 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-5.22 et est omise. $\square$

Remarquons que, par définition, si W est distribuée suivant la loi de Fisher $F_{q,r}$ alors $1/W$ est distribuée suivant la loi de Fisher $F_{r,q}$. 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 $(X_1, \dots, X_n)$ des variables aléatoires gaussiennes indépendantes de loi $N(\mu_1, \sigma^2)$ et $(Y_1, \dots, Y_m)$ des variables aléatoires gaussiennes indépendantes de loi $N(\mu_2, \sigma^2)$, indépendantes de $(X_1, \dots, X_n)$. 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,

$$\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2(n-1), \quad \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(m-1),$$

d'où immédiatement :

$$\frac{S_1^2}{S_2^2} \sim F(n-1, m-1).$$

Chapitre IV-6

Famille Exponentielle

Définition IV-6.1 (Famille exponentielle canonique). Soit $(X, \mathcal{X})$ un espace mesurable et $\mu \in \mathbb{M}_+(\mathcal{X}, \mathcal{X})$ une mesure σ -finie. Nous appelons famille exponentielle canonique de dimension d engendrée par $\mathbf{T} = (T^{(1)}, \dots, T^{(d)})$ et la fonction mesurable h la famille de lois indexées par $\eta = (\eta_1, \dots, \eta_d)$ de densité

$$q(\eta, x) = h(x) \exp(\eta^T \mathbf{T}(x) - A(\eta)) \quad , \quad x \in X, \quad (\text{IV-6.1})$$

par rapport à une mesure μ où $A(\eta)$ est défini par :

$$A(\eta) = \log \int h(x) \exp(\eta^T \mathbf{T}(x)) \mu(dx). \quad (\text{IV-6.2})$$

L'espace des paramètres naturels de la famille canonique associée à $(\mathbf{T}, h)$ est l'ensemble

$$\Xi = \left\{ \eta = (\eta_1, \dots, \eta_d) \in \mathbb{R}^d : |A(\eta)| < \infty \right\}. \quad (\text{IV-6.3})$$

Exemple IV-6.2. 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

$$A(\eta) = \log \int_0^\infty e^{\eta x} dx = \begin{cases} \log(-1/\eta), & \eta < 0 \\ \infty, & \eta \geq 0. \end{cases}$$

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. $\diamond$

Il est parfois souhaitable de ne pas se limiter à la paramétrisation canonique.

Définition IV-6.3 (Famille Exponentielle). Soit $(X, \mathcal{X})$ un espace mesurable. Soient μ une mesure σ -finie sur $(X, \mathcal{X})$, Θ un ouvert de $\mathbb{R}^d$, $\mathbf{T} = (T^{(1)}, \dots, T^{(d)})$ des fonctions mesurables de $X \rightarrow \mathbb{R}$, $\boldsymbol{\varphi} = (\varphi^{(1)}, \dots, \varphi^{(d)})$ des fonctions de $\Theta \rightarrow \mathbb{R}$, et h une fonction positive mesurable de $X \rightarrow \mathbb{R}^+$.

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

$$p(\theta, x) = h(x) \exp[\boldsymbol{\varphi}(\theta)^T \mathbf{T} - B(\theta)], \quad x \in X. \quad (\text{IV-6.4})$$

Si nous disposons d'un n -échantillon $X_1, \dots, X_n$ i.i.d. de cette famille exponentielle, la loi jointe est donnée par

$$p(\theta, x_1, \dots, x_n) = \prod_{i=1}^n h(x_i) \exp \left[\sum_{j=1}^d \varphi_j(\theta) T_{n,j}(\mathbf{x}) - nB(\theta) \right]$$

où $T_{n,j}(x_1, \dots, x_n) = \sum_{i=1}^n T_j(x_i)$, et donc la loi du n -échantillon appartient encore à une famille exponentielle de dimension d , indépendamment du nombre n d'échantillons.

Exemple IV-6.4 (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 $x \in \mathbb{X} = \mathbb{N}$, l'ensemble des entiers naturels,

$$p(\theta, x) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!} \exp[x \log(\theta) - \theta], \quad \theta \in \Theta,$$

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

$$q(\eta; x) = \frac{1}{x!} \exp[x\eta - \exp(\eta)]$$

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-6.5 (Loi binômiale). Soit $\{\mathbb{P}_\theta, \theta \in \Theta =]0, 1[\}$ la distribution de la loi binômiale sur $\mathbb{X} = \{0, \dots, n\}$. Pour $\theta \in \Theta$ et $x \in \{0, \dots, n\}$, 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-6.6 (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-6.7 (IV-6.6, 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, $\mathcal{E} = \mathbb{R} \times (\mathbb{R}^- \setminus \{0\})$. ◇

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 $\varphi(\Theta)$ est inclus dans un sous-espace de dimension $l \leq (d-1)$.

IV-6.0.1 Quelques propriétés de la famille exponentielle

Lemme IV-6.8. 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-6.1). Alors

- (i) L'espace des paramètres naturel Ξ est convexe.
- (ii) La fonction $A : \Xi \rightarrow \mathbb{R}$ définie par (IV-6.2) 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 :

$$\int u^\lambda(x) v^{1-\lambda}(x) h(x) \mu(dx) \leq \left(\int u(x) h(x) \mu(dx) \right)^\lambda \left(\int v(x) h(x) \mu(dx) \right)^{1-\lambda}.$$

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

$$\begin{aligned} \int \exp(\{\lambda \eta + (1-\lambda) \tilde{\eta}\}^T \mathbf{T}(x)) h(x) \mu(dx) \\ \leq \left(\int \exp(\eta^T \mathbf{T}(x)) h(x) \mu(dx) \right)^\lambda \left(\int \exp(\tilde{\eta}^T \mathbf{T}(x)) h(x) \mu(dx) \right)^{1-\lambda}, \end{aligned}$$

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 $\lambda \in]0, 1[$:

$$A(\lambda \eta + (1-\lambda) \tilde{\eta}) \leq \lambda A(\eta) + (1-\lambda) A(\tilde{\eta}). \quad \square$$

Théorème IV-6.9. Soit ϕ une fonction mesurable telle que, pour tout $(\eta_1, \dots, \eta_d) \in \Xi$,

$$\int |\phi(x)| \exp\left(\sum_{j=1}^k \eta_j T_j(x)\right) \mu(dx) < \infty. \quad (\text{IV-6.5})$$

Alors la fonction des variables complexes $(s_1, \dots, s_d) \in \mathbb{C}^d$

$$(s_1, \dots, s_d) \mapsto \Phi(s_1, \dots, s_d) = \int \phi(x) \exp\left(\sum_{j=1}^k s_j T_j(x)\right) \mu(dx)$$

est définie en tous points $(s_1, \dots, s_d)$ tels que $(\text{Re}(s_1), \dots, \text{Re}(s_d)) \in \Xi$. De plus

- (i) La fonction Φ est analytique sur le domaine

$$R = \left\{ (s_1, \dots, s_d) \in \mathbb{C}^d : (\text{Re}(s_1), \dots, \text{Re}(s_d)) \in \Xi^o \right\}$$

- (ii) Pour tout $p \in \mathbb{N}$, et tout d -uplet $(i_1, \dots, i_d)$ d'entiers naturels vérifiant $i_1 + \dots + i_d = p$, nous avons

$$\begin{aligned} \int |\phi(x)| |T_1^{i_1}(x) \dots T_d^{i_d}(x)| \exp\left(\sum_{j=1}^k \text{Re}(s_j) T_j(x)\right) \mu(dx) < \infty \\ \frac{\partial^p \Phi(s_1, \dots, s_d)}{\partial s_1^{i_1} \dots \partial s_d^{i_d}} = \int \phi(x) T_1^{i_1}(x) \dots T_d^{i_d}(x) \exp\left(\sum_{j=1}^k s_j T_j(x)\right) \mu(dx). \end{aligned}$$

Démonstration. Nous démontrons le résultat pour $p = 1$, $i_1 = 1$ et $i_2 = \dots = i_d = 0$. Soit $(s_1^0, \dots, s_d^0) \in R$. En décomposant le facteur $\phi(x) \exp(s_2^0 T_2(x) + \dots + s_k^0 T_k(x))$ en partie réelle et imaginaire puis chacune de celles-ci en parties positives et négatives, nous pouvons réécrire

$$\begin{aligned} \Phi(s_1, s_2^0, \dots, s_d^0) &= \int \exp(s_1 T_1(x)) \mu_1(dx) - \int \exp(s_1 T_1(x)) \mu_2(dx) \\ &\quad + i \int \exp(s_1 T_1(x)) \mu_3(dx) - i \int \exp(s_1 T_1(x)) \mu_4(dx) . \end{aligned}$$

Il suffit donc d'établir le résultat pour une intégrale de la forme

$$\psi(s_1) = \int \exp(s_1 T_1(x)) \mu(dx) .$$

Comme $(\operatorname{Re}(s_1^0), \dots, \operatorname{Re}(s_d^0)) \in \Xi^o$, 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

$$\frac{\psi(s_1) - \psi(s_1^0)}{s_1 - s_1^0} = \int \frac{\exp(s_1 T_1(x)) - \exp(s_1^0 T_1(x))}{s_1 - s_1^0} \mu(dx) .$$

On peut écrire l'intégrande sous la forme

$$\exp(s_1^0 T_1(x)) \left[\frac{\exp[(s_1 - s_1^0) T_1(x)] - 1}{s_1 - s_1^0} \right]$$

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

$$\frac{1}{\delta} |\exp(s_1^0 T_1(x) + \delta |T_1(x)|)| \leq \frac{1}{\delta} |\exp[(s_1^0 + \delta) T_1(x)] + \exp[(s_1^0 - \delta) T_1(x)]|$$

pour $|s_1 - s_1^0| \leq \delta$. Comme le terme de droite est intégrable, le théorème de convergence dominée 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

$$\int T_1(x) \exp(s_1^0 T_1(x)) \mu(dx) .$$

Le raisonnement pour les dérivées d'ordre plus élevé est exactement similaire. $\square$

Dans la suite, nous notons, pour toute fonction mesurable $S: X \rightarrow \mathbb{R}$ et $\eta \in \Xi$ vérifiant $\int |S(x)| q(\eta, x) \mu(dx) < \infty$

$$Q_\eta(S) = \int S(x) q(\eta, x) \mu(dx) \tag{IV-6.6}$$

Comme conséquence immédiate du résultat précédent, nous avons :

Corollaire IV-6.10. (i) Pour tout $\eta = (\eta_1, \dots, \eta_d) \in \Xi^o$ et $i, j \in \{1, \dots, d\}$, nous avons

$$Q_\eta(T_i) = \frac{\partial A}{\partial \eta_i}(\eta)$$

$$Q_\eta\{(T_i - Q_\eta(T_i))(T_j - Q_\eta(T_j))\} = \frac{\partial^2 A}{\partial \eta_i \partial \eta_j}(\eta) .$$

(ii) Pour tout $\eta = (\eta_1, \dots, \eta_d) \in \Xi$ et tout $t = (t_1, \dots, t_d) \in \mathbb{R}^d$ tel que $\eta + t = (\eta_1 + t_1, \dots, \eta_d + t_d) \in \Xi$, la fonction génératrice des moments des statistiques $\mathbf{T}$ est donnée par

$$M_\eta(t) = Q_\eta\{\exp[\mathbf{t}^T \mathbf{T}]\} = \exp(A(\eta + t) - A(\eta)) .$$

Définition IV-6.11 (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 $H(\eta) = [H_{i,j}(\eta)]_{1 \leq i,j \leq d}$ est définie positive où

$$H_{i,j}(\eta) = \left(\frac{\partial^2 A}{\partial \eta_i \partial \eta_j}(\eta) \right)_{1 \leq i,j \leq d}. \quad (\text{IV-6.7})$$

Théorème IV-6.12. 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,

$$Q_\eta \left(\left\{ x \in X : \sum_{i=1}^d a_i T_i(X) \neq a_{d+1} \right\} \right) = 1, \quad \text{pour tout } \eta \in \Xi^o. \quad (\text{IV-6.8})$$

- (iii) La fonction $\eta \rightarrow A(\eta)$ est strictement convexe sur Ξ^o , i.e. pour tout $\lambda \in]0, 1[$, $\eta \in \Xi^o$, $\tilde{\eta} \in \Xi^o$,

$$A(\lambda \eta + (1 - \lambda) \tilde{\eta}) < \lambda A(\eta) + (1 - \lambda) A(\tilde{\eta})$$

Remarque IV-6.13. Remarquons que pour $\eta, \tilde{\eta} \in \Xi$ et $A \in \mathcal{B}(X)$,

$$Q_\eta(A) = \int_A q(x; \eta) \mu(dx) = \int_A q(x; \tilde{\eta}) \frac{\exp(\sum_{i=1}^d \eta_i T_i(x) - A(\eta))}{\exp(\sum_{i=1}^d \tilde{\eta}_i T_i(x) - A(\tilde{\eta}))} \mu(dx).$$

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 $A \in \mathcal{B}(X)$ 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

$$\frac{\exp(\sum_{i=1}^d \eta_i T_i(x) - A(\eta))}{\exp(\sum_{i=1}^d \tilde{\eta}_i T_i(x) - A(\tilde{\eta}))} > 0, \quad \text{pour tout } x \in X, \eta, \tilde{\eta} \in \Xi.$$

Par conséquent, si la propriété (IV-6.8) est vérifiée pour $\eta_0 \in \Xi$, alors elle est vérifiée pour tout $\eta \in \Xi$. ◇

Démonstration. (i) $\iff$ (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 :

$$\mathbb{P}_{q(\eta, \cdot)} \left(\sum_{i=1}^d a_i T_i(x) = c \right) = 1.$$

Alors,

$$\text{Var}_{q(\eta, \cdot)}(\mathbf{a}^T \mathbf{T}(X)) = \mathbf{a}^T \text{Cov}_{q(\eta, \cdot)}(\mathbf{T}(X)) \mathbf{a} = 0$$

et la famille exponentielle engendrée par $\mathbf{T}$ et h n'est pas de rang complet.

Réciproquement, si, pour $\mathbf{a} \neq 0$, $\text{Var}_{q(\eta, \cdot)}(\mathbf{a}^T \mathbf{T}(X)) = 0$, 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-6.8) n'est pas vérifiée.

(i) $\iff$ (iii) Si pour tout $\eta \in \Xi$ la matrice $(\partial^2 A(\eta) / \partial \eta_i \partial \eta_j)_{1 \leq i,j \leq d}$, 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 $\text{Var}_{q(\eta_0, \cdot)}(\mathbf{a}^T \mathbf{T}(X)) = 0$, ou, de façon équivalente, il existe $c \in \mathbb{R}$ tel que,

$$\begin{aligned} \mathbb{P}_{q(\eta_0, \cdot)}(\mathbf{a}^T \mathbf{T}(X) = c) &= 1 \Leftrightarrow \mathbb{P}_{q(\eta_0, \cdot)}(\lambda \mathbf{a}^T \mathbf{T}(X) = \lambda c) = 1, \quad \text{pour tout } \lambda \in [0, 1], \\ &\Leftrightarrow \int h(x) \exp(\lambda \mathbf{a}^T \mathbf{T}(x)) \exp(\eta_0^T \mathbf{T}(x) - A(\eta_0)) \mu(dx) = \exp(\lambda c), \\ &\Leftrightarrow \int h(x) \exp(\eta_0 + \lambda \mathbf{a}^T \mathbf{T}(x)) \mu(dx) = \exp(\lambda c + A(\eta_0)). \end{aligned}$$

Donc, pour tout $\lambda \in [0, 1]$, $\eta_0 + \lambda \mathbf{a} \in \Xi$ et

$$A(\eta_0 + \lambda \mathbf{a}) = A(\eta_0) + \lambda c. \quad (\text{IV-6.9})$$

La fonction A n'est donc pas strictement convexe et (iii) n'est pas vérifiée. $\square$

Théorème IV-6.14. Soit $\{q(\eta, \cdot), \eta \in \Xi\}$ la famille exponentielle canonique engendrée par $T_1, \dots, T_d$ et h . Le modèle exponentiel engendré par $\mathbf{T}$ et h est de rang complet si et seulement si l'application

$$(\eta_1, \dots, \eta_d) \mapsto \mathbf{e}(\eta_1, \dots, \eta_d) = \left(\frac{\partial A}{\partial \eta_1}(\eta_1, \dots, \eta_d), \dots, \frac{\partial A}{\partial \eta_d}(\eta_1, \dots, \eta_d) \right) \quad (\text{IV-6.10})$$

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

$$[\mathbf{J}_{\mathbf{e}}(\eta)]_{i,j} = \text{Cov}_{q(\eta, \cdot)}(T_i(X), T_j(X)) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta), \quad 1 \leq i, j \leq d. \quad (\text{IV-6.11})$$

Démonstration. Supposons que le modèle exponentiel engendré par $(\mathbf{T}, h)$ est de rang complet. Le théorème IV-6.12-(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-6.11) découle de la définition de $\mathbf{e}$ et du corollaire IV-6.10. L'expression (IV-6.11) montre que la matrice jacobienne $\mathbf{J}_{\mathbf{e}}(\eta)$ est inversible et, comme $\mathbf{e}$ est injective, le ?? 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-6.12, 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-6.13, 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-6.9), on obtient

$$A(\eta + \lambda \mathbf{a}) = \lambda c + A(\eta) \Rightarrow \mathbf{e}(\eta + \lambda \mathbf{a}) = \mathbf{e}(\eta),$$

ce qui montre que la fonction $\eta \mapsto \mathbf{e}(\eta)$ n'est pas bijective. $\square$

Exemple IV-6.15 (Vecteur gaussien p -dimensionnel). Soit $\mathbf{Y} = (Y_1, \dots, Y_p)$ un vecteur Gaussien de moyenne μ 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

$$f(\mathbf{Y}; \mu, \Sigma) = \det(\Sigma)^{-1/2} (2\pi)^{-p/2} \exp\left(-\frac{1}{2}(\mathbf{Y} - \mu)^T \Sigma^{-1}(\mathbf{Y} - \mu)\right).$$

Par conséquent

$$\log f(\mathbf{Y}; \mu, \Sigma) = -\frac{1}{2} \mathbf{Y}^T \Sigma^{-1} \mathbf{Y} + (\Sigma^{-1} \mu)^T \mathbf{Y} - \frac{1}{2} \left(\log(\det(\Sigma)) + \mu^T \Sigma^{-1} \mu \right) - \frac{p}{2} \log(2\pi),$$

ce qui montre que la famille de loi $N(\mu, \Sigma)$ est une famille exponentielle $p(p+3)/2$ -dimensionnelle, engendrée par la statistique

$$\mathbf{T}(\mathbf{Y}) = (\{Y_i\}_{1 \leq i \leq p}, \{Y_i Y_j\}_{1 \leq i \leq j \leq p}),$$

et la fonction $h(\mathbf{Y}) = 1$. La famille canonique associée à $(\mathbf{T}, h)$ est donnée par

$$q(\mathbf{X}; \beta) = \exp \left(\alpha^T \mathbf{X} + \mathbf{X}^T \Gamma \mathbf{X} - A(\alpha, \gamma) \right),$$

où $\beta = (\alpha, \gamma)$, $\alpha = (\alpha_1, \dots, \alpha_p)$, $\gamma = \{\gamma_{i,j}\}_{1 \leq i \leq j \leq p}$ 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

$$\alpha^T \mathbf{y} + \mathbf{y}^T \Gamma \mathbf{y} = (U \alpha)^T U \mathbf{y} + (U \mathbf{y})^T \Lambda (U \mathbf{y}).$$

Le changement de variable $\mathbf{y} \mapsto U \mathbf{y}$ montre que

$$\int_{\mathbb{R}^p} \exp(\alpha^T \mathbf{y} + \mathbf{y}^T \Gamma \mathbf{y}) d\mathbf{y} = \prod_{i=1}^p \int_{\mathbb{R}} \exp(\tilde{\alpha}_i z_i + \lambda_i z_i^2) dz_i,$$

où $\tilde{\alpha}_i = (U \alpha)_i$, $i \in \{1, \dots, p\}$ et λ_i est le i -ème élément diagonal de Λ . Ce produit d'intégrale est fini si et seulement si $\max_{1 \leq i \leq p} \lambda_i < 0$, et par conséquent l'ensemble des paramètres canoniques est donné par :

$$\Xi = \left\{ (\alpha, \gamma) \in \mathbb{R}^p \times \mathbb{R}^{p(p+1)/2}, \quad \Gamma \text{ est définie négative} \right\}. \quad (\text{IV-6.12})$$

L'ensemble Ξ est ouvert et nous pouvons donc appliquer le théorème IV-6.12. 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 $\mathbf{Z} = (Z_1, \dots, Z_p)$ p v.a. gaussiennes i.i.d. centrées réduites, nous avons

$$\mathbb{P}_{N(0, I)} \left(\mathbf{a}^T \mathbf{Z} + \mathbf{Z}^T B \mathbf{Z} = c \right) = 1, \quad \mathbf{a} \in \mathbb{R}^p, c \in \mathbb{R}, B \in \mathbb{R}^{p \times p},$$

si et seulement si $\|\mathbf{a}\| + \|B\| = 0$, où $\|B\|^2 = \text{Tr}(BB^T)$. Remarquons en effet que, pour $1 \leq i \leq j \leq k \leq l \leq p$,

$$\text{Cov}_{N(0, I)}(Z_i, Z_j) = \delta_{i,j}, \quad \text{Cov}_{N(0, I)}(Z_i, Z_j Z_l) = 0, \quad \text{Cov}_{N(0, I)}(Z_i Z_j, Z_k Z_l) = \delta_{i,k} \delta_{j,l} + \delta_{i,l} \delta_{j,k},$$

et donc que :

$$\text{Var}_{N(0, I)} \left(\mathbf{a}^T \mathbf{Z} + \mathbf{Z}^T B \mathbf{Z} \right) = \|\mathbf{a}\|^2 + \|B\|^2.$$

ce qui établit la relation désirée. Nous avons, pour tout vecteur $\mathbf{a} \in \mathbb{R}^p$ et toute matrice $B \in \mathbb{R}^{p \times p}$,

$$\text{Var}_{q(\beta, \cdot)} \left(\mathbf{a}^T \mathbf{Y} + \mathbf{Y}^T B \mathbf{Y} \right) = \text{Var} \left(\tilde{\mathbf{a}}^T \mathbf{Z} + \mathbf{Z}^T \tilde{B} \mathbf{Z} \right),$$

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

$$\begin{aligned} \text{Var}_{q(\beta, \cdot)} \left(\mathbf{a}^T \mathbf{Y} + \mathbf{Y}^T B \mathbf{Y} \right) &= 0 \Leftrightarrow \tilde{\mathbf{a}} = 0 \quad \text{et} \quad \tilde{B} = 0, \\ &\Leftrightarrow \mathbf{a} = 0 \quad \text{et} \quad B = 0. \end{aligned}$$

Le théorème IV-6.12 montre que la famille $\mathbf{T}$ est de rang $p(p+3)/2$. ◇

section Estimateur des moments et maximum de vraisemblance pour la famille exponentielle

IV-6.0.2 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-6.11). Nous observons un n -échantillon $X_1, \dots, X_n$ 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

$$\hat{\mathbb{P}}_n \mathbf{T} = \frac{1}{n} \sum_{i=1}^n \mathbf{T}(X_i) = Q_\eta(\mathbf{T}) := \int \mathbf{T}(x) q(\eta, x) \mu(dx).$$

En utilisant le corollaire IV-6.10, ce système d'équations s'écrit de façon équivalente

$$\hat{\mathbb{P}}_n \mathbf{T} = \nabla A(\eta). \quad (\text{IV-6.13})$$

Comme nous l'avons établi dans le théorème IV-6.14, la fonction $\mathbf{e}$ définit un difféomorphisme de Ξ^o sur $\mathbf{e}(\Xi^o)$. La loi forte des grands nombres montre que, pour $\eta_0 \in \Xi^o$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{q(\eta_0, \cdot)} \left(\limsup_{n \rightarrow \infty} \hat{\mathbb{P}}_n \mathbf{T} \in \Xi^o \right) = 1. \quad (\text{IV-6.14})$$

D'autre part, sur l'événement $\{\hat{\mathbb{P}}_n \mathbf{T} \in \mathbf{e}(\Xi^o)\}$, nous avons

$$\hat{\eta}_n = \mathbf{e}^{-1}(\hat{\mathbb{P}}_n \mathbf{T}). \quad (\text{IV-6.15})$$

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-6.14, nous obtenons

- (i) La suite d'estimateurs $\{\eta_n, n \in \mathbb{N}\}$ est fortement consistante, pour tout $\eta_0 \in \Xi^o$, $\hat{\eta}_n \xrightarrow{\mathbb{P}_{\eta_0} \text{-p.s.}} \eta_0$
- (ii) La suite d'estimateurs $\{\eta_n, n \in \mathbb{N}\}$ est asymptotiquement normale, pour tout $\eta_0 \in \Xi^o$,

$$\sqrt{n}(\hat{\eta}_n - \eta_0) \xrightarrow{\mathbb{P}_{\eta_0}} \mathcal{N}(0, \{\text{Var}_{q(\eta_0, \cdot)}(\mathbf{T}(X))\}^{-1}). \quad (\text{IV-6.16})$$

On remarque que dans ce cas particulier que, pour tout $\eta \in \Xi^o$,

$$\mathbf{J}_{\mathbf{e}}(\eta) = \text{Var}_{q(\eta, \cdot)}(\mathbf{T}(X))$$

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

$$\eta \mapsto q(\eta, X_1, \dots, X_n) = \prod_{i=1}^n q(\eta, X_i) = \prod_{i=1}^n h(X_i) \cdot \exp \left(\sum_{i=1}^n \eta^T \mathbf{T}(X_i) - nA(\eta) \right)$$

Posons $\ell_n(\eta, x) = \log q(\eta, x)$. Le théorème IV-6.9 montre que la *log-vraisemblance* est indéfiniment différentiable sur Ξ^o . Un calcul élémentaire montre que

$$\nabla \ell(\eta, x) = \mathbf{T}(x) - \nabla A(\eta). \quad (\text{IV-6.17})$$

Nous appelons dans la suite *score de Fisher* le gradient de la log-vraisemblance. En appliquant le corollaire IV-6.10, nous avons, pour tout $\eta \in \Xi^o$,

$$\mathbb{E}_{q(\eta, \cdot)} [\nabla \ell(\eta, X_1)] = \mathcal{Q}_{\eta}(\mathbf{T}) - \nabla A(\eta) = \mathbf{0}_{d \times 1}.$$

Le système d'équations (IV-6.13) définissant l'estimateur des moments peut aussi s'interpréter comme la solution du système d'équations de *vraisemblance* donné par

$$n^{-1} \sum_{i=1}^n \nabla \ell(\eta, X_i) = \mathbf{0}. \quad (\text{IV-6.18})$$

Comme la fonction $\eta \mapsto A(\eta)$ est strictement convexe sur Ξ^o , la solution de (IV-6.18) (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-6.16) montre que l'estimateur du maximum de vraisemblance est asymptotiquement normal. Nous appellerons *matrice d'information de Fisher* la covariance du score de Fisher

$$\mathbb{I}(\eta_0) = \mathcal{Q}_{\eta_0}(\ell_{\eta} \ell_{\eta}^T). \quad (\text{IV-6.19})$$

En utilisant (IV-6.17), l'information de Fisher est donnée par

$$\mathbb{I}(\eta_0) = \text{Var}_{q(\eta_0, \cdot)}(\mathbf{T}(X_1)). \quad (\text{IV-6.20})$$

L'équation (IV-6.16) 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-6.0.3 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-6.3).

$$p(\theta, x) = h(x) \exp [\varphi(\theta)^T \mathbf{T}(x) - B(\theta)], \quad x \in \mathcal{X}, \theta \in \Theta, \quad (\text{IV-6.21})$$

où $\varphi = (\varphi^{(1)}, \dots, \varphi^{(d)})$ et $\mathbf{T} = (T_1, \dots, T_d)$. Notons Ξ l'espace des paramètres naturels (voir Définition IV-6.1) Supposons que l'espace des paramètres Θ est un ouvert de $\mathbb{R}^d$ et que $\varphi = (\varphi_1, \dots, \varphi_d) : \Theta \mapsto \Xi$, est un difféomorphisme de Θ sur Ξ . Dans ce cas, on peut passer des paramètres $\theta = (\theta^{(1)}, \dots, \theta^{(d)})$ aux paramètres canoniques $\eta = (\eta_1, \dots, \eta_d)$ 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-6.2).

L'estimateur des moments associés aux statistiques $\mathbf{T}$ est donc donné par

$$\hat{\theta}_n = \varphi^{-1}(\hat{\eta}_n) \quad (\text{IV-6.22})$$

où $\hat{\eta}_n$ est l'estimateur des moments dans le modèle canonique (IV-6.15). 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-6.21). 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 $\theta \rightarrow \ell(\theta, x)$ est la log-vraisemblance d'une observation, donnée par

$$\ell(\theta, x) = \log p(\theta, x) = \log(h(x)) - \varphi(\theta)^T \mathbf{T}(x) - A(\varphi(\theta)) \quad (\text{IV-6.23})$$

Si le modèle canonique est de rang complet, nous pouvons déduire de (IV-6.16), en appliquant la méthode- δ , la distribution limite de $\sqrt{n}(\hat{\theta}_n - \theta_0)$, où $\hat{\theta}_n$ est défini par (IV-6.22) :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} \mathcal{N}(0, \{J_\varphi(\theta_0)\}^{-1} \mathbb{I}^{-1}(\varphi(\theta_0)) \{J_\varphi(\theta_0)\}^{-T}), \quad (\text{IV-6.24})$$

où

- (i) $\mathbb{I}(\eta)$ est la matrice d'information de Fisher du modèle canonique (IV-6.19) évaluée au point η ,
- (ii) $J_\varphi(\theta_0)$ est la matrice jacobienne de φ évaluée au point θ_0 .

Le score de Fisher (le gradient de la log-vraisemblance (IV-6.23)) est donné, pour tout $\theta \in \Theta$, par

$$\nabla \ell(\theta, x) = J_\varphi(\theta) \{ \mathbf{T}(x) - \nabla A(\varphi(\theta)) \}. \quad (\text{IV-6.25})$$

Notons que, comme pour le modèle exponentiel canonique, l'espérance du score est nulle :

$$\int \nabla \ell(\theta, x) p(\theta, x) \mu(dx) = 0,$$

car nous avons, en utilisant le corollaire IV-6.10

$$\int \mathbf{T}(x) p(\theta, x) \mu(dx) = \int \mathbf{T}(x) q(\varphi(\theta), x) \mu(dx) = A(\varphi(\theta)).$$

La matrice de Fisher du modèle exponentiel, défini comme la matrice de covariance du score est donnée par

$$\mathbb{J}(\theta) = \int \nabla \ell(\theta, x) \nabla \ell(\theta, x)^T p(\theta, x) \mu(dx). \quad (\text{IV-6.26})$$

En utilisant (IV-6.25), nous avons donc

$$\mathbb{J}(\theta) = \{J_\varphi(\theta)\} \mathbb{I}(\phi(\theta)) \{J_\varphi(\theta)\}^T .$$

Par conséquent, (IV-6.27) montre que nous avons

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \mathbb{J}^{-1}(\theta_0)) , \quad (\text{IV-6.27})$$

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-6.16. Soit $\{p(\theta, \cdot), \theta \in \Theta\}$ 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° 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 $\theta_0 \in \Theta$, $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0} - \text{p.s.}} \theta_0$
- (ii) la suite d'estimateurs $\{\hat{\theta}_n, n \in \mathbb{N}\}$ est asymptotiquement normale, i.e. pour tout $\theta_0 \in \Theta$, $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0} - \text{p.s.}} \theta_0$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathbb{P}_{\theta_0}} N(0, \mathbb{J}^{-1}(\theta_0))$$

où $\mathbb{J}(\theta_0)$ est la matrice d'information de Fisher définie par (IV-6.26).

References

- [Anthony and Bartlett, 1999] Anthony, M. and Bartlett, P. L. (1999). *Neural network learning : theoretical foundations*. Cambridge University Press, Cambridge.
- [Gallouët and Herbin, 2013] Gallouët, T. and Herbin, R. (2013). *Mesure, intégration, probabilités*. Ellipses.
- [Garet and Kurtzmann, 2011] Garet, O. and Kurtzmann, A. (2011). *De l'intégration aux probabilités*, volume 470. Ellipses.
- [Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.
- [Hoeffding, 1963] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58 :13–30.
- [Hoeffding, 1967] Hoeffding, W. (1967). On probabilities of large deviations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I : Statistics*, pages 203–219. Univ. California Press, Berkeley, Calif.
- [Rosenblatt, 1960] Rosenblatt, F. (1960). Perceptron simulation experiments. *Proc. I.R.E.*, 48 :301–309.
- [Rosenblatt, 1962] Rosenblatt, F. (1962). *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. Spartan Books, Washington, D.C.
- [Valiant, 1984] Valiant, L. G. (1984). A theory of the learnable. *Commun. ACM*, 27(11) :1134–1142.
- [Vapnik, 2000] Vapnik, V. N. (2000). *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition.