Statistique et apprentissage

Gérard Biau

2014-2015

Ce document contient les notes du cours Statistique et apprentissage donné dans le cadre de la spécialité Probabilités et modèles aléatoires du master Mathématiques et applications de l'UPMC. Son contenu s'inspire pour tout ou partie des ouvrages suivants :

- B. Cadre et C. Vial. (2012). *Statistique mathématique - Master 1 et Agrégation*, Ellipse, Paris.
- V. Rivoirard et G. Stoltz (2012). *Statistique en action (seconde édition)*, Vuibert, Paris.
- L. Devroye, L. Györfi et G. Lugosi (1996). *A probabilistic theory of pattern recognition*, Springer, New York.

Nous invitons le lecteur à se procurer ces trois manuels, dans lesquels il trouvera approfondissements, compléments et exercices variés.

—G. Biau

Table des matières

Chapitre 1

Modélisation statistique

1.1 Problématique

Au cours d'un sondage d'opinion, on interroge n individus :

$$\left.\begin{array}{ll} x_i = 1 & \text{(oui)} \\ x_i = 0 & \text{(non)} \end{array}\right\} \quad i = 1, \ldots, n.$$

Les réponses $x_1, \ldots, x_n$ sont considérées comme des réalisations d'une suite de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées (i.i.d.), de loi commune $\mathscr{B}(\theta)$, avec $\theta \in]0,1[$ inconnu.

Problème : estimer la valeur de θ et, éventuellement, tester si $\theta \geq 1/2$ ou pas.

Cet exemple relève de la statistique inférentielle, dont les deux volets les plus immédiats sont l'estimation et les tests. Bien comprendre ici la différence avec le calcul des probabilités : en probabilités, on suppose la loi connue précisément et on cherche à caractériser le comportement d'une variable aléatoire qui suit cette loi ; en statistique, la démarche est inverse : à partir de la connaissance des réalisations de la variable, que peut-on dire de sa loi ?

1.2 Modèle statistique

Dans la suite, l'espace des observations est $\mathscr{H}^n = \mathscr{H} \times \cdots \times \mathscr{H}$ (n fois), avec $\mathscr{H} \subset \mathbb{R}^d$.

Définition 1. *Un modèle statistique est un couple $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, où $\{P_\theta\}_{\theta\in\Theta}$ est une famille de probabilités sur $(\mathscr{H}^n, \mathscr{B}(\mathscr{H}^n))$. Quand il existe $k \in \mathbb{N}^\star$ tel que $\Theta \subset \mathbb{R}^k$, le modèle est dit paramétrique. Sinon, il est non paramétrique.*

Définition 2. *Pour $\theta \in \Theta$, une observation $\mathbb{X} = (X_1, \ldots, X_n)$ de loi P_θ est une variable aléatoire à valeurs dans $(\mathscr{H}^n, \mathscr{B}(\mathscr{H}^n))$ de loi P_θ. On note $\mathbb{X} \sim P_\theta$.*

Exemple. Dans le cas du sondage, $\mathscr{H}^n = \{0,1\}^n$, $P_\theta = \mathscr{B}(\theta)^{\otimes n}$ et $\Theta =]0,1[$.

Important : L'ensemble Θ est connu, mais le vrai paramètre θ_0 (celui qui a servi à générer l'observation) ne l'est en revanche pas. La démarche statistique consiste précisément à donner de l'information sur ce paramètre (on parle alors d'inférence ou de statistique inférentielle) à partir de l'observation $\mathbb{X} = (X_1, \ldots, X_n)$. L'hypothèse fondamentale est donc qu'il existe un certain $\theta_0 \in \Theta$ tel que $P_{\mathbb{X}} = P_{\theta_0}$.

Le cas le plus fréquent est celui où l'observation $\mathbb{X} = (X_1, \ldots, X_n)$ est issue de n répétitions indépendantes d'une même expérience. Les composantes $X_1, \ldots, X_n$ sont alors des variables aléatoires i.i.d., de même loi Q_θ sur $(\mathscr{H}, \mathscr{B}(\mathscr{H}))$. Le modèle statistique associé $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ s'écrit ainsi $(\mathscr{H}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$, et l'on dit que $\mathbb{X} = (X_1, \ldots, X_n)$ est un n-échantillon (i.i.d.) de loi (commune) Q_θ. Dans ce contexte d'échantillonnage, les variables aléatoires X_i sont souvent appelées, avec un abus de langage, *observations*.

Exemples.

1. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, avec $\theta \in \Theta =]0,1[$. Outre le sondage d'opinion, ce modèle peut aussi représenter n lancers indépendants du jeu de pile ou face.
2. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{P}(\theta)$. Dans ce cas, $\theta \in \Theta = \mathbb{R}_+^\star$.
3. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{E}(\theta)$. Ici encore, $\theta \in \Theta = \mathbb{R}_+^\star$.
4. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathcal{N}(m, \sigma^2)$. Dans ce cas, $\theta = (m, \sigma)$ et $\Theta = \mathbb{R} \times \mathbb{R}_+$. On peut aussi prendre $\theta = (m, \sigma^2)$, en fonction du contexte.

Un modèle statistique doit être construit sans ambiguïté, en n'associant à chaque loi du modèle qu'un seul paramètre. C'est par exemple le cas du modèle $(\mathbb{R}_+^n, \{\mathscr{E}(\theta)^{\otimes n}\}_{\theta>0})$, car l'application $\theta \mapsto \mathscr{E}(\theta)^{\otimes n}$ définie sur $\mathbb{R}_+^\star$

est injective. Cette propriété, appelée *identifiabilité*, ôte toute ambiguïté à l'étape d'approximation, en faisant correspondre à l'observation un, et un seul, paramètre du modèle.

Définition 3. *Le modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est dit identifiable si l'application $\theta \mapsto P_\theta$ définie sur Θ est injective.*

Exemples. Le modèle statistique gaussien $(\mathbb{R}^n, \{\mathcal{N}(m,\sigma^2)^{\otimes n}\}_{m\in\mathbb{R},\sigma>0})$ est identifiable, mais $(\mathbb{R}^n, \{\mathcal{N}(m,\sigma^2)^{\otimes n}\}_{m\in\mathbb{R},\sigma\neq 0})$ ne l'est pas car $\mathcal{N}(m,\sigma^2) = \mathcal{N}(m,(-\sigma)^2)$.

Dans toute la suite du cours, on supposera les modèles identifiables. Terminons ce chapitre introductif par une définition utile :

Définition 4. *Le modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est dit dominé s'il existe une mesure σ-finie ν sur $\mathscr{H}^n$ telle que $P_\theta \ll \nu$ pour chaque θ. La mesure ν est appelée mesure dominante du modèle.*

Exemples. Le modèle gaussien $(\mathbb{R}^n, \{\mathcal{N}(m,\sigma^2)^{\otimes n}\}_{m\in\mathbb{R},\sigma>0})$ et le modèle de Bernoulli du jeu de pile ou face $(\{0,1\}^n, \{\mathscr{B}(\theta)^{\otimes n}\}_{\theta\in]0,1[})$ sont dominés : une mesure dominante du premier est la mesure de Lebesgue sur $\mathbb{R}^n$, et une mesure dominante du second est $(\delta_0+\delta_1)^{\otimes n}$.

Chapitre 2

Estimation paramétrique

Dans tout le chapitre, $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ désigne un modèle statistique paramétrique avec $\mathscr{H} \subset \mathbb{R}^d$ et $\Theta \subset \mathbb{R}^k$. Le paramètre d'intérêt est $g(\theta)$ avec $g : \Theta \to \mathbb{R}^p$ une fonction connue. L'objectif consiste alors à estimer $g(\theta)$ à partir de l'observation $\mathbb{X} = (X_1, \ldots, X_n)$ issue du modèle.

2.1 Estimateur

Définition 5. *Une statistique est une fonction borélienne de l'observation* $\mathbb{X} = (X_1, \ldots, X_n)$*. Un estimateur est une statistique qui prend ses valeurs dans un sur-ensemble de* $g(\Theta)$*.*

Un estimateur est censé approcher le paramètre d'intérêt, le rôle plus général d'une statistique étant de fournir des informations de diverses natures. De ce fait, ils doivent être construits indépendamment du paramètre du modèle, comme l'indique la définition. Un estimateur de $g(\theta)$ est toujours de la forme $\hat{g} = h(X_1, \ldots, X_n)$, où h est une fonction borélienne définie sur $\mathscr{H}^n$.

Remarque. Dans la pratique, c'est la réalisation de l'estimateur qui fournit une estimation du paramètre d'intérêt (on l'appelle parfois l'estimée). Ainsi, si $(x_1, \ldots, x_n)$ est une réalisation de $(X_1, \ldots, X_n)$ de loi P_{θ_0}, on peut calculer, en utilisant l'estimateur $\hat{g}$, l'approximation $\hat{g}(x_1, \ldots, x_n)$ de $g(\theta_0)$.

La question est alors double : 1) Trouver des méthodes d'estimation et 2) Définir ce qu'est un bon estimateur (ça va de pair). Les deux grandes tech-

niques d'estimation paramétriques sont la méthode des moments d'une part et la méthode du maximum de vraisemblance d'autre part.

Dorénavant, $\|\cdot\|$ désigne la norme euclidienne et $\langle\cdot,\cdot\rangle$ représente le produit scalaire usuel. Les vecteurs de $\mathbb{R}^p$ sont assimilés à des matrices colonnes et le symbole $\top$ est utilisé pour la transposition des matrices.

Dans la suite, $\mathbb{E}_\theta$ désigne l'espérance sous la loi P_θ, i.e. pour une variable aléatoire intégrable Z à valeurs dans $\mathbb{R}^p$ et de loi P_θ,

$$\mathbb{E}_\theta Z = \int_{\mathscr{H}^n} Z(x) P_\theta(\mathrm{d}x).$$

De plus, $\mathbb{V}_\theta Z$ désigne la matrice de variance-covariance (ou la variance si $p = 1$) de $Z \in \mathbb{L}^2$ sous la loi P_θ, i.e.

$$\begin{aligned}\mathbb{V}_\theta Z &= \mathbb{E}_\theta \left(Z - \mathbb{E}_\theta Z\right)\left(Z - \mathbb{E}_\theta Z\right)^\top \\ &= \mathbb{E}_\theta Z\, Z^\top - \left(\mathbb{E}_\theta Z\right)\left(\mathbb{E}_\theta Z\right)^\top.\end{aligned}$$

Une statistique $S(\mathbb{X})$ est d'ordre $q \in \mathbb{N}$ si $S(\mathbb{X}) \in \mathbb{L}^q$ pour chaque $\theta \in \Theta$, i.e.

$$\mathbb{E}_\theta \|S(\mathbb{X})\|^q = \int_{\mathscr{H}^n} \|S(x)\|^q P_\theta(\mathrm{d}x) < \infty \ \forall \theta \in \Theta.$$

2.2 Construction d'estimateurs

2.2.1 Méthode des moments

Considérons le cas d'un n-échantillon i.i.d., pour lequel les composantes $X_1, \ldots, X_n$ de l'observation $\mathbb{X}$ sont indépendantes et de même loi Q_θ sur $(\mathscr{H}, \mathscr{B}(\mathscr{H}))$. Dans le cas particulier où le paramètre d'intérêt $g(\theta)$ est un moment de la loi Q_θ ou, par extension, une fonction de plusieurs moments de cette loi, la méthode des moments permet de retrouver des estimateurs naturels, en substituant à Q_θ la mesure empirique construite avec l'échantillon.

Par exemple, si pour chaque $\theta \in \Theta$, $g(\theta)$ est le moment d'ordre 1 de la loi Q_θ, autrement dit si $g(\theta) = \mathbb{E}_\theta(X_1)$, l'estimateur $\hat{g}$ construit avec cette méthode est la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

De même, si les X_i sont réels et pour chaque $\theta \in \Theta$, $g(\theta) = \mathbb{V}_\theta X_1$, i.e.

$$g(\theta) = \mathbb{E}_\theta (X_1 - \mathbb{E}_\theta X_1)^2 = \mathbb{E}_\theta X_1^2 - \mathbb{E}_\theta^2 X_1,$$

l'estimateur $\hat{g}$ obtenu par la méthode des moments est la variance empirique

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

Plus généralement, si $g(\theta) = \Psi(g_1(\theta), \ldots, g_q(\theta))$, où $g_j(\theta) = \mathbb{E}_\theta \Phi_j(X_1)$ et $\mathbb{E}_\theta \|\Phi_j(X_1)\| < \infty$, la méthode des moments suggère de travailler avec l'estimateur

$$\hat{g} = \Psi\left(\frac{1}{n}\sum_{i=1}^n \Phi_1(X_i), \ldots, \frac{1}{n}\sum_{i=1}^n \Phi_q(X_i)\right).$$

Avantage : L'estimateur a souvent de bonnes propriétés, obtenues via la loi des grands nombres ou le théorème central limite.

Inconvénient : Pour utiliser cette méthode, il faut pouvoir exprimer $g(\theta)$ comme une fonction des moments de la loi Q_θ, ce qui n'est pas toujours possible (ou facile).

Exemples.

1. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. Puisque $\theta = \mathbb{E}_\theta X_1$, la méthode des moments suggère $\hat{\theta} = \bar{X}_n$.
2. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{P}(\theta)$, $\theta > 0$. Ici encore, $\theta = \mathbb{E}_\theta X_1$, et l'on choisit donc $\hat{\theta} = \bar{X}_n$. Mais comme $\theta = \mathbb{V}_\theta X_1$, on peut aussi prendre $\hat{\theta} = S_n^2$.
3. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathcal{N}(m, \sigma^2)$, $(m, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Si σ^2 est connu, $\hat{m} = \bar{X}_n$, et si m est connu, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2$ (pourquoi ?). Dans le cas où m et σ^2 sont tous les deux inconnus, $\theta = (m, \sigma^2)$ et la méthode des moments conduit à l'estimateur $\hat{\theta} = (\bar{X}_n, S_n^2)$.
4. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{U}([0, \theta])$, $\theta > 0$. Il est facile de voir que $\theta = 2\mathbb{E}_\theta(X_1)$, d'où l'estimateur $\hat{\theta} = 2\bar{X}_n$.

2.2.2 Méthode du maximum de vraisemblance

On suppose dans ce paragraphe que $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est un modèle statistique dominé par une mesure σ-finie ν, avec $\mathscr{H} \subset \mathbb{R}^d$ et $\Theta \subset \mathbb{R}^k$. Le cas classique est celui où la mesure ν est la mesure de Lebesgue (cas continu) ou bien la mesure de comptage (cas discret).

Lorsque $\mathscr{H}^n$ est discret, la probabilité que l'observation $\mathbb{X} = (X_1, \ldots, X_n)$ soit égale à $(x_1, \ldots, x_n) \in \mathscr{H}^n$ représente le degré de vraisemblance de cette observation pour la loi P_θ. En étendant ce point de vue à tous les modèles, on obtient le concept de vraisemblance défini ci-dessous.

Définition 6. *La vraisemblance du modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est l'application $L_n : \mathscr{H}^n \times \Theta \to \mathbb{R}_+$ telle que, pour chaque $\theta \in \Theta$, $L_n(\cdot;\theta) : \mathscr{H}^n \to \mathbb{R}_+$ est un élément de la classe d'équivalence de la densité de P_θ par rapport à ν.*

Dans un modèle à échantillonnage i.i.d., l'expression de la vraisemblance se simplifie, comme en témoigne le résultat qui suit.

Proposition 1. *Soit L la vraisemblance du modèle $(\mathscr{H}, \{Q_\theta\}_{\theta\in\Theta})$ dominé par la mesure μ. Si, pour chaque $\theta \in \Theta$, $P_\theta = Q_\theta^{\otimes n}$, alors la fonction*

$$\begin{array}{rccc} L_n & : & \mathscr{H}^n \times \Theta & \to \mathbb{R} \\ & & (x_1, \ldots, x_n, \theta) & \mapsto \prod_{i=1}^n L(x_i;\theta) \end{array}$$

est la vraisemblance du modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ pour la mesure dominante $\nu = \mu^{\otimes n}$.

Démonstration. Il suffit de remarquer que, pour chaque $\theta \in \Theta$, l'application

$$(x_1, \ldots, x_n) \mapsto \prod_{i=1}^n L(x_i;\theta),$$

définie sur $\mathscr{H}^n$, est une version de la densité de $Q_\theta^{\otimes n}$ par rapport à $\nu = \mu^{\otimes n}$. □

Les deux cas les plus classiques en échantillonnage i.i.d. sont ceux où μ est la mesure de comptage sur $\mathscr{H}$ (cas discret) ou la mesure de Lebesgue (cas

continu). On utilise alors souvent la notation $\mathbb{P}_\theta(X_1 = x)$ (cas discret) ou $f_\theta(x)$ (cas continu) en lieu et place de $L(x;\theta)$, de sorte que

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

pour le cas discret, et

$$L_n(x_1,\dots,x_n,\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

pour le cas continu.

Exemples.

1. $\mathbb{X} = (X_1,\dots,X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. Le modèle est dominé par la mesure de comptage, et la vraisemblance L_n vaut

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

   pour $(x_1,\dots,x_n) \in \{0,1\}^n$ et $\theta \in]0,1[$.
2. $\mathbb{X} = (X_1,\dots,X_n)$ i.i.d., de loi commune $\mathcal{N}(m,\sigma^2)$, $(m,\sigma^2) \in \mathbb{R} \times \mathbb{R}_+^\star$. Le modèle est dominé par la mesure de Lebesgue sur $\mathbb{R}^n$, et la vraisemblance L_n s'écrit

   $$L_n(x_1,\dots,x_n;m,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^n (x_i-m)^2}{2\sigma^2}\right),$$

   pour $(x_1,\dots,x_n) \in \mathbb{R}^n$, $m \in \mathbb{R}$ et $\sigma^2 > 0$.

L'intuition nous amène à choisir comme estimateur du paramètre du modèle un paramètre qui maximise la vraisemblance. C'est le concept d'estimateur du maximum de vraisemblance.

Définition 7. *Un estimateur du maximum de vraisemblance (EMV) est un estimateur $\hat{\theta}$ qui vérifie*

$$L_n\left(X_1,\dots,X_n;\hat{\theta}(\cdot)\right) = \sup_{\theta\in\Theta} L_n(X_1,\dots,X_n;\theta).$$

Remarques.

1. En pratique, la vraisemblance se maximise en θ à $X_1, \ldots, X_n$ "fixés", et l'éventuel EMV s'écrit comme une fonction de $X_1, \ldots, X_n$.
2. Ni l'existence, ni l'unicité des EMV ne sont en général acquises. De plus, sous réserve d'existence, l'EMV peut ne pas avoir de représentation explicite ; dans ce cas, le recours à une méthode d'optimisation numérique est nécessaire afin de déterminer sa valeur en l'observation.
3. L'EMV, noté $\hat{\theta}$, est donc un estimateur du paramètre θ du modèle. Si le paramètre d'intérêt est $g(\theta)$, avec g une fonction borélienne connue définie sur Θ, on dit par abus que $g(\hat{\theta})$ est l'EMV de $g(\theta)$.
4. Lorsque $\mathbb{X} = (X_1, \ldots, X_n)$ est un n-échantillon i.i.d., on préfère parfois calculer l'EMV en maximisant la log-vraisemblance, c'est-à-dire le logarithme de la vraisemblance, plutôt que la vraisemblance. En effet, d'après la Proposition 1, si L désigne la vraisemblance du modèle $(\mathscr{H}, \{Q_\theta\}_{\theta\in\Theta})$, la log-vraisemblance du modèle $(\mathscr{H}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$ s'écrit
$$\ln L_n(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \ln L(x_i; \theta),$$
pour $(x_1, \ldots, x_n) \in \mathscr{H}^n$ et $\theta \in \Theta$. L'intérêt pratique est clair, l'étape de maximisation étant en principe plus facile à mener.
5. Sous certaines conditions de régularité du modèle, l'EMV possède de bonnes propriétés (existence, unicité, convergence, etc.).

Exemples.

1. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. Dans ce cas,
$$\ln L_n(x_1, \ldots, x_n; \theta) = n\bar{x}_n \ln\theta + n(1-\bar{x}_n)\ln(1-\theta),$$
pour $(x_1, \ldots, x_n) \in \{0,1\}^n$ et $\theta \in]0,1[$. Une étude rapide montre que $\bar{x}_n$ réalise le maximum de la fonction
$$\theta \mapsto \ln L_n(x_1, \ldots, x_n; \theta),$$
définie sur $]0,1[$. Ainsi, $\hat{\theta} = \bar{X}_n$.
2. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{P}(\theta)$, $\theta > 0$. Ici,
$$L_n(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \left(e^{-\theta}\,\frac{\theta^{x_i}}{x_i!}\right) = e^{-n\theta}\theta^{n\bar{x}_n} \prod_{i=1}^{n} \frac{1}{x_i!},$$

et

$$\ln L_n(x_1,\dots,x_n;\theta) = -n\theta + (n\bar{x}_n)\ln\theta - \sum_{i=1}^{n}\ln x_i!,$$

pour $(x_1,\dots,x_n) \in (\mathbb{R}^+)^n$ et $\theta > 0$. Le maximum de cette fonction est atteint en $\bar{x}_n$, de sorte que $\hat{\theta} = \bar{X}_n$.

3. $\mathbb{X} = (X_1,\dots,X_n)$ i.i.d., de loi commune $\mathcal{N}(m,\sigma^2)$, $(m,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+^\star$. On trouve

$$\ln L_n(x_1,\dots,x_n;\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2,$$

pour $(x_1,\dots,x_n) \in \mathbb{R}^n$, $m \in \mathbb{R}$ et $\sigma^2 > 0$. Si σ^2 est connu, $\hat{m} = \bar{X}_n$, tandis que si m est connu, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2$. Lorsque les deux paramètres sont inconnus, on pose $\theta = (m,\sigma^2)$ et on trouve facilement [1] en maximisant la log-vraisemblance que $\hat{\theta} = (\bar{X}_n, S_n^2)$.

4. $\mathbb{X} = (X_1,\dots,X_n)$ i.i.d., de loi commune $\mathscr{U}([0,\theta])$, $\theta > 0$. On écrit, pour $(x_1,\dots,x_n) \in [0,\theta]$ et $\theta > 0$,

$$\begin{aligned} L(x_1,\dots,x_n;\theta) = \theta^{-n}\prod_{i=1}^{n}\mathbb{1}_{[0,\theta]}(x_i) &= \theta^{-n}\prod_{i=1}^{n}\mathbb{1}_{[x_i,+\infty[}(\theta) \\ &= \theta^{-n}\mathbb{1}_{[x_{(n)},+\infty[}(\theta), \end{aligned}$$

où $x_{(n)} = \max(x_1,\dots,x_n)$. On trouve ainsi $\hat{\theta} = X_{(n)}$. Noter que, pour ce modèle, l'estimateur obtenu par la méthode des moments est différent de l'EMV.

5. $\mathbb{X} = (X_1,\dots,X_n)$ i.i.d., de densité commune sur $\mathbb{R}$

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

(modèle de translation de Cauchy). Alors, pour $(x_1,\dots,x_n) \in \mathbb{R}^n$ et $\theta \in \mathbb{R}$,

$$\ln L(x_1,\dots,x_n;\theta) = -\sum_{i=1}^{n}\ln\left(1+(x_i-\theta)^2\right) - n\ln\pi.$$

Cette fonction, qui est dérivable en θ, converge vers $-\infty$ lorsque $\theta \to \pm\infty$. Elle admet donc certainement un maximum, difficile à localiser, qui est un point où sa dérivée s'annule.

1. Ne pas se lancer dans un calcul de matrice Hessienne mais observer que, à σ^2 fixé, le maximum en m est toujours atteint au point $\bar{x}_n$, indépendamment de σ^2.

2.3 Critères de performance en moyenne

La première propriété que l'on puisse exiger d'un estimateur est qu'il se comporte en moyenne comme le paramètre qu'il est censé approcher. C'est le concept de *biais*, qui fait l'objet de la prochaine définition.

Définition 8. *Soit $\hat{g}$ un estimateur d'ordre 1. On appelle biais la fonction $\theta \mapsto \mathbb{E}_\theta \hat{g} - g(\theta)$. L'estimateur $\hat{g}$ est dit sans biais lorsque $\mathbb{E}_\theta \hat{g} = g(\theta)\ \forall \theta \in \Theta$. Il est asymptotiquement sans biais lorsque $\lim_{n\to\infty} \mathbb{E}_\theta \hat{g} = g(\theta)\ \forall \theta \in \Theta$.*

Remarques.

1. Le caractère non biaisé d'un estimateur doit se vérifier pour chaque θ dans Θ.
2. Il s'agit d'un critère parmi d'autres, sur lequel on peut discuter.

Exemples. Dans les exemples qui suivent, on se place dans le cadre d'un n-échantillon $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi $P_\theta = Q_\theta^{\otimes n}$.

1. Si Q_θ admet un moment d'ordre 1, la moyenne empirique $\bar{X}_n$ est toujours un estimateur sans biais de $g(\theta) = \mathbb{E}_\theta X_1$, puisque

$$\mathbb{E}_\theta \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta X_i = \mathbb{E}_\theta X_1 = g(\theta).$$

2. Supposons que $\mathscr{H} \subset \mathbb{R}$ et que la probabilité Q_θ admette un moment d'ordre 2. La variance empirique S_n^2 est alors un estimateur biaisé de $g(\theta) = \mathbb{V}_\theta X_1$. En effet, on a la décomposition :

$$nS_n^2 = \sum_{i=1}^{n} X_i^2 + n(\bar{X}_n)^2 - 2n(\bar{X}_n)^2 = \sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2.$$

Les variables aléatoires $X_1, \ldots, X_n$ étant indépendantes et de même loi, $\mathbb{V}_\theta \bar{X}_n = \mathbb{V}_\theta X_1 / n = g(\theta)/n$ car la variance de X_1 vaut $g(\theta)$. Par suite,

$$\begin{aligned} n\mathbb{E}_\theta S_n^2 &= n\mathbb{E}_\theta X_1^2 - n\mathbb{E}_\theta (\bar{X}_n)^2 \\ &= ng(\theta) + n\mathbb{E}_\theta^2 X_1 - n\mathbb{V}_\theta \bar{X}_n - n\mathbb{E}_\theta^2 \bar{X}_n \\ &= (n-1)g(\theta), \end{aligned}$$

ce qui montre que

$$\mathbb{E}_\theta S_n^2 = \frac{n-1}{n} g(\theta).$$

Voilà pourquoi, lorsque $n > 1$, on considère plutôt l'estimateur

$$S_n^{\star 2} = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

(appelé *variance empirique corrigée*) qui, lui, estime sans biais $g(\theta) = \mathbb{V}_\theta X_1$. Notons également que la variance empirique S_n^2 est asymptotiquement sans biais.

3. Supposons que chaque X_i suive la loi $\mathscr{U}([0,\theta])$, $\theta > 0$. Dans ce modèle, l'estimateur $\hat{\theta} = 2X_n$ obtenu par la méthode des moments est sans biais. Il n'en est pas de même pour l'EMV $\hat{\theta} = X_{(n)}$, car $\mathbb{E}_\theta \hat{\theta} < \theta$ (pourquoi ?). On montre d'ailleurs facilement que $\mathbb{E}_\theta \hat{\theta} = \frac{n}{n+1}\theta$.

La proximité entre l'estimateur et le paramètre d'intérêt peut être évaluée par leur distance dans $\mathbb{L}^2$.

Définition 9. *Soit $\hat{g}$ un estimateur d'ordre 2.*

1. *Pour $\theta \in \Theta$, le risque quadratique de $\hat{g}$ sous P_θ est*

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

2. *$\hat{g}$ est dit préférable à l'estimateur $\hat{g}'$ d'ordre 2 lorsque*

$$\mathcal{R}(\hat{g};\theta) \leq \mathcal{R}(\hat{g}';\theta) \ \forall \theta \in \Theta.$$

On a la relation fondamentale suivante (dite *décomposition biais-variance*) :

$$\mathcal{R}(\hat{g};\theta) = \left\| \mathbb{E}_\theta \hat{g} - g(\theta) \right\|^2 + \mathbb{E}_\theta \left\| \hat{g} - \mathbb{E}_\theta \hat{g} \right\|^2 \ \forall \theta \in \Theta. \tag{2.1}$$

Ainsi,

$$\mathcal{R}(\hat{g};\theta) = \text{biais}^2(\theta) + \mathbb{V}_\theta \hat{g}.$$

L'intérêt de cette décomposition est qu'elle montre que, pour un risque quadratique donné, abaisser le biais revient à augmenter le terme de variation $\mathbb{E}_\theta \| \hat{g} - \mathbb{E}_\theta \hat{g} \|^2$, et réciproquement. Il est alors naturel de s'intéresser aux estimateurs qui minimisent uniformément la variance parmi les estimateurs sans biais de $g(\theta)$. On dit ainsi qu'un estimateur $\hat{g}$ d'ordre 2 est

de *variance uniformément minimum parmi les estimateurs sans biais* (VUMSB) s'il est sans biais et préférable à tout autre estimateur sans biais d'ordre 2. L'existence d'un estimateur VUMSB n'est en général pas acquise. Nous reviendrons sur ce problème dans le Chapitre 6, en retenant pour l'instant l'idée qu'un estimateur est d'autant meilleur que son risque quadratique est faible.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. L'estimateur $\bar{X}_n$ de θ est sans biais, et sa variance vaut

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

car les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{B}(\theta)$. Par suite, d'après la décomposition biais-variance (2.1) :

$$\mathcal{R}(\bar{X}_n; \theta) = \frac{\theta(1-\theta)}{n}.$$

En augmentant n, l'estimateur $\bar{X}_n$ gagne donc en précision. Ce n'est pas le cas pour l'estimateur X_1, de risque quadratique $\mathcal{R}(X_1; \theta) = \theta(1-\theta)$. Comme on pouvait s'y attendre, $\bar{X}_n$ est donc préférable à X_1. Nous montrerons dans le Chapitre 6 que $\bar{X}_n$ est VUMSB.

2.4 Critères de performance asymptotique

L'observation $\mathbb{X} = (X_1, \ldots, X_n)$ contient de plus en plus d'information sur la vraie valeur du paramètre à mesure que sa taille n croît. De ce fait, on est amené à s'intéresser aux propriétés asymptotiques des estimateurs. Dans la suite, sauf mention explicite du contraire, toute propriété de convergence sera entendue pour une taille d'échantillon n qui tend vers l'infini.

Définition 10. *L'estimateur $\hat{g}$ est dit consistant lorsque*

$$\hat{g} \xrightarrow{\mathbb{P}} g(\theta) \ \forall \theta \in \Theta.$$

Exemple. L'estimateur $\bar{X}_n$ de $g(\theta) = \mathbb{E}_\theta X_1$ construit avec un n-échantillon i.i.d. $\mathbb{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{dn}$ satisfaisant $\mathbb{E}_\theta \|X_1\| < \infty$ est consistant car, d'après la loi faible des grands nombres :

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}_\theta X_1 \ \forall \theta \in \Theta.$$

Pour $d = 1$, il en est de même de la variance empirique dès que $\mathbb{E}_\theta X_1^2 < \infty$ puisque, toujours par la loi faible des grands nombres,

$$S_n^2 \xrightarrow{\mathbb{P}} \mathbb{V}_\theta X_1 \ \forall \theta \in \Theta.$$

Remarque. Consistance et absence de biais asymptotique ne sont pas les mêmes notions. Par exemple, pour se convaincre qu'un estimateur consistant n'est pas nécessairement asymptotiquement sans biais, considérons le modèle statistique $(\mathbb{R}^n, \{\mathcal{N}(\theta,1)^{\otimes n}\}_{\theta\in]0,1[})$ et l'estimateur $\hat{\theta}$ de θ issu de $\mathbb{X} = (X_1, \ldots, X_n) \sim P_\theta = \mathcal{N}(\theta,1)^{\otimes n}$ défini par

$$\hat{\theta} = \bar{X}_n + \frac{1}{\Phi(-\sqrt{n})}\, \mathbb{1}_{[\bar{X}_n \leq 0]},$$

où Φ désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$. L'estimateur $\bar{X}_n$ est consistant d'après la loi faible des grands nombres. En outre, comme $\theta > 0$, pour chaque $\varepsilon > 0$:

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{1}{\Phi(-\sqrt{n})}\, \mathbb{1}_{[\bar{X}_n\leq 0]} > \varepsilon\right) = \lim_{n\to+\infty} \mathbb{P}(\bar{X}_n \leq 0) = 0.$$

On en déduit la consistance de $\hat{\theta}$. Or, comme $\bar{X}_n$ suit la loi $\mathcal{N}(\theta, 1/n)$ et $\theta \leq 1$:

$$\mathbb{P}(\bar{X}_n \leq 0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\theta\sqrt{n}} e^{-t^2/2} \mathrm{d}t = \Phi(-\theta\sqrt{n}) \geq \Phi(-\sqrt{n}).$$

En conséquence,

$$\mathbb{E}_\theta\, \hat{\theta} = \mathbb{E}_\theta \bar{X}_n + \frac{1}{\Phi(-\sqrt{n})} \mathbb{P}(\bar{X}_n \leq 0) \geq \theta + 1,$$

donc $\hat{\theta}$ est biaisé, et même asymptotiquement biaisé.

La consistance ne doit être vue que comme une propriété minimale que doit satisfaire un estimateur. Elle ne permet cependant pas de préciser l'erreur commise, d'où la définition qui suit.

Définition 11. *Soit $(v_n)_{n\geq 1}$ une suite de réels positifs telle que $v_n \to +\infty$. L'estimateur $\hat{g}$ est dit de vitesse v_n si, pour chaque $\theta \in \Theta$, il existe une loi $\ell(\theta)$ sur $\mathbb{R}^p$ différente de la loi de Dirac en 0, appelée loi limite de $\hat{g}$, telle que*

$$v_n\, (\hat{g} - g(\theta)) \xrightarrow{\mathscr{L}} \ell(\theta).$$

Si toutes les lois $\ell(\theta)$ sont normales, $\hat{g}$ est dit asymptotiquement normal.

La qualité d'un estimateur est ainsi évaluée sur sa vitesse car il est alors d'autant plus proche de $g(\theta)$ qu'elle est rapide, mais aussi sur la variance de la loi limite, qui doit idéalement être faible afin que l'estimateur se concentre sur le paramètre d'intérêt.

Exemples.

1. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. L'estimateur $\hat{\theta} = \bar{X}_n$ de θ est consistant. Il est aussi asymptotiquement normal de vitesse $\sqrt{n}$ car, pour chaque $\theta \in]0,1[$:

   $$\sqrt{n}\,(\bar{X}_n - \theta) \xrightarrow{\mathscr{L}} \mathcal{N}\,(0, \theta(1-\theta))\,,$$

   d'après le théorème central limite. Noter que la variance de la loi limite prend ses valeurs les plus faibles lorsque θ est proche de 0 ou de 1 et ses valeurs les plus grandes lorsque θ est proche de $1/2$. De ce fait, l'estimation de θ par $\bar{X}_n$ est d'autant meilleure que θ est proche de 0 ou de 1 car la loi limite de l'estimateur $\bar{X}_n$ est alors très peu dispersée.
2. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{P}(\theta)$, $\theta > 0$. Ici encore, l'estimateur $\hat{\theta} = \bar{X}_n$ est consistant et asymptotiquement normal de vitesse $\sqrt{n}$, car

   $$\sqrt{n}\,(\bar{X}_n - \theta) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \theta),$$

   toujours d'après le théorème central limite.
3. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{U}([0,\theta])$, $\theta > 0$. Il est facile de voir (exercice) que l'EMV $\hat{\theta} = X_{(n)}$ est consistant et de vitesse n, car

   $$n(\bar{X}_{(n)} - \theta) \xrightarrow{\mathscr{L}} Z,$$

   où $-Z \sim \mathscr{E}(\theta)$. De ce point de vue, il est plus performant (malgré son caractère biaisé) que l'estimateur $2\bar{X}_n$ obtenu par la méthode des moments, qui ne converge qu'à la vitesse $\sqrt{n}$.

2.5 Asymptotique de l'erreur d'estimation

Pour fixer les idées, supposons dans cette section que l'estimateur $\hat{\theta}$ de θ est de vitesse v_n, i.e.

$$v_n(\hat{\theta} - \theta) \xrightarrow{\mathscr{L}} \ell(\theta)\ \forall \theta \in \Theta,$$

avec $\ell(\theta)$ une loi sur $\mathbb{R}^k$ différente de la mesure de Dirac en 0 et $v_n \to +\infty$.

La loi de l'erreur renormalisée $v_n(\hat{\theta} - \theta)$ est proche de la loi $\ell(\theta)$ pour les grandes valeurs de n. Or, $\ell(\theta)$ est inconnu car θ est inconnu, donc comment peut-on préciser cette erreur d'approximation ? De plus, comment exploiter cette propriété asymptotique lorsque le paramètre d'intérêt est $g(\theta)$? Sous réserve d'hypothèses supplémentaires, nous allons examiner de quelle manière il est possible d'apporter des réponses à ces questions. Commençons au préalable par énoncer le lemme très utile suivant dont la preuve, facile, est laissée au lecteur.

Lemme 1 (Lemme de Slutsky)**.** *Soient $(Z_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ des suites de variables aléatoires à valeurs dans $\mathbb{R}^k$ et $\mathbb{R}^q$ telles que $(Z_n)_{n\geq 1}$ converge en loi vers une variable aléatoire Z et $(Y_n)_{n\geq 1}$ converge en probabilité vers $y \in \mathbb{R}^q$. Alors, $((Z_n, Y_n))_{n\geq 1}$ converge en loi vers (Z, y).*

Le plus souvent, on applique à cette convergence jointe $(Z_n, Y_n) \xrightarrow{\mathscr{L}} (Z, y)$ une fonction continue h (somme, multiplication, etc.), et l'on en tire que $h(Z_n, Y_n) \xrightarrow{\mathscr{L}} h(Z, y)$.

Estimation de θ. Supposons qu'il existe une fonction connue $\sigma : \Theta \to \mathbb{R}^\star$ et une loi connue τ sur $\mathbb{R}^k$ telles que pour chaque $\theta \in \Theta$, $\ell(\theta) = \sigma(\theta)\tau$. Pourvu que l'on dispose d'un estimateur consistant $\hat{\sigma}$ de $\sigma(\theta)$, on déduit du lemme de Slutsky que

$$\left(v_n(\hat{\theta} - \theta), \hat{\sigma}\right) \xrightarrow{\mathscr{L}} (\sigma(\theta)W, \sigma(\theta)) \ \ \forall \theta \in \Theta,$$

si W est une variable aléatoire de loi τ. Comme la convergence en loi est préservée par la composition des fonctions continues,

$$\frac{v_n}{\hat{\sigma}}(\hat{\theta} - \theta) \xrightarrow{\mathscr{L}} W \ \forall \theta \in \Theta.$$

Ainsi, la loi de l'erreur renormalisée $(v_n/\hat{\sigma})(\hat{\theta} - \theta)$ est proche de celle de τ pour les grandes valeurs de n.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0, 1[$. Le théorème central limite donne

$$\sqrt{n}\,(\bar{X}_n - \theta) \xrightarrow{\mathscr{L}} \mathcal{N}\,(0, \theta(1-\theta)) \ \ \forall \theta \in]0, 1[.$$

De plus, $\sqrt{\bar{X}_n(1-\bar{X}_n)}$ est un estimateur consistant de $\sqrt{\theta(1-\theta)}$ d'après la loi des grands nombres. La loi asymptotique de l'erreur renormalisée est donc $\mathcal{N}(0,1)$, car

$$\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}\,(\bar{X}_n-\theta) \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)\ \forall\theta\in\,]0,1[.$$

Ce résultat peut alors être exploité pour encadrer le paramètre inconnu θ.

Estimation de $g(\theta)$. Revenons au problème plus général de l'estimation du paramètre $g(\theta)$. Comme l'indique le résultat qui suit, le calcul de la vitesse de l'estimateur $g(\hat{\theta})$ est immédiat, sous réserve que g possède les propriétés analytiques adéquates.

Théorème 1 (δ-méthode)**.** *Soient $(v_n)_{n\geq 1}$ une suite de réels qui tend vers $+\infty$, $z\in\mathbb{R}^k$ et $(Z_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^k$ telle que $v_n(Z_n-z)$ converge en loi vers une loi ℓ. Si $g:\mathbb{R}^k\to\mathbb{R}^p$ est de classe $\mathscr{C}^1$, de matrice jacobienne J_g, $v_n(g(Z_n)-g(z))$ converge en loi vers $\mathrm{J}_g(z)\ell$.*

Ainsi, si la fonction g est de classe $\mathscr{C}^1$, on a, pour tout $\theta\in\Theta$,

$$v_n\left(g(\hat{\theta})-g(\theta)\right)\xrightarrow{\mathscr{L}}\mathrm{J}_g(\theta)\ell(\theta),$$

$J_g(\theta)$ désignant la matrice jacobienne de g évaluée en θ. De ce fait, $g(\hat{\theta})$ est, comme $\hat{\theta}$, un estimateur de vitesse v_n dès que la loi $\mathrm{J}_g(\theta)\ell(\theta)$ est différente de la loi de Dirac en 0.

Comme dans la partie précédente, on peut préciser l'erreur commise en approchant $g(\theta)$ par $g(\hat{\theta})$ au moins lorsqu'il existe une fonction $\sigma:\Theta\to\mathbb{R}^\star$ et une loi τ sur $\mathbb{R}^k$ telles que, pour chaque $\theta\in\Theta$, $\ell(\theta)=\sigma(\theta)\tau$. En effet, si $J_g(\theta)$ est une matrice carrée (donc $k=p$) inversible pour chaque $\theta\in\Theta$ et $\hat{\sigma}$ est un estimateur consistant de $\sigma(\theta)$, on déduit du lemme de Slutsky que

$$\frac{v_n}{\hat{\sigma}}J_g(\hat{\theta})^{-1}\left(g(\hat{\theta})-g(\theta)\right)\xrightarrow{\mathscr{L}}\tau\ \forall\theta\in\Theta,$$

car, g étant de classe $\mathscr{C}^1$, $J_g(\hat{\theta})$ est un estimateur consistant de $J_g(\theta)$. La loi de l'erreur renormalisée $(v_n/\hat{\sigma})J_g(\hat{\theta})^{-1}(g(\hat{\theta})-g(\theta))$ est donc proche de celle de τ pour les grandes valeurs de n.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. Si $g(\theta) = \theta(1-\theta)$ est le paramètre d'intérêt, la méthode des moments nous conduit à considérer l'estimateur $g(\bar{X}_n)$. Le théorème central limite et la δ-méthode donnent alors

$$\sqrt{n}\left(g(\bar{X}_n) - g(\theta)\right) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, \theta(1-\theta)(1-2\theta)^2\right) \ \forall \theta \in]0,1[.$$

Puis, le lemme de Slutsky et la loi des grands nombres montrent que

$$\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)(1-2\bar{X}_n)^2}}\left(g(\bar{X}_n) - g(\theta)\right) \xrightarrow{\mathscr{L}} \mathcal{N}(0,1) \ \forall \theta \in]0,1[.$$

Remarque. L'utilisation de la δ-méthode ne se limite pas à l'obtention de lois limites pour les estimateurs de $g(\theta)$. Pour s'en convaincre, considérons un n-échantillon $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{P}(\theta)$, $\theta > 0$. Dans ce contexte, la variance empirique $\hat{\theta} = S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2$ est un estimateur convergent de θ. Le théorème central limite multivarié (appliqué au couple de variables aléatoires $(\frac{1}{n}\sum_{i=1}^n X_i^2, \bar{X}_n)$) et la δ-méthode (appliquée avec la fonction $g(x,y) = x - y^2$) conduisent alors (exercice) à

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \theta + 2\theta^2) \ \forall \theta > 0.$$

Démonstration du Théorème 1. Notons Z une variable aléatoire de loi ℓ et ψ la fonction définie pour tout $y \in \mathbb{R}^k$ par

$$\psi(y) = \int_0^1 \mathrm{J}_g\left(z + u(y-z)\right) \mathrm{d}u.$$

Comme $(1/v_n, v_n(Z_n - Z))$ converge en loi vers le couple $(0, Z)$ d'après le lemme de Slutsky, $Z_n - z = (1/v_n)\, v_n(Z_n - z)$ tend vers 0 en probabilité. Or, ψ est continue car g est de classe $\mathscr{C}^1$, d'où $\psi(Z_n)$ converge en probabilité vers $\psi(z) = \mathrm{J}_g(z)$. Par suite, d'après le lemme de Slutsky,

$$(\psi(Z_n), v_n(Z_n - Z)) \xrightarrow{\mathscr{L}} (\mathrm{J}_g(z), Z).$$

La formule de Taylor avec reste intégral nous donne, pour tout $y \in \mathbb{R}^k$:

$$g(y) - g(z) = \psi(y)(y - z).$$

La convergence en loi étant préservée par la composition par des fonctions continues,

$$v_n\left(g(Z_n) - g(z)\right) = \psi(Z_n) v_n(Z_n - z) \xrightarrow{\mathscr{L}} \mathrm{J}_g(z) Z,$$

d'où le lemme. □

Chapitre 3

Intervalles de confiance

3.1 Principe général

Ce chapitre est consacré à la construction d'intervalles contenant le paramètre inconnu supposé réel, avec un niveau de confiance fixé. La théorie se généralise sans problème au cas multivarié, en remplaçant les intervalles par des régions de confiance.

Dans la suite, $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est un modèle statistique paramétrique avec $\mathscr{H} \subset \mathbb{R}^k$ et $\Theta \subset \mathbb{R}^d$. Le paramètre d'intérêt est $g(\theta)$, avec $g : \Theta \to \mathbb{R}$ une fonction connue.

Définition 12. *Soit $\alpha \in]0,1[$. Un intervalle de confiance pour $g(\theta)$ de niveau $(1-\alpha)$ est une statistique I à valeurs dans les intervalles de $\mathbb{R}$ telle que pour chaque $\theta \in \Theta$,*

$$\mathbb{P}\left(g(\theta) \in I\right) \geq 1-\alpha.$$

On distingue parfois les intervalles de confiance de niveau *exactement* $(1-\alpha)$ des intervalles de confiance de niveau supérieur ou égal à $(1-\alpha)$, qui sont alors dits *par excès*. Noter que les deux critères de qualité d'un intervalle de confiance, i.e. sa longueur et son niveau, s'opposent, et qu'il est donc impératif de réaliser un compromis. En pratique, pour un niveau de confiance raisonnable (souvent 90, 95 ou 99%), on cherche un intervalle de confiance de plus petite longueur.

On veillera à ne pas confondre l'intervalle de confiance (qui est aléatoire)

et sa réalisation (qui ne l'est pas). Une erreur fréquente consiste à écrire

$$\mathbb{P}\left(\theta \in [-1.3, 2.5]\right) = 0.95,$$

ce qui n'a évidemment pas de sens, cette probabilité valant en fait 0 ou 1.

L'un des ingrédients de base pour construire un intervalle de confiance est le quantile d'une loi sur $\mathbb{R}$.

Définition 13. *Soit F la fonction de répartition d'une loi ν sur $\mathbb{R}$. Le quantile d'ordre $p \in]0,1[$ de la loi ν est défini par*

$$q_p = \inf\left\{x \in \mathbb{R} : F(x) \geq p\right\}.$$

Il n'est en général pas possible de calculer un quantile de façon exacte, mais la plupart des logiciels de mathématiques sont dotés de fonctions consacrées au calcul des valeurs approchées des quantiles des lois usuelles. Pour la loi $\mathcal{N}(0,1)$, il faut se souvenir que $q_{0.975} \approx 1.960$ et $q_{0.95} \approx 1.645$.

La recherche d'une variable aléatoire pivotale, i.e. une variable aléatoire dont la loi est indépendante de celle de l'observation, joue un rôle essentiel dans le mécanisme de construction des intervalles de confiance.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. Si $\alpha \in]0,1[$, construisons un intervalle de confiance de niveau $(1-\alpha)$ pour le paramètre θ. Soit Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$ et $q_{1-\frac{\alpha}{2}}$ son quantile d'ordre $(1-\frac{\alpha}{2})$. Comme $\sqrt{n}(\bar{X}_n - \theta)$ est une variable aléatoire pivotale de loi $\mathcal{N}(0,1)$,

$$\begin{aligned}\mathbb{P}\Big(\sqrt{n}\,|\bar{X}_n - \theta| \leq q_{1-\frac{\alpha}{2}}\Big) &= \Phi(q_{1-\frac{\alpha}{2}}) - \Phi(-q_{1-\frac{\alpha}{2}}) \\ &= 2\Phi(q_{1-\frac{\alpha}{2}}) - 1 = 1-\alpha,\end{aligned}$$

car la densité de la loi $\mathcal{N}(0,1)$ est paire. Ainsi,

$$\mathbb{P}\left(\theta \in \left[\bar{X}_n - \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X}_n + \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right]\right) = 1-\alpha,$$

c'est-à-dire que $[\bar{X}_n - \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X}_n + \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}]$ est un intervalle de confiance, dit *bilatère*, de niveau $(1-\alpha)$ pour le paramètre θ. D'autres intervalles possibles sont $]-\infty, \bar{X}_n + \frac{q_{1-\alpha}}{\sqrt{n}}]$ (*unilatère à gauche*) et $[\bar{X}_n - \frac{q_{1-\alpha}}{\sqrt{n}}, +\infty[$ (*unilatère à droite*).

L'exemple ci-dessus illustre la méthode dite du *pivot*, qui se résume de la façon suivante :

1. On trouve un estimateur du paramètre dont on connaît la loi, fonction de θ.
2. On transforme cet estimateur pour obtenir une statistique pivotale, dont la loi ne dépend plus de θ.
3. A partir des quantiles de cette loi, on essaie de trouver un intervalle de confiance de niveau adéquat pour θ.

3.2 Utilisation des inégalités de probabilité

Pour toute la suite de cette section, on se place dans le cadre d'un n-échantillon $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune Q_θ sur $\mathscr{H} \subset \mathbb{R}$, de support $[a, b]$ indépendant de θ. On suppose de plus que le paramètre d'intérêt vérifie $g(\theta) = \mathbb{E}_\theta X_1$, et on utilise la moyenne empirique pour estimer $g(\theta)$ (méthode des moments).

A partir de l'inégalité de Bienaymé-Tchebytchev, on montre facilement que

$$I_1 = \left[\bar{X}_n - \frac{b-a}{2\sqrt{n\alpha}}, \bar{X}_n + \frac{b-a}{2\sqrt{n\alpha}} \right]$$

est un intervalle de confiance (par excès) pour $g(\theta)$ de niveau $(1-\alpha)$, avec $\alpha \in]0, 1[$.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0, 1[$. Ici, $a = 0$, $b = 1$, et on trouve

$$I_1 = \left[\bar{X}_n - \frac{1}{2\sqrt{n\alpha}}, \bar{X}_n + \frac{1}{2\sqrt{n\alpha}} \right].$$

Typiquement, $1 - \alpha = 0.95$ et

$$\frac{1}{2\sqrt{n\alpha}} = \frac{\sqrt{5}}{\sqrt{n}} \leq \frac{2.24}{\sqrt{n}}.$$

L'intervalle I_1 peut être amélioré en basant sa construction sur une inégalité plus précise, par exemple l'inégalité de Hoeffding, qui fait l'objet du prochain théorème.

Théorème 2 (Inégalité de Hoeffding). *Soit $Z_1, \ldots, Z_n$ des variables aléatoires réelles indépendantes telles que $a_i \leq Z_i \leq b_i$ $\mathbb{P}$-p.s. ($a_i < b_i$). Alors, pour tout $\varepsilon > 0$,*

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

Construisons avec cette inégalité un intervalle de confiance par excès de niveau $(1-\alpha)$ pour le paramètre $g(\theta)$. Puisque les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi avec $X_i \in [a, b]$ $\mathbb{P}$-p.s. et $\mathbb{E}_\theta X_1 = g(\theta)$, l'inégalité de Hoeffding donne

$$\begin{aligned}\mathbb{P}\left(|\bar{X}_n - g(\theta)| \geq \varepsilon\right) &= \mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}(X_i - \mathbb{E}_\theta X_i)\right| \geq \varepsilon\right) \\ &\leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right),\end{aligned}$$

pour chaque $\varepsilon > 0$. Avec le choix

$$\varepsilon = (b-a)\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$$

on trouve $\mathbb{P}(|\bar{X}_n - g(\theta)| \geq \varepsilon) \leq \alpha$. Par suite,

$$I_2 = \left[\bar{X}_n - (b-a)\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}, \bar{X}_n + (b-a)\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}\right]$$

est un intervalle de confiance (par excès) pour $g(\theta)$ de niveau $(1-\alpha)$. Comparé à l'intervalle I_1 obtenu avec l'inégalité de Bienaymé-Tchebytchev, les contributions de la taille de l'échantillon, en $1/\sqrt{n}$, et de la longueur du support de Q_θ sont les mêmes. En revanche, l'amélioration est nette en ce qui concerne l'influence de α car

$$\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}} \ll \frac{1}{\sqrt{2n\alpha}}$$

pour les petites valeurs de α.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. On obtient

$$I_2 = \left[\bar{X}_n - \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}, \bar{X}_n + \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}\right].$$

Avec $1-\alpha = 0.95$,

$$\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}} \leq \frac{1.36}{\sqrt{n}}.$$

Preuve du Théorème 2. Supposons pour simplifier que les Z_i sont centrées, et notons $S_n = \sum_{i=1}^n Z_i$. Pour tout $r > 0$,

$$\begin{aligned}\mathbb{P}\left(|S_n| \geq \varepsilon\right) &= \mathbb{P}(S_n \geq \varepsilon) + \mathbb{P}(-S_n \geq \varepsilon) \\ &= \mathbb{P}\left(e^{rS_n} \geq e^{r\varepsilon}\right) + \mathbb{P}\left(e^{-rS_n} \geq e^{r\varepsilon}\right).\end{aligned}$$

On en déduit, en utilisant l'inégalité de Markov que

$$\mathbb{P}\left(|S_n| \geq \varepsilon\right) \leq e^{-r\varepsilon}\left\{\mathbb{E}e^{rS_n} + \mathbb{E}e^{-rS_n}\right\} \leq e^{-r\varepsilon}\left\{\sum_{i=1}^n \mathbb{E}e^{rZ_i} + \sum_{i=1}^n \mathbb{E}e^{-rZ_i}\right\},$$

$\mathbb{E}$ désignant l'espérance sous la probabilité $\mathbb{P}$. Majorons maintenant chaque terme $\mathbb{E}e^{sZ_i}$, pour $s = r$ ou $s = -r$. Par convexité de la fonction exponentielle et comme $Z_i \in [a_i, b_i]$ $\mathbb{P}$-p.s.,

$$e^{sZ_i} = \exp\left(\frac{Z_i - a_i}{b_i - a_i}\, sb_i + \frac{b_i - Z_i}{b_i - a_i}\, sa_i\right) \leq \frac{Z_i - a_i}{b_i - a_i}\, e^{sb_i} + \frac{b_i - Z_i}{b_i - a_i}\, e^{sa_i}.$$

Puisque Z_i est centrée, il vient

$$\mathbb{E}e^{sZ_i} \leq -\frac{a_i}{b_i - a_i}\, e^{sb_i} + \frac{b_i}{b_i - a_i}\, e^{sa_i}.$$

Or, en posant $p_i = -a_i/(b_i - a_i)$, on trouve la représentation :

$$-\frac{a_i}{b_i - a_i}\, e^{sb_i} + \frac{b_i}{b_i - a_i}\, e^{sa_i} = \exp\left[-p_i s(b_i - a_i) + \ln\left(1 - p_i + p_i e^{s(b_i - a_i)}\right)\right].$$

Par suite, si $\phi(x) = -p_i x + \ln(1 - p_i + p_i e^x)$ pour tout $x \in \mathbb{R}$,

$$\mathbb{E}e^{sZ_i} \leq e^{\phi(s(b_i - a_i))}.$$

La fonction ϕ est de classe $\mathscr{C}^2$ et vérifie $\phi(0) = \phi'(0) = 0$ et $\phi''(x) \leq 1/4$ pour tout x. D'après la formule de Taylor-Lagrange, il existe donc $\kappa \in [0, s(b-a)]$ (si $s = r$) ou $\kappa \in [s(b-a), 0]$ (si $s = -r$) tel que

$$\phi\left(s(b_i - a_i)\right) = \frac{s^2(b_i - a_i)^2}{2}\phi''(\kappa),$$

d'où $\phi\left(s(b_i-a_i)\right) \leq s^2(b_i-a_i)^2/8$ et $\mathbb{E}e^{sZ_i} \leq e^{r^2(b_i-a_i)^2/8}$ car $s^2 = r^2$. Il s'ensuit que pour chaque $r > 0$,

$$\mathbb{P}\left(|S_n| \geq \varepsilon\right) \leq 2\exp\left(-r\varepsilon + \frac{r^2\sum_{i=1}^n (b_i-a_i)^2}{8}\right).$$

Finalement, le choix $r = 4\varepsilon/\sum_{i=1}^n (b_i-a_i)^2$, qui minimise le terme de droite dans l'inégalité ci-dessus, nous donne l'inégalité annoncée. □

3.3 Intervalles de confiance asymptotiques

A défaut d'informations suffisantes ou appropriées sur la loi de la variable aléatoire utilisée pour la construction de l'intervalle de confiance, une alternative consiste à se retrancher sur une propriété asymptotique.

Définition 14. *Soit $\alpha \in]0,1[$. Un intervalle de confiance asymptotique pour $g(\theta)$ de niveau $(1-\alpha)$ est une statistique I_n à valeurs dans les intervalles de $\mathbb{R}$ telle que pour chaque $\theta \in \Theta$,*

$$\lim_{n\to+\infty} \mathbb{P}\left(g(\theta) \in I_n\right) \geq 1-\alpha.$$

En pratique, on écrit $\mathbb{P}(g(\theta) \in I_n) \approx 1-\alpha$ et on raisonne comme si n était fixé. Il s'agit d'un abus.

Supposons maintenant que l'on veuille construire un intervalle de confiance asymptotique de niveau $(1-\alpha)$ dans le cas où l'estimateur $\hat{g}$ de $g(\theta)$ est asymptotiquement normal et de vitesse $(v_n)_{n\geq 1}$: pour chaque $\theta \in \Theta$, il existe $\sigma^2(\theta) > 0$ tel que

$$v_n\left(\hat{g} - g(\theta)\right) \xrightarrow{\mathscr{L}} \mathcal{N}\big(0, \sigma^2(\theta)\big).$$

Par suite,

$$\frac{v_n}{\sigma(\theta)}\left(\hat{g} - g(\theta)\right) \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

La variable aléatoire $v_n(\hat{g} - g(\theta))/\sigma(\theta)$ est dite asymptotiquement pivotale, car sa loi limite est indépendante de celle de l'observation. Cependant, un tel résultat ne permet pas en général de construire un intervalle de

confiance asymptotique pour $g(\theta)$. Si $\hat{\sigma}$ est un estimateur consistant de $\sigma(\theta)$, le lemme de Slutsky montre que pour chaque $\theta \in \Theta$,

$$\frac{v_n}{\hat{\sigma}} (\hat{g} - g(\theta)) \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

En désignant par $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi $\mathcal{N}(0,1)$, on en déduit que

$$\begin{aligned}\lim_{n\to+\infty} \mathbb{P}\left(\frac{v_n}{\hat{\sigma}}|\hat{g} - g(\theta)| \leq q_{1-\frac{\alpha}{2}}\right) &= \Phi(q_{1-\frac{\alpha}{2}}) - \Phi(-q_{1-\frac{\alpha}{2}}) \\ &= 2\Phi(q_{1-\frac{\alpha}{2}}) - 1 \\ &= 1-\alpha,\end{aligned}$$

avec Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$. Nous avons ainsi montré que $[\hat{g} - q_{1-\frac{\alpha}{2}}\hat{\sigma}/v_n, \hat{g} + q_{1-\frac{\alpha}{2}}\hat{\sigma}/v_n]$ est un intervalle de confiance asymptotique de niveau $(1-\alpha)$ pour $g(\theta)$.

Exemples.

1. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. On sait que

$$\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}} (\bar{X}_n - \theta) \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

Par suite, en notant $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi $\mathcal{N}(0,1)$,

$$I_3 = \left[\bar{X}_n - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

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

Avantage. Marge d'erreur plus faible :

$$1.96\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq \frac{1}{\sqrt{n}}.$$

Inconvénient. Caractère asymptotique.

Application. Marge d'erreur en $1/\sqrt{n}$: pour un échantillon de 1000 personnes (cas des sondages politiques), la précision est de l'ordre de $\pm 3\%$ (seulement).

2. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{E}(\theta)$, $\theta > 0$. Comme $\theta = 1/\mathbb{E}_\theta X_1$, on estime θ à l'aide de l'estimateur $1/\bar{X}_n$ (méthode des moments). Le théorème central limite et la δ-méthode conduisent alors à
$$\frac{\sqrt{n}}{\theta}\left(\frac{1}{\bar{X}_n} - \theta\right) \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$
On s'en sort sans estimer la variance en encadrant θ. On trouve ainsi, en notant $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi $\mathcal{N}(0,1)$, que
$$\left[\frac{1}{\bar{X}_n}\frac{1}{1+\frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}}, \frac{1}{\bar{X}_n}\frac{1}{1-\frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}}\right]$$
est un intervalle de confiance asymptotique de niveau $(1-\alpha)$ pour θ.

En supposant toujours
$$v_n\left(\hat{g} - g(\theta)\right) \xrightarrow{\mathscr{L}} \mathcal{N}\big(0, \sigma^2(\theta)\big),$$
on peut essayer d'identifier une fonction φ de classe $\mathscr{C}^1$ sur Θ telle que $\varphi'(\theta) = 1/\sigma(\theta)$ pour tout θ. En effet, d'après la δ-méthode,
$$\sqrt{n}\big(\varphi(\hat{g}) - \varphi\left(g(\theta)\right)\big) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0,1\right),$$
et cela conduit donc à l'intervalle de confiance asymptotique de niveau $1-\alpha$ pour $g(\theta)$
$$\left[\varphi^{-1}\left(\varphi(\hat{g}) - \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right), \varphi^{-1}\left(\varphi(\hat{g}) + \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)\right].$$
Cette technique porte le nom de *stabilisation de la variance*.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. Comme
$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, \theta(1-\theta)\right),$$
on cherche une fonction φ de classe $\mathscr{C}^1$ sur $]0,1[$ telle que
$$\varphi'(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} \ \forall \theta \in]0,1[.$$
On trouve $\varphi(\theta) = 2\arcsin\sqrt{\theta}$, ce qui conduit à l'intervalle de confiance asymptotique
$$\left[\sin^2\left(\arcsin\sqrt{\bar{X}_n} - \frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right), \sin^2\left(\arcsin\sqrt{\bar{X}_n} + \frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right)\right].$$

Chapitre 4

Tests statistiques

4.1 Formalisme et démarche expérimentale

La notion de test, très importante, est sans doute aussi l'une des plus difficiles à appréhender. Elle est liée au problème suivant : prendre une décision qui ne permet que deux choix possibles, comme "oui" ou "non" et que l'on peut toujours désigner, par commodité, par 0 et 1. Pour une élection avec seulement deux candidats, on souhaite déterminer lequel sera élu. Pour un contrôle de qualité, il s'agit de décider si un processus de fabrication correspond ou non aux spécifications requises. Pour un essai médical, on veut savoir si un médicament est suffisamment efficace, ou plus efficace qu'un autre, ou si ses effets secondaires sont admissibles, etc.

Dans le cadre du modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, on se donne deux sous-ensembles Θ_0 et Θ_1, disjoints et inclus dans Θ (on n'impose pas que leur union soit égale à Θ). Au vu d'une observation $\mathbb{X} = (X_1, \ldots, X_n) \sim P_\theta$, on veut décider si θ_0 (le vrai paramètre) appartient à Θ_0 ou pas. Dans la négative, on considère alors que $\theta_0 \in \Theta_1$.

On définit respectivement

- ▷ $H_0 : \theta \in \Theta_0$ (l'hypothèse *nulle*) ;
- ▷ $H_1 : \theta \in \Theta_1$ (l'hypothèse *alternative*).

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0, 1[$. S'il s'agit d'un jeu de pile ou face, le problème de test associé à la question "La pièce est-elle équilibrée ou pas ?" s'écrit $H_0 : \theta = 1/2$ (i.e. $\Theta_0 = \{1/2\}$) contre $H_1 : \theta \neq 1/2$ (i.e. $\Theta_1 = \Theta_0^c$).

Définition 15. *Dans le cadre du problème de test de H_0 contre H_1, un test est une statistique T à valeurs dans $\{0,1\}$ associée à la stratégie suivante : pour l'observation $\mathbb{X} = (X_1, \ldots, X_n)$, H_0 est conservée (respectivement rejetée) si $T(\mathbb{X}) = 0$ (respectivement $T(\mathbb{X}) = 1$). La région de rejet du test est $T^{-1}(\{1\})$, i.e. l'ensemble des $x \in \mathscr{H}^n$ tels que $T(x) = 1$.*

Un test peut donc toujours s'écrire $T(\mathbb{X}) = \mathbb{1}_{[\mathbb{X} \in R]}$, où R est la région de rejet. Il est parfois plus naturel de l'écrire sous la forme $T(\mathbb{X}) = \mathbb{1}_{[h(\mathbb{X}) \in R']}$, où h est une fonction mesurable appelée statistique de test. On dit aussi souvent, en commettant un abus de dénomination, que R' est la région de rejet associée à la statistique de test $h(\mathbb{X})$.

En pratique, puisqu'on ne dispose que d'une réalisation x de $\mathbb{X}$, le protocole est alors le suivant : si x tombe dans la région de rejet R (ou si $h(x) \in R'$), on conserve H_0, sinon on la rejette.

4.2 Risques d'un test

Un test est construit à partir d'une probabilité d'erreur, ou risque. Le premier type de risque que l'on peut dégager est la probabilité de rejeter H_0 à tort.

Définition 16. *Le risque de première espèce du test T est l'application qui à chaque élément $\theta \in \Theta$ donne la probabilité de prendre la mauvaise décision :*

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

La *taille* du test est le réel $\alpha^\star$ défini par

$$\alpha^\star = \sup_{\theta \in \Theta_0} \underline{\alpha}(\theta).$$

On dit que le test T est de *niveau* $\alpha \in]0,1[$ si sa taille est inférieure ou égale à α.

Pour un test de niveau suffisamment proche de 0, si la décision de rejeter l'hypothèse nulle est convaincante, la décision de la conserver est, en

revanche, plus contestable. Par exemple, le test nul $T \equiv 0$, pour lequel l'hypothèse nulle est toujours choisie, possède un niveau nul. Pour autant, il n'apporte aucune information car la décision rendue est toujours la même. Ce phénomène nous conduit à distinguer un autre type de risque, en l'occurrence la probabilité de conserver H_0 à tort.

Définition 17. *Le risque de seconde espèce du test T est l'application qui à chaque élément $\theta \in \Theta_1$ donne la probabilité de prendre la mauvaise décision :*

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

A partir de là, on définit la puissance du test comme la fonction $1 - \underline{\beta}$, c'est-à-dire l'application qui à chaque élément de Θ_1 associe la probabilité de prendre la bonne décision.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. On souhaite tester

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

ou, de façon équivalente, en définissant $\Theta_0 = [1, +\infty[$ et $\Theta_1 =]-\infty, 1[$,

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

L'intuition conduit à prendre $T(\mathbb{X}) = \mathbb{1}_{[\bar{X}_n < 1]}$: ce test est fondé sur la statistique de test $\bar{X}_n$ et la région de rejet $]-\infty, 1[$. En notant Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$, on voit facilement que, $\forall \theta \geq 1$,

$$\underline{\alpha}(\theta) = \Phi\left(\sqrt{n}(1-\theta)\right)$$

et, $\forall \theta < 1$,

$$\underline{\beta}(\theta) = 1 - \Phi\left(\sqrt{n}(1-\theta)\right).$$

Par conséquent, la taille du test est $\Phi(0) = 1/2$, ce qui n'est pas très convaincant...

4.3 Dissymétrie des rôles de H_0 et H_1

L'exemple précédent montre qu'il est impossible de contrôler l'une et l'autre des hypothèses (c'est intuitivement clair, penser au cas dégénéré $T \equiv 0$).

On choisit donc de dissymétriser le problème, en faisant jouer un rôle particulier à H_0 et en maîtrisant exclusivement le risque de première espèce. Ainsi, on prendra pour H_0 :

- ▷ Une hypothèse communément admise ;
- ▷ Une hypothèse de prudence, conservative ;
- ▷ Une hypothèse facile à formuler...

... en imposant que le risque de première espèce soit inférieur à $\alpha \in]0,1[$, où α (le niveau) est **fixé à l'avance par le statisticien**. Reprenons l'exemple du test gaussien à la lumière de ce nouveau paradigme.

Exemple. $\mathbb{X} = (X_1, \dots, X_n)$ i.i.d., de loi commune $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. On souhaite tester

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

La statistique de test naturelle est toujours $\bar{X}_n$ et l'on cherche, au vu de H_1, une région de rejet de la forme $]-\infty, k_\alpha[$, soit $T(\mathbb{X}) = \mathbb{1}_{[\bar{X}_n < k_\alpha]}$. Comment déterminer k_α ? En écrivant simplement la contrainte de niveau :

$$\sup_{\theta \geq 1} \mathbb{P}(\bar{X}_n < k_\alpha) \leq \alpha.$$

Or, si N est une variable aléatoire de loi $\mathcal{N}(0,1)$, on a, $\forall \theta \geq 1$,

$$\begin{aligned} \mathbb{P}(\bar{X}_n < k_\alpha) &= \mathbb{P}\left(\sqrt{n}(\bar{X}_n - \theta) < \sqrt{n}(k_\alpha - \theta)\right) \\ &= \mathbb{P}\left(N < \sqrt{n}(k_\alpha - \theta)\right). \end{aligned}$$

Dès lors,

$$\sup_{\theta \geq 1} \mathbb{P}(\bar{X}_n < k_\alpha) = \Phi\left(\sqrt{n}(k_\alpha - 1)\right).$$

Il suffit donc de choisir k_α tel que $\Phi(\sqrt{n}(k_\alpha - 1)) = \alpha$, soit

$$k_\alpha = 1 + \frac{q_\alpha}{\sqrt{n}},$$

où q_α désigne le quantile d'ordre α de la loi $\mathcal{N}(0,1)$. En conclusion, on rejette l'hypothèse H_0 si $\bar{X}_n < 1 + \frac{q_\alpha}{\sqrt{n}}$.

Par ailleurs, pour $\theta < 1$, on a $1 - \underline{\beta}(\theta) = \mathbb{P}(T(\mathbb{X}) = 1)$. Ainsi,

$$\begin{aligned} 1 - \underline{\beta}(\theta) &= \mathbb{P}\Big(\bar{X}_n < 1 + \frac{q_\alpha}{\sqrt{n}}\Big) \\ &= \mathbb{P}\left(\sqrt{n}(\bar{X}_n - \theta) < \sqrt{n}(1-\theta) + q_\alpha\right) \\ &= \mathbb{P}\left(N < \sqrt{n}(1-\theta) + q_\alpha\right), \end{aligned}$$

la dernière égalité provenant du fait que, pour $\theta < 1$, $\bar{X}_n \sim \mathcal{N}(\theta, \frac{1}{\sqrt{n}})$ (et non pas $\mathcal{N}(1, \frac{1}{\sqrt{n}})$: erreur classique !). La fonction puissance est donc croissante, minorée par α et tend vers 1 lorsque θ tend vers l'infini.

4.4 Propriétés éventuelles d'un test

Lorsque la fonction puissance du test T passe strictement en dessous de sa taille, la stratégie de décision qui en résulte devient incohérente. En effet, il existe alors $\theta_1 \in \Theta_1$ et $\theta_0 \in \Theta_0$ tels que $\mathbb{E}_{\theta_1} T < \mathbb{E}_{\theta_0} T$. On se retrouve alors dans la situation paradoxale où la probabilité de rejeter H_0 à raison est plus petite que la probabilité de rejeter H_0 à tort ! Dans un tel contexte, le test ne sépare pas bien les hypothèses H_0 et H_1, d'où la nécessité de définir un concept délimitant cette situation. C'est l'objet de la définition qui suit.

Définition 18. *Un test T de niveau $\alpha \in]0,1[$ est dit sans biais si sa puissance est supérieure à α, i.e.*

$$\mathbb{E}_\theta T \geq \alpha \ \forall \theta \in \Theta_1.$$

Rien ne nous certifie, en général, qu'un test sans biais existe. Nous reviendrons sur ce problème ultérieurement.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de densité commune sur $\mathbb{R}$

$$f_\theta(x) = e^{-(x-\theta)} \mathbb{1}_{[\theta, +\infty[}(x),$$

où θ est un paramètre réel. Fixons $\alpha \in]0,1[$. Dans le cadre du problème de test de $H_0 : \theta \leq 0$ contre $H_1 : \theta > 0$, nous allons montrer que le test $T = \mathbb{1}_{[\mathbb{X} \in R]}$ associé à la région de rejet

$$R = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \min_{1 \leq i \leq n} x_i \geq -\frac{\ln \alpha}{n} \right\}$$

est de niveau α et sans biais. En effet, comme $X_1, \ldots, X_n$ sont indépendantes et de même loi, on a lorsque $\theta \leq 0$:

$$\begin{aligned} \mathbb{E}_\theta T &= \mathbb{P}\Big(\min_{1 \leq i \leq n} X_i \geq -\frac{\ln \alpha}{n} \Big) = \mathbb{P}\Big(X_1 \geq -\frac{\ln \alpha}{n} \Big)^n \\ &= \left\{ \int_{-\ln \alpha / n}^{+\infty} e^{-(t-\theta)} \mathrm{d}t \right\}^n = \alpha e^{n\theta}. \end{aligned}$$

Ainsi, $\mathbb{E}_\theta T \leq \alpha$ (avec égalité si $\theta = 0$), i.e. le test T est de niveau α. Calculons maintenant sa puissance : pour $\theta \geq -\ln\alpha/n$, $\mathbb{E}_\theta T = 1$ car toutes les variables aléatoires X_i sont $\mathbb{P}$-p.s. plus grandes que $\theta \geq -\ln\alpha/n$; pour $\theta \in]0, -\ln\alpha/n[$, $\mathbb{E}_\theta T = \alpha e^{n\theta} \geq \alpha$. Le test T est donc sans biais.

Définition 19. *Un test T de niveau $\alpha \in]0,1[$ est dit uniformément plus puissant parmi tous les tests de niveau α (UPPα) si, pour tout autre test T' de niveau α, on a*

$$\mathbb{E}_\theta T \geq \mathbb{E}_\theta T' \ \forall \theta \in \Theta_1.$$

La notion d'optimalité envisagée est claire, un test UPP étant de puissance maximale pour un niveau fixé. La question délicate de la caractérisation des tests UPP fait l'objet de la section qui suit.

4.5 Tests de Neyman-Pearson

Dans cette section, $\mathscr{H} \subset \mathbb{R}^d$ et $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est un modèle statistique dominé par une mesure σ-finie ν et de vraisemblance L_n. Fixons deux paramètres $\theta_0 \neq \theta_1 \in \Theta$ et considérons le problème de test simple suivant :

$$H_0 : \theta = \theta_0 \quad \text{contre} \quad H_1 : \theta = \theta_1.$$

L'objectif de cette section est de donner, dans ce cas simple, des conditions suffisantes pour qu'un test soit UPP. Le bon contexte est ici celui des tests de *rapport de vraisemblance*.

Définition 20. *Un test T est dit de Neyman-Pearson (ou de rapport de vraisemblance) s'il existe $k \in \mathbb{R}_+$ tel que*

$$T = \mathbb{1}_{[L_n(\mathbb{X};\theta_1) > kL_n(\mathbb{X};\theta_0)]}.$$

On note $T = T_k$ et on désigne par $\mathscr{T}$ l'ensemble des tests de Neyman-Pearson.

Un test de Neyman-Pearson T_k s'interprète comme suit : si, par exemple, l'observation réalisée $x \in \mathscr{H}^n$ est issue de la loi P_{θ_0}, la vraisemblance de x en θ_0 est en principe plus grande que la vraisemblance de x en θ_1, ce qui est exprimé par la propriété $L_n(x;\theta_1) \leq kL_n(x;\theta_0)$ pour un certain $k \in \mathbb{R}_+$. Bien sûr, dans ce cas, $T(x) = 0$, i.e. H_0 n'est pas rejetée.

Le résultat qui suit, appelé lemme de Neyman-Pearson, est relatif à l'existence d'un test UPP dans la famille des tests de Neyman-Pearson.

Théorème 3. *Soit $\alpha \in]0,1[$. Si un test de $\mathscr{T}$ est de taille α, alors il est UPPα.*

Démonstration. Soient T un test de niveau α et $T_k^\star \in \mathscr{T}$ un test de taille α avec $k \in \mathbb{R}_+$. Alors, pour tout $x \in \mathscr{H}^n$,

$$\begin{cases} T^\star(x) - T(x) > 0 \Rightarrow T^\star(x) = 1 \Rightarrow L_n(x;\theta_1) > kL_n(x;\theta_0); \\ T^\star(x) - T(x) < 0 \Rightarrow T^\star(x) = 0 \Rightarrow L_n(x;\theta_1) \leq kL_n(x;\theta_0). \end{cases}$$

Il s'ensuit,

$$(T^\star(x) - T(x))\, L_n(x;\theta_1) \geq k\,(T^\star(x) - T(x))\, L_n(x;\theta_0),$$

et, par conséquent,

$$\mathbb{E}_{\theta_1} T^\star - \mathbb{E}_{\theta_1} T = \int_{\mathscr{H}^n} (T^\star - T) L_n(\cdot;\theta_1) \mathrm{d}\nu \geq k \int_{\mathscr{H}^n} (T^\star - T) L_n(\cdot;\theta_0) \mathrm{d}\nu.$$

On a donc $\mathbb{E}_{\theta_1} T^\star - \mathbb{E}_{\theta_1} T \geq k(\mathbb{E}_{\theta_0} T^\star - \mathbb{E}_{\theta_0} T)$. Or, $\mathbb{E}_{\theta_0} T^\star \geq \mathbb{E}_{\theta_0} T$ car $T^\star$ est de taille α et T est de niveau α, ce qui montre que

$$\mathbb{E}_{\theta_1} T^\star \geq \mathbb{E}_{\theta_1} T,$$

i.e. $T^\star$ est UPPα. □

Il existe de nombreuses extensions et raffinements divers du lemme de Neyman-Pearson, notamment en termes de forme des hypothèses. Nous renvoyons le lecteur à l'ouvrage de Cadre et Vial (2012) pour un traitement plus en profondeur du problème.

4.6 Tests asymptotiques

A défaut d'informations suffisantes ou appropriées sur la loi de la statistique de test, on est amené, à l'instar des intervalles de confiance asymptotiques, à définir la notion de *test asymptotique*.

Définition 21. *Un test asymptotique T_n de niveau $\alpha \in]0,1[$ est un test qui vérifie*

$$\sup_{\theta \in \Theta_0} \lim_{n \to +\infty} \mathbb{E}_\theta T_n \leq \alpha.$$

La procédure de décision est calquée sur celle des tests à taille d'échantillon finie. La seule différence est qu'un test asymptotique est construit pour contrôler le risque de première espèce, mais seulement asymptotiquement. Dans ce contexte, il est raisonnable d'exiger une puissance asymptotique maximale. C'est le concept de *convergence* décrit ci-dessous.

Définition 22. *Un test asymptotique T_n est dit convergent si*

$$\lim_{n\to+\infty} \mathbb{E}_\theta T_n = 1 \ \forall \theta \in \Theta_1.$$

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. On souhaite confronter les hypothèses

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

Pour ce faire, nous allons construire un test asymptotique de niveau 5%. Si $\theta = 1/2$, d'après le théorème central limite :

$$2\sqrt{n}\,(\bar{X}_n - 0.5) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

En notant $q_{0.975}$ le quantile d'ordre 0.975 de la loi $\mathcal{N}(0,1)$, on en déduit que

$$\begin{aligned}\lim_{n\to+\infty} \mathbb{P}\left(2\sqrt{n}\,|\bar{X}_n - 0.5| > q_{0.975}\right) &= 1 - \Phi(q) + \Phi(-q) \\ &= 2\,(1 - \Phi(q)) = 0.05,\end{aligned}$$

si Φ désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$. Par suite, le test $T_n = \mathbb{1}_{[\bar{X}_n \in R]}$ défini par

$$R = \{x \in \mathbb{R} : 2\sqrt{n}|x - 0.5| > q_{0.975}\}$$

est un test asymptotique de niveau 5%.

Ce test est de plus convergent car, si $\theta \neq 1/2$,

$$\begin{aligned}\lim_{n\to+\infty} \mathbb{E}_\theta T_n &= \lim_{n\to+\infty} \mathbb{P}(\bar{X}_n \in R) \\ &= \lim_{n\to+\infty} \mathbb{P}\left(2\left|\sqrt{n}(\bar{X}_n - \theta) + \sqrt{n}(\theta - 0.5)\right| > q_{0.975}\right) \\ &= 1,\end{aligned}$$

puisque $\sqrt{n}(\theta - 0.5)$ tend vers l'infini.

Application numérique : Supposons que $n = 1000$. Comme $q_{0.975} = 1.960$, on rejette H_0 dès lors que $|\bar{X}_n - 0.5| > 0.03$. Si, par exemple, $\bar{x}_n = 0.52$, on conserve H_0 au niveau (asymptotique) 5%.

Chapitre 5

Echantillons gaussiens et modèle linéaire

5.1 Rappels sur les vecteurs gaussiens

Cas réel. Une variable aléatoire réelle X est dite *gaussienne* (ou de *loi normale*) de paramètres $m \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}_+$ ($\sigma \geq 0$) si sa fonction caractéristique s'écrit

$$\mathbb{E}\exp(iuX) = \exp\left(ium - \frac{\sigma^2 u^2}{2}\right) \ \forall u \in \mathbb{R}.$$

La loi de X est notée $\mathcal{N}(m, \sigma^2)$, et l'on a $\mathbb{E}X = m$ et $\mathbb{V}X = \sigma^2$. Lorsque $\sigma = 0$, on dit que X est dégénérée ; dans le cas contraire, elle admet la densité par rapport à la mesure de Lebesgue

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

Cas vectoriel. Plus généralement, une variable aléatoire X à valeurs dans $\mathbb{R}^d$ est un *vecteur gaussien* de $\mathbb{R}^d$ s'il existe $M \in \mathbb{R}^d$ et Σ une matrice $d \times d$ réelle, symétrique et positive, tels que la fonction caractéristique de X s'écrive

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

(Les vecteurs sont considérés comme des matrices colonnes.) La loi de X est notée $\mathcal{N}_d(M, \Sigma)$. Alors M est la moyenne de X, i.e. $\mathbb{E}X = M$, et Σ est la

matrice de variance-covariance de X, i.e.

$$\Sigma = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top.$$

Lorsque la matrice Σ est inversible (i.e., définie positive), X admet la densité par rapport à la mesure de Lebesgue

$$\frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma}} \exp\Big(-\frac{1}{2}(x-M)^\top\Sigma^{-1}(x-M)\Big)\ \forall x \in \mathbb{R}^d.$$

Lorsque Σ n'est pas inversible, on montre facilement que la loi de X est $\mathbb{P}$-p.s. concentrée sur le sous-espace affine de $\mathbb{R}^d$ d'origine M et engendré par les vecteurs propres correspondant aux valeurs propres non nulles de Σ.

Un moyen simple de montrer qu'un vecteur aléatoire est gaussien est d'utiliser la définition équivalente (exercice) suivante :

Définition 23. *Un vecteur aléatoire est gaussien si et seulement si toute combinaison linéaire de ses composantes est une variable aléatoire réelle gaussienne.*

Mentionnons pour finir deux résultats d'utilité constante dans la manipulation des vecteurs gaussiens, dont la preuve est laissée en exercice.

Proposition 2.

(**i**) **Transformation affine.** *Si A est une matrice réelle de format $k \times d$, $b \in \mathbb{R}^k$ et X est un vecteur de loi $\mathcal{N}_d(M, \Sigma)$, alors $AX + b$ suit la loi $\mathcal{N}_k(AM + b, A\Sigma A^\top)$.*

(**ii**) **Caractérisation de l'indépendance.** *Soit X un vecteur gaussien. Les composantes de X sont des variables aléatoires réelles indépendantes si et seulement si la matrice de variance-covariance de X est diagonale.*

Ainsi, lorsque $X = (X_1, \ldots, X_d)$ est un vecteur gaussien, on a l'équivalence :

$$\forall i \neq j,\ X_i \text{ et } X_j \text{ indépendantes} \Leftrightarrow \mathrm{Cov}(X_i, X_j) = 0.$$

Rappelons que seule l'implication $\Rightarrow$ est vraie dans le cas général.

Exemple. Soient $Z \sim \mathcal{N}(0,1)$ et $\varepsilon \sim \mathscr{B}(1/2)$ deux variables aléatoires indépendantes. Alors $X_1 = Z$ et $X_2 = (2\varepsilon - 1)Z$ sont des variables aléatoires réelles gaussiennes (pourquoi ?), mais $X = (X_1, X_2)$ n'est pas un vecteur gaussien, puisque $X_1 + X_2 = 2\varepsilon Z$ prend avec probabilité $1/2$ la valeur 0. On notera que $\mathrm{Cov}(X_1, X_2) = 0$ mais que X_1 et X_2 ne sont pas indépendantes (sans quoi X serait un vecteur gaussien...)

5.2 Théorème de Cochran

Dans le monde des vecteurs gaussiens, orthogonalité et indépendance se confondent. Ce lien entre la géométrie et les probabilités a pour conséquence le théorème ci-dessous, qui constitue la pierre angulaire de toute la statistique des échantillons gaussiens.

Théorème 4 (Cochran). *Soient $\sigma > 0$, $X \sim \mathcal{N}_n(0, \sigma^2 \mathrm{Id})$ et $V_1, \dots, V_p$ des sous-espaces vectoriels orthogonaux de dimensions respectives $r_1, \dots, r_p$ tels que*

$$V_1 \oplus \cdots \oplus V_p = \mathbb{R}^n.$$

Alors les projections orthogonales $\pi_1, \dots, \pi_p$ de X sur $V_1, \dots, V_p$ sont des vecteurs gaussiens indépendants et, pour chaque $i = 1, \dots, p$,

$$\frac{1}{\sigma^2} \|\pi_i\|^2 \sim \chi^2(r_i).$$

Démonstration. Soit $(e_j^i)_{i,j}$ une base orthonormée de $\mathbb{R}^n$ telle que pour chaque $i = 1, \dots, p$, $(e_j^i)_{1 \le j \le r_i}$ est une base orthonormée de V_i. Si $i = 1, \dots, p$, $\pi_i = M_i X$, où M_i est la matrice symétrique de format $n \times n$ définie par

$$M_i = (e_1^i \cdots e_{r_i}^i) \, (e_1^i \cdots e_{r_i}^i)^\top.$$

Noter que puisque les vecteurs $(e_j^i)_{i,j}$ sont normés et orthogonaux, M_i est idempotente et de plus $M_i M_j = 0$ pour tout $i \neq j$.

Montrons la première assertion du théorème. Tout d'abord, X étant gaussien, toute combinaison linéaire de ses composantes est gaussienne, donc $(\pi_1, \dots, \pi_p)$ est gaussien. De plus, la covariance entre les vecteurs aléatoires π_i et π_j est nulle pour tout $i \neq j$. En effet, ces vecteurs aléatoires étant centrés,

$$\mathbb{C}(\pi_i, \pi_j) = \mathbb{E} \, (\pi_i - \mathbb{E}\pi_i) \, \big(\pi_j - \mathbb{E}\pi_j\big)^\top = \mathbb{E}\pi_i \pi_j^\top,$$

$\mathbb{C}$ et $\mathbb{E}$ désignant respectivement la matrice de covariance et l'espérance sous la probabilité $\mathbb{P}$. Il vient,

$$\begin{aligned} \mathbb{C}(\pi_i, \pi_j) &= \mathbb{E} M_i X (M_j X)^\top \\ &= M_i \mathbb{E} X X^\top M_j \\ &= \sigma^2 M_i M_j, \end{aligned}$$

d'où $\mathbb{C}(\pi_i, \pi_j) = 0$. Par suite, $\pi_1, \ldots, \pi_p$ sont des vecteurs gaussiens indépendants.

Pour montrer la seconde assertion, fixons $i = 1, \ldots, p$ et remarquons que comme M_i est symétrique et idempotente :

$$\pi_i \sim \mathcal{N}_n\big(0, \sigma^2 M_i \,\mathrm{Id}\, M_i\big) = \mathcal{N}_n(0, \sigma^2 M_i).$$

En notant E_i la matrice de format $n \times r_i$ définie par $E_i = (e_1^i \cdots e_{r_i}^i)$, on a donc

$$\pi_i \sim \sigma E_i \mathcal{N}_{r_i}(0, \mathrm{Id}).$$

Or, si Z est un vecteur aléatoire de loi $\mathcal{N}_{r_i}(0, \mathrm{Id})$, $\|E_i Z\|^2 = \|Z\|^2 \sim \chi^2_{r_i}$ car $E_i^\top E_i = \mathrm{Id}$, d'où le théorème. □

5.3 Echantillons gaussiens

Rappelons que pour une suite $X_1, \ldots, X_n$ de variables aléatoires réelles, on note

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

Le théorème ci-dessous met en évidence le rôle tenu par la loi de Student et la loi du χ^2 lorsque $X_1, \ldots, X_n$ sont indépendantes et de même loi gaussienne.

Théorème 5. *Soient $m \in \mathbb{R}$, $\sigma > 0$ et $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. Alors :*

(**i**) *$\bar{X}_n$ et S_n^2 sont indépendantes.*

(**ii**) $(n-1)S_n^{\star 2}/\sigma^2 \sim \chi^2(n-1)$.

(**iii**) $\sqrt{n}(\bar{X}_n - m)/S_n^\star \sim \mathcal{T}(n-1)$.

Remarque. Dans ce théorème, (**iii**) est à comparer à la propriété classique $\sqrt{n}(\bar{X}_n - m)/\sigma \sim \mathcal{N}(0,1)$ satisfaite par la suite de variables aléatoires indépendantes $X_1, \ldots, X_n$ de même loi $\mathcal{N}(m, \sigma^2)$.

Démonstration. Pour simplifier les écritures, considérons le cas $m = 0$ et $\sigma = 1$. Soit V le sous-espace vectoriel de $\mathbb{R}^n$ engendré par $e = (1 \cdots 1)^\top$ et $X = (X_1 \dots X_n)^\top$. Le projecteur orthogonal P sur V est la matrice $n \times n$ dont tous les coefficients valent $1/n$. De ce fait,

$$PX = \bar{X}_n e \quad \text{et} \quad (\mathrm{Id} - P)X = \begin{pmatrix} X_1 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{pmatrix}.$$

Comme $(\mathrm{Id} - P)X$ est la projection orthogonale de X sur l'orthogonal de V et X suit la loi $\mathcal{N}_n(0, \mathrm{Id})$, on déduit du théorème de Cochran que PX est indépendant de $(\mathrm{Id} - P)X$, et donc en particulier que $\bar{X}_n$ est indépendant de $S_n^{\star 2}$, d'où (**i**). De plus, comme V est de dimension 1,

$$(n-1)S_n^{\star 2} = \|(\mathrm{Id} - P)X\|^2 \sim \chi^2(n-1)$$

d'après le théorème de Cochran, d'où (**ii**). Enfin, (**iii**) se déduit des résultats précédents, car $\sqrt{n}\bar{X}_n$ et $(n-1)S_n^{\star 2}$ sont indépendantes, et de lois respectives $\mathcal{N}(0,1)$ et $\chi^2(n-1)$. □

Le Théorème 5 a des conséquences simples mais fondamentales pour le traitement des échantillons gaussiens i.i.d. Nous détaillons dans les paragraphes qui suivent deux exemples (le test de Student et le test de Fisher), mais bien d'autres extensions sont possibles. A partir de maintenant, on considère un n-échantillon $\mathbb{X} = (X_1, \dots, X_n)$ i.i.d., de loi commune $\mathcal{N}(m, \sigma^2)$, avec $m \in \mathbb{R}$ et $\sigma > 0$.

Test de Student. Construisons un test de niveau $\alpha \in]0,1[$ dans le problème de test de

$$H_0 : m \geq m_1 \quad \text{contre} \quad H_1 : m < m_1,$$

avec m_1 un réel fixé. Un protocole naturel de rejet pour ce problème est de la forme $\bar{X}_n < k_\alpha$, avec k_α un seuil à préciser, car H_0 est rejetée lorsque la moyenne des observations prend une valeur anormalement faible. Cependant, ce test n'est pas utilisable car la loi de la statistique $\bar{X}_n$, en l'occurrence $\mathcal{N}(m, \sigma^2/n)$, fait intervenir les paramètres inconnus m et σ^2.

Adaptons la construction du test à cette contrainte : si $t_\alpha^{(n-1)}$ est le quantile d'ordre α de la loi $\mathcal{T}(n-1)$ alors, sous H_0 (i.e. $m \geq m_1$),

$$\mathbb{P}\Big(\bar{X}_n < m_1 + t_\alpha^{(n-1)} \frac{S_n^\star}{\sqrt{n}}\Big) \leq \mathbb{P}\Big(\bar{X}_n < m + t_\alpha^{(n-1)} \frac{S_n^\star}{\sqrt{n}}\Big),$$

et donc d'après le Théorème 5,

$$\mathbb{P}\Big(\bar{X}_n < m_1 + t_\alpha^{(n-1)} \frac{S_n^\star}{\sqrt{n}}\Big) \leq \mathbb{P}\Big(\sqrt{n}\,\frac{\bar{X}_n - m}{S_n^\star} < t_\alpha^{(n-1)}\Big) = \alpha,$$

avec égalité lorsque $m = m_1$. Ainsi, le test de région de rejet

$$R_{\text{Student}} = \left\{(x_1, \dots, x_n) \in \mathbb{R}^n : \bar{x}_n < m_1 + t_\alpha^{(n-1)} \frac{s_n^\star}{\sqrt{n}}\right\},$$

appelé *test de Student*, est de niveau (de taille) α. La procédure de décision consiste donc à rejeter H_0 au niveau α lorsque $(X_1, \dots, X_n) \in R_{\text{Student}}$.

Test de Fisher. Construisons un test de niveau $\alpha \in]0,1[$ dans le problème de test de

$$H_0 : \sigma \geq \sigma_1 \quad \text{contre} \quad H_1 : \sigma < \sigma_1,$$

avec $\sigma_1 > 0$ fixé. Une région de rejet naturelle pour ce problème de test est de la forme $s_n^\star < k_\alpha$ avec k_α un seuil à préciser, car H_0 est rejetée lorsque la variance empirique prend une valeur anormalement faible. Soit $\chi^2_\alpha(n-1)$ le quantile d'ordre α de la loi $\chi^2(n-1)$. Sous H_0 (i.e. $\sigma \geq \sigma_1$),

$$\mathbb{P}\Big(S_n^{\star 2} < \frac{\chi^2_\alpha(n-1)}{n-1}\,\sigma_1^2\Big) \leq \mathbb{P}\Big(S_n^{\star 2} < \frac{\chi^2_\alpha(n-1)}{n-1}\,\sigma^2\Big).$$

D'après le Théorème 5,

$$\mathbb{P}\Big(S_n^{\star 2} < \frac{\chi^2_\alpha(n-1)}{n-1}\,\sigma_1^2\Big) \leq \mathbb{P}\Big(\frac{(n-1)S_n^{\star 2}}{\sigma^2} < \chi^2_\alpha(n-1)\Big) = \alpha,$$

avec égalité lorsque $\sigma = \sigma_1$. Le *test de Fisher* est le test de région de rejet

$$R_{\text{Fisher}} = \left\{(x_1, \dots, x_n) \in \mathbb{R}^n : s_n^{\star 2} < \frac{\chi^2_\alpha(n-1)}{n-1}\,\sigma_1^2\right\}.$$

Ce test est de niveau (de taille) α, et la procédure de décision consiste à rejeter H_0 au niveau α lorsque $(X_1, \dots, X_n) \in R_{\text{Fisher}}$.

5.4 Régression linéaire multiple

Modèle statistique. De manière générale, il s'agit de modéliser une expérience dont chaque observation est influencée par des régresseurs (déterministes), représentés par k vecteurs de $\mathbb{R}^n$ connus notés $R_1, \dots, R_k$. On impose l'hypothèse d'homoscédasticité du modèle selon laquelle la matrice de variance-covariance de la loi dont l'observation est issue est proportionnelle à la matrice identité. En désignant par $R = (R_1 \dots R_k)$ la matrice des régresseurs de format $n \times k$, le modèle statistique admet la formulation suivante :

$$X = R\theta + \varepsilon,$$

avec $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathrm{Id})$, pour des paramètres inconnus $\theta \in \mathbb{R}^k$ et $\sigma > 0$. Par ailleurs, il admet la formulation équivalente suivante :

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

En réduisant au besoin leur nombre, on peut toujours considérer que les régresseurs sont linéairement indépendants et que, par conséquent, la matrice des régresseurs R est de rang k.

Estimation des paramètres. Dans ce qui suit, E désigne l'espace vectoriel engendré par les vecteurs $R_1, \dots, R_k$. Prenons la projection orthogonale X_E [1] de X sur E ; elle s'écrit $X_E = R\hat{\theta}$ avec $\hat{\theta}$ un estimateur de θ. On peut décrire explicitement $\hat{\theta}$ en remarquant que, comme $X - R\hat{\theta}$ est dans l'orthogonal de E, pour tout $u \in \mathbb{R}^k$:

$$\langle Ru, X - R\hat{\theta} \rangle = 0.$$

Par suite, $\langle u, R^\top X - R^\top R\hat{\theta} \rangle = 0$ pour tout $u \in \mathbb{R}^k$ et donc $R^\top X = R^\top R\hat{\theta}$. Puisque R est de rang plein, la matrice $R^\top R$ est inversible (pourquoi ?), d'où

$$\hat{\theta} = (R^\top R)^{-1} R^\top X.$$

L'estimateur $\hat{\theta}$ est sans biais car, si $\mathbb{E}_{\theta,\sigma}$ désigne l'espérance sous la loi de X :

$$\mathbb{E}_{\theta,\sigma}\hat{\theta} = (R^\top R)^{-1} R^\top \mathbb{E}_{\theta,\sigma} X = (R^\top R)^{-1} R^\top R\theta = \theta.$$

1. u_F désigne la projection orthogonale de $u \in \mathbb{R}^n$ sur le sous-espace vectoriel F de $\mathbb{R}^n$.

Construisons maintenant un estimateur de σ^2. Comme $X_E = R\theta + \varepsilon_E$, $X - X_E = \varepsilon - \varepsilon_E$, d'où $\|X - X_E\|^2 \sim \sigma^2\chi^2(n-k)$ d'après le théorème de Cochran. La moyenne de la loi $\chi^2(n-k)$ valant $n-k$, l'estimateur

$$\hat{\sigma}^2 = \frac{\|X - X_E\|^2}{n-k}$$

de σ^2 est donc sans biais.

Test de l'utilité des régresseurs. Dans le cadre d'une modélisation trop complète, tous les régresseurs n'ont pas la même influence, et certains n'ont qu'une contribution mineure. Nous allons construire un test dans le but de supprimer ces régresseurs à l'influence réduite.

Fixons $q = 0, \ldots, k-1$. S'interroger sur l'utilité des $(k-q)$ derniers régresseurs mène au problème de test suivant :

$$H_0 : \forall i = q+1, \ldots, k,\ \theta_i = 0 \quad \text{contre} \quad H_1 : \exists i = q+1, \ldots, k,\ \theta_i \neq 0.$$

Sous H_0, la matrice des régresseurs utiles $\bar{R} = (R_1 \ldots R_q)$ est la restriction de R à ses q premiers régresseurs. L'effet moyen $R\theta$ se trouve alors dans l'espace vectoriel V engendré par $R_1, \ldots, R_q$, dont la dimension est q car $R_1, \ldots, R_q$ sont linéairement indépendants par hypothèse. Avec ces notations, le problème de test se réécrit de la manière suivante :

$$H_0 : R\theta \in V \quad \text{contre} \quad H_1 : R\theta \in E \setminus V.$$

Le principe de construction d'un tel test est de rejeter H_0 lorsque les projections orthogonales de l'observation sur E et sur V sont significativement différentes. Selon ce principe, une région de rejet naturelle est de la forme $\{x \in \mathbb{R}^n : \|x_E - x_V\| \geq s\}$ avec s un seuil à préciser. Mais la loi de $\|X_E - X_V\|$ dépend du paramètre inconnu σ. En effet, sous H_0, $X_V = R\theta + \varepsilon_V$ car $R\theta \in V$ et donc, d'après le théorème de Cochran appliqué au vecteur gaussien ε,

$$\|X_E - X_V\|^2 = \|\varepsilon_E - \varepsilon_V\|^2 \sim \sigma^2\chi^2(k-q).$$

Or, le théorème de Cochran montre aussi que sous H_0, le vecteur aléatoire $\varepsilon_E - \varepsilon_V = X_E - X_V$ est indépendant de $\varepsilon - \varepsilon_E = X - X_E$. Enfin, $\|X -$

$X_E\|^2 \sim \sigma^2\chi^2(n-k)$. En réunissant ces observations, et en notant pour $u \in \mathbb{R}^n$:

$$F(x) = \frac{\|x_E - x_V\|^2/(k-q)}{\|x - x_E\|^2/(n-k)},$$

on trouve $F(\mathbb{X}) \sim \mathscr{F}(k-q, n-k)$ sous H_0. Si $f_{1-\alpha}^{(k-q,n-k)}$ désigne le quantile d'ordre $(1-\alpha)$ de la loi de Fisher $\mathscr{F}(k-q, n-k)$ alors, sous H_0,

$$\mathbb{P}\left(F(\mathbb{X}) \geq f_{1-\alpha}^{(k-q,n-k)}\right) = \alpha.$$

La région de rejet

$$R_{\text{regress}} = \left\{x \in \mathbb{R}^n : F(x) \geq f_{1-\alpha}^{(k-q,n-k)}\right\}$$

nous donne donc un test de niveau (de taille) α pour le problème de test de H_0 contre H_1. La procédure de décision consiste à rejeter H_0 au niveau α si l'observation $\mathbb{X} = (X_1, \ldots, X_n)$ tombe dans R_{regress}.

Chapitre 6

Information et exhaustivité

6.1 Information de Fisher

Dans la suite, $\nabla F(\theta)$ désigne le gradient de $F : \Theta \to \mathbb{R}$ évalué en $\theta \in \Theta$. Par convention, le gradient d'une fonction n'est calculé en θ que si la fonction est de classe $\mathscr{C}^1$ sur un voisinage de θ. Par ailleurs, Cov_θ (respectivement $\mathbb{V}_\theta$) désigne la covariance (respectivement la matrice de variance-covariance) sous la loi P_θ.

On suppose dans tout le chapitre que $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est un modèle statistique dominé par une mesure σ-finie ν, avec $\mathscr{H} \subset \mathbb{R}^d$ et Θ un ouvert de $\mathbb{R}^k$. On note L_n la vraisemblance du modèle et on désigne comme d'habitude par $\mathbb{X} = (X_1, \ldots, X_n)$ une observation générique issue de P_θ.

Définition 24. *Supposons que $\nabla \ln L_n(\mathbb{X};\theta) \in \mathbb{L}^2$ pour chaque $\theta \in \Theta$. L'information de Fisher est la fonction I_n définie sur Θ telle que, pour chaque $\theta \in \Theta$,*

$$I_n(\theta) = \mathbb{V}_\theta \left(\nabla \ln L_n(\mathbb{X};\theta) \right).$$

Dorénavant, l'information de Fisher n'est évoquée que lorsque les conditions ci-dessus sont implicitement satisfaites.

Dans cette définition, si $\partial/\partial\theta_i$ désigne la dérivée partielle relativement à la i-ème composante de θ :

$$I_n(\theta) = \left(\mathrm{Cov}_\theta \left(\frac{\partial}{\partial\theta_i} \ln L_n(\mathbb{X};\theta), \frac{\partial}{\partial\theta_j} \ln L_n(\mathbb{X};\theta) \right) \right)_{i,j=1,\ldots,k}.$$

L'information de Fisher est donc une fonction à valeurs dans l'ensemble des matrices symétriques et positives. On notera en particulier que dans le cas réel ($k = 1$)

$$I_n(\theta) = \mathbb{V}_\theta \left(\nabla \ln L_n(\mathbb{X}; \theta) \right) = \mathbb{V}_\theta \left(\frac{\partial}{\partial \theta} \ln L_n(\mathbb{X}; \theta) \right).$$

L'information précise donc le pouvoir de discrimination du modèle entre deux valeurs proches du paramètre du modèle : dans le cas $k = 1$, une grande valeur pour $I_n(\theta)$ traduit une variation importante de la nature des probabilités du modèle au voisinage de P_θ, d'où une discrimination de la vraie valeur du paramètre inconnu facilitée. A l'inverse, si $I_n(\theta)$ est petit, la loi est piquée et on est amené à rechercher le maximum de la vraisemblance dans une région plus vaste.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. La vraisemblance L_n vaut, si $\theta \in]0,1[$ et $(x_1, \ldots, x_n) \in \{0,1\}^n$:

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

Il vient facilement

$$I_n(\theta) = \mathbb{V}_\theta \left(\nabla \ln L_n(\mathbb{X}; \theta) \right) = \frac{n^2}{\theta^2 (1-\theta)^2} \mathbb{V}_\theta \left(\bar{X}_n \right) = \frac{n}{\theta(1-\theta)}.$$

Dans ce modèle, l'incertitude est faible pour θ proche de 0 et 1 alors qu'elle est d'autant plus grande que θ est proche de $1/2$, ce qui se traduit par une information $I_n(\theta)$ maximale pour θ proche de 0 et 1, et minimale pour $\theta = 1/2$.

Dans une situation d'échantillonnage i.i.d., l'information de Fisher est proportionnelle à la taille de l'échantillon :

Proposition 3. *Soit I l'information de Fisher du modèle $(\mathscr{H}, \{Q_\theta\}_{\theta \in \Theta})$ dominé par la mesure μ. Si, pour chaque $\theta \in \Theta$, $P_\theta = Q_\theta^{\otimes n}$, l'information de Fisher I_n du modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta \in \Theta})$ pour la mesure dominante $\nu = \mu^{\otimes n}$ vaut*

$$I_n(\theta) = nI(\theta) \ \forall \theta \in \Theta.$$

Démonstration. Si L désigne la vraisemblance du modèle $(\mathscr{H}, \{Q_\theta\}_{\theta \in \Theta})$, la vraisemblance L_n du modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta \in \Theta})$ vérifie, d'après la Proposition 1 :

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

On en déduit la relation

$$I_n(\theta) = \mathbb{V}_\theta\left(\nabla \ln L_n(\mathbb{X};\theta)\right) = \sum_{i=1}^{n} \mathbb{V}_\theta\left(\nabla \ln L(X_i;\theta)\right) = n\mathbb{V}_\theta\left(\nabla \ln L(X_1;\theta)\right),$$

car les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi. Comme $I(\theta) = \mathbb{V}_\theta\left(\nabla \ln L(X_1;\theta)\right)$, la proposition est démontrée. □

Du point de vue des calculs, on se réfèrera souvent à la proposition qui suit, dont l'objectif principal est de donner une forme simplifiée pour l'information de Fisher. Dans la suite, $\nabla^2 F(\theta)$ désigne la matrice Hessienne de $F : \Theta \to \mathbb{R}$ évaluée en $\theta \in \Theta$. Par convention, l'utilisation de la notation $\nabla^2 F(\theta)$ signifie implicitement que F est de classe $\mathscr{C}^2$ sur un voisinage de θ.

Proposition 4. *Soit I_n l'information de Fisher du modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$. Supposons que pour chaque $\theta \in \Theta$, il existe un voisinage $V \subset \Theta$ de θ tel que $\sup_{\alpha\in V} \|\nabla L_n(\cdot;\alpha)\| \in \mathbb{L}(\nu)$. Alors :*

(**i**) $\mathbb{E}_\theta(\nabla \ln L_n(\mathbb{X};\theta)) = 0$.

(**ii**) *Si, en outre, $\sup_{\alpha\in V} \|\nabla^2 L_n(\cdot;\alpha)\| \in \mathbb{L}^1(\nu)$, on a*

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

Les conditions de cette proposition ne sont pas aussi restrictives qu'elles peuvent le sembler, car elles sont satisfaites par bon nombre de modèles statistiques. Pour la suite, on notera que, pour tout $\theta \in \Theta$, $L_n(\cdot;\theta) > 0$ P_θ-p.s. En effet, puisque $\mathrm{d}P_\theta = L_n(\cdot;\theta)\mathrm{d}\nu$,

$$P_\theta\left(L_n(\cdot;\theta) = 0\right) = \int_{\{L_n(\cdot;\theta)=0\}} L_n(\cdot;\theta)\mathrm{d}\nu = 0.$$

Les dénominateurs en $L_n(\cdot;\theta)$ éventuellement nuls ne posent donc aucun problème lorsque l'on intègre contre $\mathrm{d}P_\theta = L_n(\cdot;\theta)\mathrm{d}\nu$.

Démonstration. Sous la condition $\sup_{\alpha\in V} \|\nabla L_n(\cdot;\alpha)\| \in \mathbb{L}^1(\nu)$, on a, d'après le théorème de dérivation sous l'intégrale,

$$\int_{\mathscr{H}^n} \nabla L_n(\cdot;\theta)\mathrm{d}\nu = \nabla \int_{\mathscr{H}^n} L_n(\cdot;\theta)\mathrm{d}\nu,$$

d'où, puisque $\mathrm{d}P_\theta = L_n(\cdot;\theta)\mathrm{d}\nu$ et $P_\theta(\mathscr{H}^n) = 1$:

$$\int_{\mathscr{H}^n} \nabla L_n(\cdot;\theta)\mathrm{d}\nu = \nabla P_\theta\big(\mathscr{H}^n\big) = 0.$$

Or, comme $\mathrm{d}P_\theta = L_n(\cdot;\theta)\mathrm{d}\nu$ et $\nabla \ln L_n(\cdot;\theta) = \nabla L_n(\cdot;\theta)/L_n(\cdot;\theta)$,

$$\mathbb{E}_\theta\left(\nabla \ln L_n(\mathbb{X};\theta)\right) = \int_{\mathscr{H}^n} \nabla \ln L_n(\cdot;\theta)\, L_n(\cdot;\theta)\mathrm{d}\nu = \int_{\mathscr{H}^n} \nabla L_n(\cdot;\theta)\mathrm{d}\nu,$$

d'où (**i**). Montrons maintenant (**ii**). Si $F : \Theta \to \mathbb{R}$ est de classe $\mathscr{C}^2$, notons pour tout $i, j = 1, \ldots, k$ et $\theta \in \Theta$:

$$\nabla_i F(\theta) = \frac{\partial F}{\partial \theta_i}(\theta) \quad \text{et} \quad \nabla^2_{ij} F(\theta) = \frac{\partial^2 F}{\partial \theta_i \partial \theta_j}(\theta).$$

Le théorème de dérivation sous l'intégrale montre que

$$\int_{\mathscr{H}^n} \nabla^2_{ij} L_n(\cdot;\theta)\mathrm{d}\nu = \nabla^2_{ij} \int_{\mathscr{H}^n} L_n(\cdot;\theta)\mathrm{d}\nu$$

sous l'hypothèse $\sup_{\alpha\in V} \|\nabla^2 L_n(\cdot;\alpha)\| \in \mathbb{L}^1(\nu)$, et donc

$$\int_{\mathscr{H}^n} \nabla^2_{ij} L_n(\cdot;\theta)\mathrm{d}\nu = \nabla_{ij} P_\theta(\mathscr{H}^n) = 0.$$

Par ailleurs, on vérifie que

$$\nabla^2_{ij} \ln L_n(\cdot;\theta) = \frac{\nabla^2_{ij} L_n(\cdot;\theta)}{L_n(\cdot;\theta)} - \frac{\nabla_i L_n(\cdot;\theta)\, \nabla_j L_n(\cdot;\theta)}{L^2_n(\cdot;\theta)}.$$

De ce fait,

$$\begin{aligned}
\mathbb{E}_\theta\big(\nabla^2_{ij} \ln L_n(\mathbb{X};\theta)\big) &= \int_{\mathscr{H}^n} \nabla^2_{ij} \ln L_n(\cdot;\theta)\, L_n(\cdot;\theta)\mathrm{d}\nu \\
&= -\int_{\mathscr{H}^n} \frac{\nabla_i L_n(\cdot;\theta)\nabla_j L_n(\cdot;\theta)}{L_n(\cdot;\theta)}\mathrm{d}\nu \\
&= -\mathbb{E}_\theta\left(\nabla_i \ln L_n(\mathbb{X};\theta)\, \nabla_j \ln L_n(\mathbb{X};\theta)\right),
\end{aligned}$$

car $\nabla \ln L_n(\cdot;\theta) = \nabla L_n(\cdot;\theta)/L_n(\cdot;\theta)$. Or, par définition de l'information de Fisher,

$$\begin{aligned}
I_n(\theta)_{ij} &= \mathrm{Cov}_\theta\left(\nabla_i \ln L_n(\mathbb{X};\theta), \nabla_j \ln L_n(\mathbb{X};\theta)\right) \\
&= \mathbb{E}_\theta\left(\nabla_i \ln L_n(\mathbb{X};\theta)\nabla_j \ln L_n(\mathbb{X};\theta)\right),
\end{aligned}$$

puisque $\mathbb{E}_\theta(\nabla \ln L_n(\mathbb{X};\theta)) = 0$, d'où (**ii**). □

Cette proposition légitime la définition qui suit, en donnant des conditions suffisantes au concept de *régularité* d'un modèle statistique.

Définition 25. *Le modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est dit régulier si les propriétés suivantes sont vérifiées en chaque $\theta \in \Theta$:*

1. *Son information de Fisher I_n en θ existe et est inversible.*
2. $\mathbb{E}_\theta(\nabla \ln L_n(\mathbb{X};\theta)) = 0$ *et* $I_n(\theta) = -\mathbb{E}_\theta(\nabla^2 \ln L_n(\mathbb{X};\theta))$.

Exemple. Le modèle statistique $(\{0,1\}^n, \{\mathscr{B}(\theta)^{\otimes n}\}_{\theta\in\,]0,1[})$ du jeu de pile ou face constitue un exemple de modèle régulier : d'une part, on constate que son information de Fisher est inversible pour chaque $\theta \in]0,1[$; d'autre part, sa vraisemblance, qui vérifie les hypothèses de la Proposition 4, répond donc au second jeu de conditions imposé dans la définition ci-dessus.

6.2 Efficacité

Dans cette section, on suppose en que le modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est régulier, dominé par ν, de vraisemblance L_n et d'information de Fisher I_n. Par ailleurs, le paramètre d'intérêt est $g(\theta)$, où $g : \Theta \to \mathbb{R}$ est une fonction connue, supposée réelle.

Définition 26. *L'estimateur $\hat{g}$ est dit régulier s'il est d'ordre 2 et*

$$\int_{\mathscr{H}^n} \hat{g}\, \nabla L_n(\cdot;\theta)\mathrm{d}\nu = \nabla\Big\{ \int_{\mathscr{H}^n} \hat{g}\, L_n(\cdot;\theta)\mathrm{d}\nu \Big\}.$$

Remarque. En pratique, l'intérêt de cette définition réside dans la remarque suivante : si l'estimateur régulier $\hat{g}$ est sans biais, $\mathbb{E}_\theta\hat{g} = g(\theta)$ pour chaque $\theta \in \Theta$ et donc, comme $\mathrm{d}P_\theta = L_n(\cdot;\theta)\mathrm{d}\nu$:

$$\int_{\mathscr{H}^n} \hat{g}\, \nabla L_n(\cdot;\theta)\mathrm{d}\nu = \nabla\Big\{ \int_{\mathscr{H}^n} \hat{g}\, \mathrm{d}P_\theta \Big\} = \nabla\mathbb{E}_\theta\hat{g} = \nabla g(\theta).$$

Comme le montre le résultat qui suit, le risque quadratique est uniformément minoré dans la famille des estimateurs réguliers et sans biais, nous donnant ainsi une vitesse plancher.

Théorème 6 (Cramér-Rao)**.** *Si $\hat{g}$ est un estimateur régulier et sans biais, alors, pour tout $\theta \in \Theta$,*

$$\mathcal{R}(\hat{g};\theta) \geq \nabla g(\theta)^\top I_n(\theta)^{-1} \nabla g(\theta).$$

Le minorant $\nabla g(\theta)^\top I_n(\theta)^{-1} \nabla g(\theta)$ s'appelle borne de Cramér-Rao.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. Le paramètre d'intérêt est le paramètre du modèle, et l'estimateur sans biais $\bar{X}_n$ construit à partir du n-échantillon vérifie :

$$\mathcal{R}(\bar{X}_n;\theta) = \frac{\theta(1-\theta)}{n}.$$

Nous savons également que l'information de Fisher I_n de ce modèle vaut

$$I_n(\theta) = \frac{n}{\theta(1-\theta)}.$$

Ainsi, le risque quadratique de l'estimateur $\bar{X}_n$ atteint la borne de Cramér-Rao du modèle. Par ailleurs, l'espace des observations étant fini, tout estimateur (d'ordre 2) du paramètre de ce modèle est régulier. En conséquence, tout estimateur d'ordre 2 sans biais $\hat{\theta}$ du paramètre du modèle vérifie

$$\mathcal{R}(\hat{\theta};\theta) \geq \mathcal{R}(\bar{X}_n;\theta) \ \forall \theta \in \Theta,$$

i.e. $\bar{X}_n$ est VUMSB.

Démonstration du Théorème 6. Soit $\theta \in \Theta$. L'estimateur $\hat{g}$ étant régulier et sans biais,

$$\nabla g(\theta) = \nabla \mathbb{E}_\theta \hat{g} = \int_{\mathscr{H}^n} \hat{g}\, \nabla L_n(\cdot;\theta) \mathrm{d}\nu.$$

Comme $\nabla \ln L_n(\cdot;\theta) = \nabla L_n(\cdot;\theta)/L_n(\cdot;\theta)$ et $\mathbb{E}_\theta(\nabla \ln L_n(\mathbb{X};\theta)) = 0$ par régularité du modèle, il vient

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

Pour $u \in \mathbb{R}^d$, on déduit de l'inégalité de Cauchy-Schwarz que

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

Or, par définition de l'information de Fisher,

$$\begin{aligned}\mathbb{E}_\theta\langle u, \nabla \ln L_n(\mathbb{X};\theta)\rangle^2 &= u^\top \left\{\mathbb{E}_\theta\Big(\nabla \ln L_n(\mathbb{X};\theta)\nabla \ln L_n(\mathbb{X};\theta)^\top\Big)\right\} u \\ &= u^\top \mathbb{V}_\theta\left(\nabla \ln L_n(\mathbb{X};\theta)\right) u \\ &= u^\top I_n(\theta) u,\end{aligned}$$

et donc, si $u = I_n(\theta)^{-1}\nabla g(\theta)$:

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

De plus, ce même choix pour u donne $\langle u, \nabla g(\theta)\rangle = \nabla g(\theta)^\top I_n(\theta)^{-1}\nabla g(\theta)$. En conclusion,

$$\mathcal{R}(\hat{g};\theta) \geq \nabla g(\theta)^\top I_n(\theta)^{-1}\nabla g(\theta),$$

d'où le théorème. □

Définition 27. *L'estimateur $\hat{g}$ sans biais d'ordre 2 est dit efficace si son risque quadratique atteint la borne de Cramér-Rao du modèle, i.e. pour tout $\theta \in \Theta$:*

$$\mathcal{R}(\hat{g};\theta) = \nabla g(\theta)^\top I_n(\theta)^{-1}\nabla g(\theta).$$

Lorsqu'un estimateur efficace existe, la borne de Cramér-Rao fournit un majorant pour la plus petite erreur quadratique dans la famille des estimateurs sans biais d'ordre 2 (pas nécessairement réguliers). En particulier, si $\hat{g}$ est un estimateur VUMSB,

$$\mathcal{R}(\hat{g};\theta) \leq \nabla g(\theta)^\top I_n(\theta)^{-1}\nabla g(\theta)\ \forall \theta \in \Theta.$$

Les estimateurs efficaces sont souvent simples à caractériser. On peut par exemple montrer qu'un estimateur $\hat{g}$ régulier et sans biais est efficace si et seulement si il existe une fonction $\psi : \Theta \to \mathbb{R}^k$ telle que pour tout $\theta \in \Theta$:

$$\hat{g} = g(\theta) + \langle \psi(\theta), \nabla \ln L_n(\mathbb{X};\theta)\rangle \ \mathbb{P}\text{-p.s.}$$

6.3 Exhaustivité

Dans le modèle statistique $(\{0,1\}^n, \{\mathscr{B}(\theta)^{\otimes n}\}_{\theta\in]0,1[})$ du jeu de pile ou face, l'observation $(X_1, \ldots, X_n)$ issue de la loi $\mathscr{B}(\theta)^{\otimes n}$ peut être résumée par sa

moyenne $\bar{X}_n$, sans perte d'information sur θ. Pour s'en convaincre, il suffit de voir que pour chaque $(x_1, \ldots, x_n) \in \{0,1\}^n$, on a :

$$\begin{aligned}\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \,\middle|\, \bar{X}_n = \bar{x}_n\right) &= \frac{\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right)}{\mathbb{P}(\bar{X}_n = \bar{x}_n)} \\ &= \frac{\theta^{n\bar{x}_n}(1-\theta)^{n-n\bar{x}_n}}{C_n^{n\bar{x}_n}\,\theta^{n\bar{x}_n}(1-\theta)^{n-n\bar{x}_n}} \\ &= \frac{1}{C_n^{n\bar{x}_n}},\end{aligned}$$

car $n\bar{X}_n$ suit la loi $\mathscr{B}(n,\theta)$. La loi conditionnelle de $(X_1, \ldots, X_n)$ sachant $\bar{X}_n$ est donc indépendante du paramètre θ. En d'autres termes, toute l'information sur θ contenue dans l'échantillon $(X_1, \ldots, X_n)$ est en fait contenue dans $\bar{X}_n$. On dit alors que $\bar{X}_n$ est une *statistique exhaustive*.

Définition 28. *La statistique $S(\mathbb{X})$ à valeurs dans $\mathbb{R}^q$ est dite exhaustive lorsque, pour chaque $\theta \in \Theta$, la loi conditionnelle de $\mathbb{X} = (X_1, \ldots, X_n)$ sachant $S(\mathbb{X})$ ne dépend pas de θ.*

En clair, une statistique exhaustive élimine toute l'information superflue contenue dans l'observation, en ne retenant que la partie informative sur le paramètre du modèle.

Remarques.

1. Du point de vue technique, la statistique $S(\mathbb{X})$ à valeurs dans $\mathbb{R}^q$ est exhaustive s'il existe un noyau de transition K sur $\mathbb{R}^q \times \mathscr{B}(\mathscr{H}^n)$ tel que pour tout $\theta \in \Theta$, $A \in \mathscr{B}(\mathbb{R}^q)$ et $B \in \mathscr{B}(\mathscr{H}^n)$,

   $$\mathbb{P}\left(\{S(\mathbb{X}) \in A\} \cap \{\mathbb{X} \in B\}\right) = \int_A K(\cdot, B)\, \mathrm{d}P_\theta \circ S^{-1}.$$

   Pour une telle statistique exhaustive, si $\hat{g}$ est un estimateur d'ordre 1, la quantité

   $$\mathbb{E}_\theta\left(\hat{g} | S(\mathbb{X})\right) = \int_{\mathscr{H}^n} \hat{g}(x) K(S(\mathbb{X}), \mathrm{d}x)$$

   est indépendante de θ, donc $\mathbb{E}_\theta(\hat{g}|S(\mathbb{X}))$ est une statistique.

2. Une statistique exhaustive peut prendre ses valeurs dans $\mathbb{R}^q$ avec $q > 1$. Il peut donc, en particulier, s'agir d'un vecteur.

3. L'observation $\mathbb{X} = (X_1, \ldots, X_n)$ est toujours une statistique exhaustive car la loi conditionnelle de $(X_1, \ldots, X_n)$ sachant $(X_1, \ldots, X_n)$ est la mesure de Dirac en $(X_1, \ldots, X_n)$. Cependant, cette statistique, qui ne résume pas l'information sur le paramètre contenue dans l'observation, doit être délaissée au profit de n'importe quelle autre statistique exhaustive.
4. Il n'y a pas unicité des statistiques exhaustives. En effet, si $S(\mathbb{X}) = \varphi(T(\mathbb{X}))$ est exhaustive, avec φ une fonction borélienne quelconque, il en va de même de $T(\mathbb{X})$ (pourquoi ?).

Malgré son apparence, la propriété d'exhaustivité est souvent simple à établir, comme le montre le résultat suivant.

Théorème 7 (Neyman-Fisher)**.** *Supposons le modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ dominé par ν, de vraisemblance L_n. Une statistique $S(\mathbb{X})$ à valeurs dans $\mathbb{R}^q$ est exhaustive si et seulement si il existe deux fonctions boréliennes $\psi : \mathbb{R}^q \times \Theta \to \mathbb{R}_+$ et $\gamma : \mathscr{H}^n \to \mathbb{R}_+$ telles que, pour tout $\theta \in \Theta$,*

$$L_n(\cdot;\theta) = \gamma(\cdot)\,\psi\big(S(\cdot),\theta\big) \ \nu\text{-p.p.}$$

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. La moyenne empirique $\bar{X}_n$ est une statistique exhaustive car la vraisemblance L_n pour la mesure de Lebesgue vaut

$$\begin{aligned} L_n(x_1, \ldots, x_n;\theta) &= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2\Big) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{n}{2}(\bar{x}_n-\theta)^2\Big) \exp\Big(-\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x}_n)^2\Big), \end{aligned}$$

pour tout $(x_1, \ldots, x_n) \in \mathbb{R}^n$ et $\theta \in \mathbb{R}$.

Démonstration du Théorème 7. Supposons, pour simplifier la preuve, que la mesure dominante ν appartient au convexifié [1] de $\{P_\theta\}_{\theta\in\Theta}$, i.e.

$$\nu = \sum_{i\geq 1} a_i P_{\theta_i},$$

1. On peut effectivement montrer qu'il existe une probabilité du convexifié qui domine le modèle.

avec $\theta_i \in \Theta$, $a_i \geq 0$ pour chaque $i \geq 1$ et $\sum_{i\geq 1} a_i = 1$.

Dans ce cadre, nous allons montrer que $S(\mathbb{X})$ est exhaustive si et seulement si il existe une fonction borélienne $\psi : \mathbb{R}^q \times \Theta \to \mathbb{R}_+$ telle que, pour tout $\theta \in \Theta$,

$$L_n(\cdot;\theta) = \psi(S,\theta) \ \nu\text{-p.p.}$$

Fixons $\theta \in \Theta$ et supposons que $L_n(\cdot;\theta) = \psi(S,\theta)$ ν-p.p. Comme $\mathrm{d}P_\theta = \psi(S,\theta)\mathrm{d}\nu$, par définition de l'espérance conditionnelle pour la probabilité ν, notée $\mathbb{E}_\nu$, on a

$$\begin{aligned}\mathbb{P}\left(\{S(\mathbb{X}) \in A\} \cap \{\mathbb{X} \in B\}\right) &= \int_{\mathscr{H}^n} \mathbb{1}_A(S)\, \mathbb{1}_B\, \psi(S,\theta)\, \mathrm{d}\nu \\ &= \int_{\mathscr{H}^n} \mathbb{E}_\nu\left(\mathbb{1}_A(S)\, \mathbb{1}_B\, \psi(S,\theta) \mid S\right) \mathrm{d}\nu,\end{aligned}$$

pour tout $A \in \mathscr{B}(\mathbb{R}^q)$ et $B \in \mathscr{B}(\mathscr{H}^n)$. Puis, $\mathbb{1}_A(S)\, \psi(S,\theta)$ étant une fonction borélienne de S, on a en notant $\nu(B|S) = \mathbb{E}_\nu(\mathbb{1}_B|S)$:

$$\begin{aligned}\mathbb{P}\left(\{S(\mathbb{X}) \in A\} \cap \{\mathbb{X} \in B\}\right) &= \int_{\mathscr{H}^n} \nu(B\,|S)\, \mathbb{1}_A(S)\, \psi(S,\theta)\, \mathrm{d}\nu \\ &= \int_{\mathbb{R}^q} \nu(B \mid S = y)\, \mathbb{1}_A(y)\, \psi(y,\theta)\, \nu \circ S^{-1}(\mathrm{d}y),\end{aligned}$$

la dernière égalité provenant du théorème de transfert. Or, $P_\theta \circ S^{-1}$ est absolument continue par rapport à $\nu \circ S^{-1}$, et sa densité est $\mathbb{E}_\nu(L_n(\cdot;\theta)|S = \cdot)$ [2]. De ce fait, $\mathrm{d}P_\theta \circ S^{-1} = \psi(\cdot,\theta)\, \mathrm{d}\nu \circ S^{-1}$ car $L_n(\cdot,\theta) = \psi(S,\theta)$ est $\sigma(S)$-mesurable, d'où

$$\mathbb{P}\left(\{S(\mathbb{X}) \in A\} \cap \{\mathbb{X} \in B\}\right) = \int_A \nu(B\,|S = y)\, P_\theta \circ S^{-1}(\mathrm{d}y).$$

L'application $(y, B) \mapsto \nu(B|S = y)$ définie sur $\mathbb{R}^q \times \mathscr{B}(\mathscr{H}^n)$ est donc le noyau de transition de la loi conditionnelle sachant S. Comme il est indépendant de θ, S est une statistique exhaustive.

2. En effet, pour tout $A \in \mathscr{B}(\mathbb{R}^q)$, par définition de l'espérance conditionnelle :

$$P_\theta \circ S^{-1}(A) = \int_{\{S\in A\}} L_n(\cdot;\theta)\, \mathrm{d}\nu = \int_{\{S\in A\}} \mathbb{E}_\nu(L_n(\cdot;\theta) \mid S)\, \mathrm{d}\nu.$$

D'après le théorème de transfert, $P_\theta \circ S^{-1}(A) = \int_A \mathbb{E}_\nu(L_n(\cdot;\theta) \mid S = y)\, \nu \circ S^{-1}(\mathrm{d}y)$, d'où le résultat annoncé.

Réciproquement, supposons que $S(\mathbb{X})$ est exhaustive. Pour tout $\theta \in \Theta$, la loi conditionnelle $P_\theta(\cdot \mid S = \cdot)$ de l'observation sachant S est indépendante de θ ; notons-là $P(\cdot \mid S = \cdot)$. Si $A \in \mathscr{B}(\mathbb{R}^q)$ et $B \in \mathscr{B}(\mathscr{H}^n)$, on a d'une part

$$\begin{aligned}\nu\left(\{S \in A\} \cap B\right) &= \sum_{i\geq 1} a_i P_{\theta_i}\left(\{S \in A\} \cap B\right) \\ &= \sum_{i\geq 1} a_i \int_A P_{\theta_i}(B|S = y)\, \nu(\mathrm{d}y) \\ &= \int_A P(B|S = y)\, \nu(\mathrm{d}y),\end{aligned}$$

car $\sum_{i\geq 1} a_i = 1$. D'autre part, en désignant par $\nu(\cdot|S = \cdot)$ la loi conditionnelle sachant S, on sait que

$$\nu\left(\{S \in A\} \cap B\right) = \int_A \nu(B|S = y)\, \nu(\mathrm{d}y).$$

Puisque ces relations sont vraies pour tout $A \in \mathscr{B}(\mathbb{R}^q)$ et $B \in \mathscr{B}(\mathscr{H}^n)$, on en déduit par unicité que les lois conditionnelles $P(\cdot \mid S = \cdot)$ et $\nu(\cdot \mid S = \cdot)$ sont les mêmes, d'où

$$\mathbb{P}\left(\{S(\mathbb{X}) \in A\} \cap \{\mathbb{X} \in B\}\right) = \int_A \nu(B \mid S = y)\, P_\theta \circ S^{-1}(\mathrm{d}y).$$

On a déjà remarqué que $\psi(\cdot, \theta) = \mathbb{E}_\nu(L_n(\cdot;\theta) \mid S = \cdot)$ est la densité de $P_\theta \circ S^{-1}$ par rapport à $\nu \circ S^{-1}$. De ce fait,

$$\mathbb{P}\left(\{S(\mathbb{X}) \in A\} \cap \{\mathbb{X} \in B\}\right) = \int_A \nu(B \mid S = y)\psi(y, \theta)\, \nu \circ S^{-1}(\mathrm{d}y).$$

Par ailleurs, la définition de l'espérance conditionnelle et le théorème de transfert donnent

$$\begin{aligned}\mathbb{P}\left(\{S(\mathbb{X}) \in A\} \cap \{\mathbb{X} \in B\}\right) &= \int_{\{S\in A\}} \mathbb{1}_B\, L_n(\cdot;\theta)\, \mathrm{d}\nu \\ &= \int_A \mathbb{E}_\nu(\mathbb{1}_B\, L_n(\cdot;\theta) \mid S = y)\, \nu \circ S^{-1}(\mathrm{d}y).\end{aligned}$$

Ces égalités étant vraies pour tout $A \in \mathscr{B}(\mathbb{R}^q)$, il vient

$$\nu(B \mid S = \cdot)\psi(\cdot, \theta) = \mathbb{E}_\nu(\mathbb{1}_B\, L_n(\cdot;\theta) \mid S = \cdot)\ \nu \circ S^{-1}\text{-p.s.},$$

et donc

$$\mathbb{E}_\nu\left(\mathbb{1}_B\ (\psi(S, \theta) - L_n(\cdot;\theta))\ \middle|\ S\right) = 0\ \nu\text{-p.s.}$$

En particulier,

$$\begin{aligned}\mathbb{E}_\nu\left[\mathbb{1}_B\ (\psi(S,\theta)-L_n(\cdot;\theta))\right] &= \mathbb{E}_\nu\mathbb{E}_\nu\left(\mathbb{1}_B\ (\psi(S,\theta)-L_n(\cdot;\theta))\ \middle|\ S\right)\\ &= 0.\end{aligned}$$

Ceci étant vrai pour tout $B \in \mathscr{B}(\mathscr{H}^n)$, $L_n(\cdot;\theta) = \psi(S,\theta)$ ν-p.s., d'où le théorème. □

Le concept d'exhaustivité permet d'améliorer le risque d'un estimateur :

Théorème 8. [Rao-Blackwell] *Soit $S(\mathbb{X})$ une statistique exhaustive et $\hat{g}$ un estimateur d'ordre 2. Alors $\mathbb{E}_\theta(\hat{g}|S(\mathbb{X}))$ est un estimateur de même biais que $\hat{g}$ qui lui est préférable, i.e. pour tout $\theta \in \Theta$:*

$$\mathcal{R}\left(\mathbb{E}_\theta(\hat{g}|S\left(\mathbb{X}\right));\theta\right) \leq \mathcal{R}(\hat{g};\theta).$$

Exemple. $\mathbb{X} = (X_1,\ldots,X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in]0,1[$. Alors X_1 est un estimateur sans biais de θ et $\bar{X}_n$ est une statistique exhaustive. A l'aide du théorème de Rao-Blackwell, nous allons améliorer l'erreur quadratique de l'estimateur X_1 en calculant $\mathbb{E}_\theta(X_1|\bar{X}_n)$. Comme les variables aléatoires $X_1,\ldots,X_n$ sont indépendantes et de même loi, pour tout $i \in \{1,\ldots,n\}$ et $A \subset \{k/n,\, k=0,\ldots,n\}$:

$$\mathbb{E}_\theta\big(X_1\mathbb{1}_{[\bar{X}_n\in A]}\big) = \mathbb{E}_\theta\big(X_i\mathbb{1}_{[\bar{X}_n\in A]}\big).$$

Ceci étant vrai pour chaque $A \subset \{k/n,\, k=0,\ldots,n\}$, on en déduit via l'unicité de l'espérance conditionnelle que $\mathbb{E}_\theta(X_1|\bar{X}_n) = \mathbb{E}_\theta(X_i|\bar{X}_n)$ $\mathbb{P}$-p.s. Par suite,

$$\mathbb{E}_\theta(X_1|\bar{X}_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta(X_i|\bar{X}_n) = \mathbb{E}_\theta(\bar{X}_n|\bar{X}_n) = \bar{X}_n\ \mathbb{P}\text{-p.s.}$$

Ainsi, l'estimateur préférable construit avec le théorème de Rao-Blackwell n'est autre que la moyenne empirique.

Démonstration du Théorème 8. Soit $\theta \in \Theta$. Par exhaustivité de $S(\mathbb{X})$, $\hat{\eta} = \mathbb{E}_\theta[\hat{g}|S(\mathbb{X})]$, qui ne dépend pas de θ, est donc un estimateur. De plus, puisque

$$\mathbb{E}_\theta\hat{\eta} = \mathbb{E}_\theta\mathbb{E}_\theta\left(\hat{g}|S(\mathbb{X})\right) = \mathbb{E}_\theta\hat{g},$$

les estimateurs $\hat{\eta} = \mathbb{E}_\theta(\hat{g}|S(\mathbb{X}))$ et $\hat{g}$ ont même biais. Notons maintenant, pour un estimateur $\hat{\psi}$ d'ordre 2,

$$V_\theta(\hat{\psi}) = \mathbb{E}_\theta\|\hat{\psi} - \mathbb{E}_\theta\hat{\psi}\|^2.$$

Alors,

$$\begin{aligned} V_\theta\hat{g} &= \mathbb{E}_\theta \left\|(\hat{g} - \hat{\eta}) + (\hat{\eta} - \mathbb{E}_\theta\hat{g})\right\|^2 \\ &= \mathbb{E}_\theta\|\hat{g} - \hat{\eta}\|^2 + V_\theta\hat{\eta} + 2\mathbb{E}_\theta\langle\hat{g} - \hat{\eta}, \hat{\eta} - \mathbb{E}_\theta\hat{\eta}\rangle, \end{aligned}$$

car $\hat{g}$ et $\hat{\eta}$ ont même biais. Or, comme $\hat{\eta}$ est une fonction borélienne de $S(\mathbb{X})$,

$$\begin{aligned} \mathbb{E}_\theta\left(\langle\hat{g} - \hat{\eta}, \hat{\eta} - \mathbb{E}_\theta\hat{\eta}\rangle\middle|S(\mathbb{X})\right) &= \langle\mathbb{E}_\theta\left(\hat{g} - \hat{\eta}|S(\mathbb{X})\right), \hat{\eta} - \mathbb{E}_\theta\hat{\eta}\rangle \\ &= \langle\hat{\eta} - \hat{\eta}, \hat{\eta} - \mathbb{E}_\theta\hat{\eta}\rangle \\ &= 0. \end{aligned}$$

De ce fait,

$$\mathbb{E}_\theta\langle\hat{g} - \hat{\eta}, \hat{\eta} - \mathbb{E}_\theta\hat{\eta}\rangle = \mathbb{E}_\theta\mathbb{E}_\theta\left(\langle\hat{g} - \hat{\eta}, \hat{\eta} - \mathbb{E}_\theta\hat{\eta}\rangle\middle|S(\mathbb{X})\right) = 0,$$

et donc $V_\theta\hat{g} \geq V_\theta\hat{\eta}$. D'après la décomposition (2.1),

$$\mathcal{R}(\hat{\eta};\theta) = \|\mathbb{E}_\theta\hat{\eta} - g(\theta)\|^2 + V_\theta\hat{\eta} \leq \|\mathbb{E}_\theta\hat{g} - g(\theta)\|^2 + V_\theta\hat{g} = \mathcal{R}(\hat{g};\theta),$$

d'où le théorème. □

Conclusion. Chaque fois que l'on dispose d'un estimateur $\hat{g}$ et d'une statistique exhaustive $S(\mathbb{X})$, on a intérêt à remplacer $\hat{g}$ par $\mathbb{E}_\theta(\hat{g}|S(\mathbb{X}))$ qui, par définition, ne dépend pas de θ. On a alors deux cas de figure :

1. Ou bien $\mathbb{E}_\theta(\hat{g}|S(\mathbb{X})) = \hat{g}$ $\mathbb{P}$-p.s. : on n'a rien changé, mais $\hat{g}$ est déjà une fonction de $S(\mathbb{X})$.
2. Ou bien $\mathbb{E}_\theta(\hat{g}|S(\mathbb{X})) \neq \hat{g}$ $\mathbb{P}$-p.s. : on gagne.

Il est donc clair qu'il est inutile de considérer des estimateurs qui ne sont pas fonctions de $S(\mathbb{X})$. La difficulté en fait vient de ce qu'il n'y a pas unicité des statistiques exhaustives.

6.4 Comparaison des statistiques exhaustives

Un problème se pose lorsque l'on dispose de deux statistiques exhaustives, $S(\mathbb{X})$ et $T(\mathbb{X})$. Laquelle doit-on utiliser de préférence ? Considérons d'abord le cas le plus simple, lorsqu'il existe une fonction φ telle que $S(\mathbb{X}) = \varphi(T(\mathbb{X}))$. Alors toute fonction de $S(\mathbb{X})$ est aussi une fonction de $T(\mathbb{X})$ (mais la réciproque n'est pas nécessairement vraie). Si φ est inversible, l'ensemble des fonctions de $S(\mathbb{X})$ est aussi celui des fonctions de $T(\mathbb{X})$, et conditionner par rapport à l'une ou l'autre statistique conduit au même résultat. Il est donc indifférent d'utiliser l'une ou l'autre.

La situation est différente si φ n'est pas inversible car il existe alors des fonctions de $T(\mathbb{X})$ qui ne sont pas des fonctions de $S(\mathbb{X})$. En particulier, si un estimateur $\hat{g}(T(\mathbb{X}))$ n'est pas une fonction de $S(\mathbb{X})$, on a

$$\hat{h}\left(S(\mathbb{X})\right) = \mathbb{E}_\theta\left[\hat{g}\left(T(\mathbb{X})\right)\middle|S(\mathbb{X})\right] \neq \hat{g}\left(T(\mathbb{X})\right)$$

et le Théorème 8 nous dit que, pour le risque quadratique, $\hat{h}(S(\mathbb{X}))$ sera meilleur que $\hat{g}(T(\mathbb{X}))$, ce qui montre qu'il convient de préférer $S(\mathbb{X})$ à $T(\mathbb{X})$.

On a donc des statistiques exhaustives plus ou moins "grosses". Par exemple, dans un n-échantillon $\mathcal{N}(\theta, 1)$, $\bar{X}_n$ est exhaustive, mais aussi le couple $(X_1, \bar{X}_n)$, $1/\bar{X}_n$ (qui existe $\mathbb{P}$-p.s.), ou $(X_1, \ldots, X_n)$. Il est clair que $\bar{X}_n$ est meilleure que $(X_1, \bar{X}_n)$ ou $(X_1, \ldots, X_n)$ d'après ce qui précède, et équivalente à $1/\bar{X}_n$. On a ainsi intérêt à chercher des statistiques "minimales" au sens précédent. Chaque fois que l'on remplace $S(\mathbb{X})$ par $\varphi(S(\mathbb{X}))$ où φ n'est pas bijective, on gagne et clairement $\varphi(T(\mathbb{X}))$ est plus simple. C'est très clair lorsque l'on remplace un vecteur par un nombre réel.

Exemples.

1. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de densité commune (par rapport à une mesure de référence sur $\mathbb{R}^d$) de la forme

   $$f_\theta(x) = C(\theta) \exp\left[Q(\theta)T(x)\right] h(x),$$

   avec $\theta \in \Theta \subset \mathbb{R}$, $Q(\theta), T(x) \in \mathbb{R}$, et $C(\theta), h(x) \in \mathbb{R}_+$. Ce modèle porte le nom de *modèle exponentiel général* (de dimension 1) et on vérifiera que la plupart des lois classiques (lois normale, exponentielle,

gamma, Bernoulli, binomiale, de Poisson, etc.) tombent dans son escarcelle. La vraisemblance s'écrit, pour $(x_1, \ldots, x_n) \in \mathbb{R}^{dn}$ et $\theta \in \Theta$,

$$L_n(x_1, \ldots, x_n; \theta) = C^n(\theta) \exp\left[Q(\theta) \sum_{i=1}^{n} T(x_i)\right] \prod_{i=1}^{n} h(x_i),$$

et le Théorème 7 indique que $\sum_{i=1}^{n} T(X_i)$ et $\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$ sont deux statistiques exhaustives équivalentes. Dans le modèle gaussien $\mathcal{N}(m, \sigma^2)$, $m \in \mathbb{R}$ et $\sigma^2 > 0$, $\bar{X}_n$ est exhaustive si σ est connu et $\sum_{i=1}^{n}(X_i - m)^2$ l'est lorsque c'est m qui est inconnu. Pour un n-échantillon de loi de Poisson $\mathscr{P}(\theta)$, c'est $\bar{X}_n$ qui est exhaustive.

2. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de densité commune $\mathscr{U}([0, \theta])$, $\theta > 0$. Dans ce cas, comme déjà vu, pour $(x_1, \ldots, x_n) \in [0, \theta]^n$,

$$L_n(x_1, \ldots, x_n; \theta) = \theta^{-n} \mathbb{1}_{[0,\theta]}(x_{(n)}),$$

où $x_{(n)} = \max(x_1, \ldots, x_n)$, ce qui montre que $X_{(n)}$ est une statistique exhaustive. La situation est différente si l'on considère les lois uniformes $\mathscr{U}([\theta, \theta + 1])$ avec $\theta \in \mathbb{R}$, ou $\mathscr{U}([-\theta, \theta])$ avec $\theta > 0$. Dans ce dernier cas, en notant $x_{(1)} = \min(x_1, \ldots, x_n)$,

$$L_n(x_1, \ldots, x_n; \theta) = (2\theta)^{-n} \theta^{-n} \mathbb{1}_{[-\theta, +\infty[}(x_{(1)}) \mathbb{1}_{]-\infty, \theta]}(x_{(n)}),$$

et le couple $(X_{(1)}, X_{(n)})$ forme une statistique exhaustive. L'autre exemple se traite de la même façon.

3. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de densité commune $f_\theta(x)$, $\theta \in \Theta$, par rapport à une mesure de référence sur $\mathbb{R}^d$. Le vecteur des statistiques d'ordre $(X_{(1)}, \ldots, X_{(n)})$ est toujours exhaustif car

$$L_n(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_{(i)}), \text{ avec } x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Dans un modèle d'échantillonnage, seul compte l'ensemble des valeurs (avec éventuelles répétitions) prises par les observations, pas l'ordre dans lequel on les a observées.

Chapitre 7

Tests d'adéquation et d'indépendance

Les tests d'adéquation tels que le test du χ^2 d'adéquation et le test de Kolmogorov-Smirnov ont pour objectif de préciser la loi dont l'observation est issue. Les tests d'indépendance, comme le test du χ^2 d'indépendance, étudient l'absence de lien entre deux observations. La construction et l'étude de ces trois tests fait l'objet de ce chapitre.

7.1 Test du χ^2 d'adéquation

Dans cette section, $\mathscr{H}$ est un ensemble fini identifié à $\mathscr{H} = \{1, \ldots, d\}$. Considérons le modèle statistique $(\mathscr{H}^n, \{P^{\otimes n}\}_{P \in \mathscr{P}})$, $\mathscr{P}$ désignant l'ensemble des probabilités sur $\mathscr{H}$ de support $\mathscr{H}$, c'est-à-dire que pour chaque $P \in \mathscr{P}$, $P(j) > 0 \ \forall j \in \mathscr{H}$. Ce modèle est paramétrique car chaque loi de $\mathscr{P}$ est à support dans l'ensemble fini $\mathscr{H}$.

Fixons $P_0 \in \mathscr{P}$. Nous allons construire un test asymptotique dans le cadre du problème de test de

$$H_0 : P = P_0 \quad \text{contre} \quad H_1 : P \neq P_0.$$

Selon le principe habituel, l'enjeu est de trouver une statistique mesurant la proximité entre P_0 et une version empirique de la loi de l'échantillon. Cette statistique doit être de surcroît asymptotiquement libre sous H_0, ce

qui signifie que sa loi asymptotique sous H_0 ne doit pas dépendre de la loi de l'échantillon. Dans le cadre du test du χ^2 d'adéquation, elle est donnée par la *statistique de Pearson*, définie pour l'observation $\mathbb{X} = (X_1, \dots, X_n)$ par

$$\hat{D}_n(\mathbb{X}, P_0) = n \sum_{j=1}^{d} \frac{\left(\hat{P}_n(j) - P_0(j)\right)^2}{P_0(j)},$$

où $\hat{P}_n$ désigne la mesure définie par

$$\hat{P}_n(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{j\}}(X_i) \ \forall j \in \mathscr{H}.$$

Cette statistique, également appelée *pseudo-distance du* χ^2, est un bon candidat pour construire un test asymptotique dans le cadre du problème de test de H_0 contre H_1.

Théorème 9. *Soient* $P \in \mathscr{P}$ *et* $\mathbb{X} = (X_1, \dots, X_n) \sim P^{\otimes n}$.

(**i**) *Sous* H_0, *i.e.* $P = P_0$,

$$\hat{D}_n(\mathbb{X}, P_0) \xrightarrow{\mathcal{L}} \chi^2(d-1).$$

(**ii**) *Sous* H_1, *i.e.* $P \neq P_0$,

$$\hat{D}_n(\mathbb{X}, P_0) \xrightarrow{\mathbb{P}} +\infty.$$

Dans le problème de test de H_0 contre H_1, la région de rejet d'un test basé sur la statistique de Pearson est de la forme $\{x \in \mathscr{H}^n : \hat{D}_n(x, P_0) \geq s\}$, car H_0 est rejetée lorsque la statistique de test prend des valeurs anormalement grandes. Le test du χ^2 d'adéquation est le test de région de rejet

$$R(\alpha) = \left\{ x \in \mathscr{H}^n : \hat{D}_n(x, P_0) \geq \chi^2_{1-\alpha}(d-1) \right\},$$

avec $\alpha \in]0,1[$ et $\chi^2_{1-\alpha}(d-1)$ le quantile d'ordre $(1-\alpha)$ de la loi $\chi^2(d-1)$. Ce test asymptotique est de niveau α puisque, d'après la propriété (**i**) du Théorème 9, sous H_0,

$$\lim_{n \to +\infty} \mathbb{P}\left(\mathbb{X} \in R(\alpha)\right) = \lim_{n \to +\infty} \mathbb{P}\left(\hat{D}_n(\mathbb{X}, P_0) \geq \chi^2_{1-\alpha}(d-1)\right) = \alpha.$$

Il est de plus convergent d'après l'assertion (**ii**). La procédure de test est donc définie ainsi : H_0 est rejetée au niveau asymptotique α lorsque $(X_1, \dots, X_n) \in R(\alpha)$.

Remarques.

1. Il s'agit d'un test asymptotique. En pratique, on demande souvent que $n \geq 30$ et $nP_0(j) \geq 5$ pour chaque $j \in \mathscr{H}$. Si cette dernière condition n'est pas vérifiée, on effectue des regroupements de classes voisines.
2. Ce test est en pratique souvent utilisé dans le cadre plus général où l'espace des observations n'est pas fini : le principe est de se ramener au cas fini en considérant une partition finie de l'espace des observations ; toute la difficulté est alors de choisir convenablement la partition.

Exemple. Une usine fabrique des bonbons de 8 couleurs différentes : bleu, vert, marron, rose, rouge, orange, jaune, violet. L'objectif est de savoir si le procédé de mise en sachets des bonbons produit une répartition uniforme des 8 couleurs. Le tableau ci-dessous donne la répartition dans un sachet contenant $n = 83$ bonbons :

Bleu	Vert	Marron	Rose	Rouge	Orange	Jaune	Violet
15	11	11	10	10	10	8	8

En attribuant un chiffre à chaque couleur et en considérant que les couleurs des bonbons sont indépendantes, le modèle statistique associé à cette expérience est $(\mathscr{H}^n, \{P^{\otimes n}\}_{P\in\mathscr{P}})$, où $\mathscr{H} = \{1,\dots,8\}$ et $\mathscr{P}$ est l'ensemble des probabilités de support $\mathscr{H}$. Le problème de test s'écrit $H_0 : P = P_0$ contre $H_1 : P \neq P_0$, avec P_0 la loi uniforme sur $\mathscr{H}$. Le quantile d'ordre 95% de la loi $\chi^2(7)$ vaut environ 14.07 donc, pour un niveau $\alpha = 5\%$, la région de rejet est

$$R(0.05) = \left\{x \in \mathscr{H}^n : \hat{D}_n(x, P_0) \geq 14.07\right\}.$$

Avec l'observation $x = (x_1,\dots,x_n) \in \mathscr{H}^n$ résumée dans le tableau ci-dessus, la statistique de Pearson vaut $\hat{D}_n(x, P_0) = 3.27$. Par suite, $x \notin R(0.05)$, i.e. H_0 est conservée au niveau 5%.

Preuve du Théorème 9. Montrons tout d'abord (**ii**). Si $P \neq P_0$, il existe $j \in \mathscr{H}$ tel que $P(j) \neq P_0(j)$. La propriété annoncée se déduit de l'inégalité

$$\hat{D}_n(\mathbb{X}, P_0) \geq n\frac{\left(\hat{P}_n(j) - P_0(j)\right)^2}{P_0(j)},$$

car $\hat{P}_n(j) \xrightarrow{\mathbb{P}} P(j)$ d'après la loi des grands nombres.

Montrons maintenant (**i**). Soient p_0, $\sqrt{p_0}$ et $Y_1, \ldots, Y_n$ les vecteurs de $\mathbb{R}^d$ définis par

$$p_0 = \begin{pmatrix} P_0(1) \\ \vdots \\ P_0(d) \end{pmatrix}, \quad \sqrt{p_0} = \begin{pmatrix} \sqrt{P_0(1)} \\ \vdots \\ \sqrt{P_0(d)} \end{pmatrix} \quad \text{et} \quad Y_i = \begin{pmatrix} \mathbb{1}_{\{1\}}(X_i) \\ \vdots \\ \mathbb{1}_{\{d\}}(X_i) \end{pmatrix},$$

pour $i = 1, \ldots, n$. En désignant par $\bar{Y}_n$ la moyenne empirique de $Y_1, \ldots, Y_n$, on trouve la représentation :

$$\hat{D}_n(\mathbb{X}, P_0) = \|Z_n\|^2, \text{ avec } Z_n = \sqrt{n}\,\mathrm{diag}(\sqrt{p_0})^{-1}(\bar{Y}_n - p_0),$$

avec $\mathrm{diag}(\alpha)$ la matrice diagonale dont les éléments diagonaux sont ceux du vecteur $\alpha = (\alpha_1 \ldots \alpha_d)^\top$. Comme Y_1 a (sous H_0) pour moyenne p_0 et pour matrice de variance-covariance $\mathbb{V} = \mathrm{diag}(p_0) - p_0 p_0^\top$, on obtient, avec le théorème central limite,

$$\sqrt{n}(\bar{Y}_n - p_0) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \mathbb{V}).$$

Par suite,

$$Z_n \xrightarrow{\mathcal{L}} \mathrm{diag}(\sqrt{p_0})^{-1}\mathcal{N}_d(0, \mathbb{V}).$$

Or, un calcul montre que $\mathrm{diag}(\sqrt{p_0})^{-1}\mathbb{V}\mathrm{diag}(\sqrt{p_0})^{-1} = \mathrm{Id} - \sqrt{p_0}\sqrt{p_0}^\top$ d'où, si $\Gamma = \mathrm{Id} - \sqrt{p_0}\sqrt{p_0}^\top$:

$$Z_n \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Gamma).$$

Il reste à identifier la loi du carré de la norme euclidienne d'un vecteur aléatoire de loi $\mathcal{N}_d(0, \Gamma)$. La matrice Γ est le projecteur orthogonal dans l'orthogonal du sous-espace vectoriel E engendré par le vecteur $\sqrt{p_0}$. Elle est en particulier symétrique et idempotente. Si N est un vecteur aléatoire de loi $\mathcal{N}_d(0, \mathrm{Id})$, sa projection orthogonale ΓN sur l'orthogonal de E a donc pour loi $\mathcal{N}_d(0, \Gamma)$. L'orthogonal de E formant un espace vectoriel de dimension $d - 1$, le théorème de Cochran montre que $\|\Gamma N\|^2$ suit la loi $\chi^2(d-1)$. Par suite,

$$\hat{D}_n(\mathbb{X}, P_0) = \|Z_n\|^2 \xrightarrow{\mathcal{L}} \chi^2(d-1),$$

d'où le théorème. □

Remarque. Le test du χ^2 d'ajustement peut être étendu au cas où l'on sait uniquement que la loi sous H_0 appartient à une famille de lois de référence paramétrées par $\theta \in \Theta \subset \mathbb{R}^k$. On se ramène alors au cas classique en estimant le paramètre inconnu sous H_0 par maximum de vraisemblance. On peut établir dans ce contexte un analogue du Théorème 9, avec une loi limite égale à $\chi^2(d-1-k)$.

7.2 Test du χ^2 d'indépendance

Dans cette section, $\mathscr{H}$ est le produit de deux ensembles finis, identifié à $\mathscr{H} = \mathscr{H}_1 \times \mathscr{H}_2$ avec $\mathscr{H}_1 = \{1, \dots, d\}$ et $\mathscr{H}_2 = \{1, \dots, \ell\}$. Considérons le modèle statistique $(\mathscr{H}^n, \{P^{\otimes n}\}_{P \in \mathscr{P}})$, où $\mathscr{P}$ désigne l'ensemble des probabilités sur $\mathscr{H}$ de support $\mathscr{H}$, i.e. $P(j,k) > 0 \ \forall (j,k) \in \mathscr{H}$. Ce modèle est paramétrique car chaque loi de $\mathscr{P}$ est à support dans l'ensemble fini $\mathscr{H}$.

Basé sur une observation $((X_1, Y_1), \dots, (X_n, Y_n)) \in \mathscr{H}^n$, un test d'indépendance précise si les extractions $\mathbb{X} = (X_1, \dots, X_n)$ et $\mathbb{Y} = (Y_1, \dots, Y_n)$ sont indépendantes. En notant $\mathscr{P}_{\text{indep}}$ l'ensemble des probabilités produits de $\mathscr{P}$ du type $P_1 \otimes P_2$, où P_1 et P_2 sont des lois marginales à support dans $\mathscr{H}_1$ et $\mathscr{H}_2$, le problème de test se formule de la manière suivante :

$$H_0 : P \in \mathscr{P}_{\text{indep}} \quad \text{contre} \quad H_1 : P \notin \mathscr{P}_{\text{indep}}.$$

Le test est fondé sur une comparaison asymptotique entre la loi empirique de l'échantillon $((X_1, Y_1), \dots, (X_n, Y_n))$ et le produit des lois empiriques des sous-échantillons $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$. Soient donc $\hat{P}_n$ la probabilité jointe

$$\hat{P}_n(j,k) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{j,k\}}(X_i, Y_i) \ \forall (j,k) \in \mathscr{H},$$

et $\hat{P}_{n,1}$, $\hat{P}_{n,2}$ les probabilités marginales

$$\hat{P}_{n,1}(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{j\}}(X_i) \text{ et } \hat{P}_{n,2}(k) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{k\}}(Y_i) \ \forall (j,k) \in \mathscr{H}.$$

La statistique de test compare les lois empiriques $\hat{P}_n$ et $\hat{P}_{n,1} \otimes \hat{P}_{n,2}$ de la manière suivante :

$$\hat{D}_n(\mathbb{X}, \mathbb{Y}) = n \sum_{j=1}^{d} \sum_{k=1}^{\ell} \frac{\left(\hat{P}_n(j,k) - \hat{P}_{n,1}(j)\hat{P}_{n,2}(k)\right)^2}{\hat{P}_n(j,k)},$$

avec la convention $0/0 = 0$.

Comme le montre le théorème ci-dessous, cette statistique est asymptotiquement libre sous H_0. Combinée avec le fait que la statistique $\hat{D}_n(\mathbb{X}, \mathbb{Y})$ compare les lois empiriques $\hat{P}_n$ et $\hat{P}_{n,1} \otimes \hat{P}_{n,2}$, cette propriété en fait un candidat pour la construction du test de H_0 contre H_1.

Théorème 10. *Soient $P \in \mathscr{P}$ et $((X_1, Y_1), \ldots, (X_n, Y_n)) \sim P^{\otimes n}$.*

(**i**) *Sous H_0, i.e. $P \in \mathscr{P}_{\text{indep}}$,*

$$\hat{D}_n(\mathbb{X}, \mathbb{Y}) \xrightarrow{\mathcal{L}} \chi^2(d-1)(\ell-1).$$

(**ii**) *Sous H_1, i.e. $P \notin \mathscr{P}_{\text{indep}}$,*

$$\hat{D}_n(\mathbb{X}, \mathbb{Y}) \xrightarrow{\mathbb{P}} +\infty.$$

Dans le problème de test de H_0 contre H_1, un test basé sur la statistique $\hat{D}_n$ est associé à une région de rejet de la forme $\{(x, y) \in \mathscr{H}_1^n \times \mathscr{H}_2^n : \hat{D}_n(x, y) \geq s\}$, car H_0 est rejetée lorsque $\hat{D}_n$ prend des valeurs anormalement grandes. Le test du χ^2 d'indépendance est le test asymptotique de région de rejet

$$R(\alpha) = \left\{(x, y) \in \mathscr{H}_1^n \times \mathscr{H}_2^n : \hat{D}_n(x, y) \geq \chi^2_{1-\alpha}(d-1)(\ell-1)\right\},$$

avec $\alpha \in]0, 1[$ et $\chi^2_{1-\alpha}(d-1)(\ell-1)$ le quantile d'ordre $(1-\alpha)$ de la loi $\chi^2(d-1)(\ell-1)$. Ce test est de niveau α puisque, d'après la propriété (**i**) du Théorème 10, pour chaque $P \in \mathscr{P}_{\text{indep}}$:

$$\lim_{n \to +\infty} P\left((\mathbb{X}, \mathbb{Y}) \in R(\alpha)\right) = \lim_{n \to +\infty} P\left(\hat{D}_n(\mathbb{X}, \mathbb{Y}) \geq \chi^2_{1-\alpha}(d-1)(\ell-1)\right) = \alpha.$$

Il est de plus convergent d'après l'assertion (**ii**) du même théorème. La procédure de test est donc définie ainsi : H_0 est rejetée au niveau asymptotique α lorsque $((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) \in R(\alpha)$.

Exemple. Pour savoir si, dans la population féminine, les couleurs des yeux et des cheveux sont indépendantes, on exploite une étude publiée en 1951 portant sur 2000 femmes parisiennes. Les données sont résumées dans le tableau suivant :

Cheveux \ Yeux	Clairs	Mixtes	Bruns	Total
Roux	14	9	8	31
Clairs	745	253	305	1303
Foncés	146	187	333	666
Total	905	449	646	2000

En attribuant un chiffre à chaque couleur et en considérant que les individus de l'étude sont indépendants, le modèle statistique associé à cette expérience est $(\mathscr{H}^n, \{P^{\otimes n}\}_{P\in\mathscr{P}})$, avec $\mathscr{H} = \{1,2,3\}\times\{1,2,3\}$, et le problème de test s'écrit $H_0 : P \in \mathscr{P}_{\text{indep}}$ contre $H_1 : P \notin \mathscr{P}_{\text{indep}}$. Pour donner une réponse à la question initiale, utilisons le test du χ^2 d'indépendance décrit plus haut, avec un niveau $\alpha = 5\%$. Le quantile d'ordre 95% de la loi $\chi^2(4)$ vaut environ 9.49, donc la région de rejet est

$$R(0.05) = \left\{(x,y) \in \{1,2,3\}^n \times \{1,2,3\}^n : \hat{D}_n(x,y) \geq 9.49\right\}.$$

Pour l'observation $((x_1,y_1),\dots,(x_n,y_n)) \in \mathscr{H}^n$ résumée dans le tableau ci-dessus, la statistique de test vaut $\hat{D}_n(x,y) = 233.28$ si $x = (x_1,\dots,x_n)$ et $y = (y_1,\dots,y_n)$. Par suite, $(x,y) \in R(0.05)$ donc H_0 est rejetée au niveau 5% : les couleurs des yeux et des cheveux sont liées.

La preuve de l'assertion (**ii**) du Théorème 10 est calquée sur celle de l'assertion (**ii**) du Théorème 9, en exhibant un couple $(j,k) \in \mathscr{H}$ tel que $P(j,k) \neq P_1(j)P_2(k)$, où P_1 et P_2 désignent les lois marginales selon la première et la seconde composante de $P \notin \mathscr{P}_{\text{indep}}$. En revanche, la démonstration complète de l'assertion (**i**) est longue et technique, et elle est admise dans le cadre de ce cours.

Le cas du test du χ^2 d'homogénéité. Il s'agit d'une situation particulière où l'une des deux variables (par exemple $\mathbb{X}$) ne peut plus raisonnablement être considérée comme aléatoire. Illustrons ceci par une étude sur des étudiants de l'UPMC, pour lesquels on cherche à savoir si le niveau d'étude (L3, M1 ou M2) est indépendant de l'opinion politique (gauche, centre, droite ou autre). Dans ce cas, si l'opinion politique peut être correctement modélisée par une variable aléatoire $\mathbb{Y}$ à quatre modalités, il n'en est pas de même du niveau d'étude à trois modalités $\mathbb{X}$, qui n'est pas aléatoire et nous est imposé par l'expérience (on connaît à l'avance les effectifs de L3,

M1 et M2). Dans ce contexte, le problème n'est donc pas de tester l'indépendance entre $\mathbb{X}$ et $\mathbb{Y}$, mais plutôt de savoir si le comportement politique des étudiants est homogène d'une année sur l'autre. Tout se passe en fait comme si l'on disposait de trois échantillons i.i.d., indépendants entre eux, issus des variables aléatoires $\mathbb{Y}_1$ (pour le L3), $\mathbb{Y}_2$ (pour le M1) et $\mathbb{Y}_3$ (pour le M2), la question posée étant désormais :"Ces trois échantillons ont-ils la même loi (H_0) ou pas (H_1) ?". On parle alors de test d'*homogénéité*, et il est facile de voir que le test du χ^2 d'indépendance s'adapte sans problème à ce contexte, avec exactement les mêmes formules. Choisir entre un test du χ^2 d'indépendance et d'homogénéité relève plus d'un problème d'interprétation de l'expérience (ou de protocole expérimental) que d'une question mathématique.

7.3 Test de Kolmogorov-Smirnov

L'objectif du test de Kolmogorov-Smirnov est de préciser, comme dans le test du χ^2 d'adéquation, si l'observation est issue d'une loi fixée par l'utilisateur. Cependant, leurs périmètres d'utilisation sont complémentaires car, dans le cas du test de Kolmogorov-Smirnov, l'espace des observations est nécessairement non fini.

Le modèle statistique de cette section est $(\mathbb{R}^n, \{P^{\otimes n}\}_{P\in\mathscr{P}})$, $\mathscr{P}$ désignant l'ensemble des probabilités sur $\mathbb{R}$ sans atomes, i.e. $P(x) = 0$ pour tout $x \in \mathbb{R}$ et $P \in \mathscr{P}$. Soit $P_0 \in \mathscr{P}$ fixé. Le problème de test s'énonce :

$$H_0 : P = P_0 \quad \text{contre} \quad H_1 : P \neq P_0.$$

L'enjeu est de trouver une statistique mesurant la proximité entre P_0 et une mesure empirique, et qui soit de surcroît libre, ou asymptotiquement libre, sous H_0. Dans cet esprit, une idée naturelle consiste à comparer la fonction de répartition de P_0 à la fonction de répartition d'un n-échantillon. Soit donc, pour chaque $\mathbb{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$, $F_{\mathbb{X},n}$ la fonction

$$F_{\mathbb{X},n}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(X_i) \ \forall t \in \mathbb{R}.$$

Il est instructif de noter que $F_{\mathbb{X},n}$, appelée *fonction de répartition empirique*, n'est autre que la fonction de répartition associée à la loi uniforme sur

l'ensemble $\{\mathbb{X}_1, \dots, \mathbb{X}_n\}$. La statistique de Kolmogorov-Smirnov compare P_0 à la probabilité associée à l'observation $\mathbb{X} = (X_1, \dots, X_n)$ par le biais de leurs fonctions de répartition. Elle est définie par

$$\hat{K}_n(\mathbb{X}, P_0) = \|F_{\mathbb{X},n} - F_0\|_\infty,$$

avec F_0 la fonction de répartition de la loi P_0, et $\|\cdot\|_\infty$ la norme uniforme, i.e.

$$\|f\|_\infty = \sup_{t \in \mathbb{R}} |f(t)|,$$

pour chaque fonction bornée $f : \mathbb{R} \to \mathbb{R}$.

Remarque. En pratique, la statistique de Kolmogorov-Smirnov est simple à calculer. Si $P \in \mathscr{P}$ et $\mathbb{X} = (X_1, \dots, X_n) \sim P^{\otimes n}$, notons $(X_{(1)}, \dots, X_{(n)})$ le vecteur des statistiques d'ordre associé à l'échantillon, c'est-à-dire le réarrangement des variables aléatoires $X_1, \dots, X_n$ tel que

$$X_{(1)} < \cdots < X_{(n)} \ \mathbb{P}\text{-p.s.}$$

Observer que ces inégalités sont strictes car P est sans atomes. Comme F_0 est continue et croissante, et comme la fonction

$$\mathbb{R} \ni t \mapsto F_{\mathbb{X},n}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(\mathbb{X}_i) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(\mathbb{X}_{(i)})$$

prend ses valeurs dans $\{k/n,\, k = 0, \dots, n\}$, en notant $X_{(0)} = 0$ et $X_{(n+1)} = +\infty$:

$$\begin{aligned} \hat{K}_n(\mathbb{X}, P_0) &= \max_{0 \le k \le n} \max_{t \in [X_{(k)}, X_{(k+1)}]} \left| \frac{k}{n} - F_0(t) \right| \\ &= \max_{1 \le k \le n} \max \left\{ \left| \frac{k}{n} - F_0(X_{(k)}) \right|, \left| \frac{k-1}{n} - F_0(X_{(k)}) \right| \right\}. \end{aligned}$$

Le calcul explicite de la statistique de Kolmogorov-Smirnov requiert donc au préalable un réarrangement des observations.

La première propriété importante de la statistique de Kolmogorov-Smirnov est énoncée dans le résultat ci-dessous.

Théorème 11. *Si $\mathbb{X} = (X_1, \dots, X_n) \sim P_0^{\otimes n}$, $\hat{K}_n(\mathbb{X}, P_0)$ est une statistique libre, i.e. sa loi ne dépend pas de P_0.*

Démonstration. Soit $F_0^{(-1)}$ le pseudo-inverse de F_0 défini pour chaque $s \in [0,1]$ par

$$F_0^{(-1)}(s) = \inf\{t \in \mathbb{R} : F_0(t) \geq s\}$$

(par convention, $\inf \mathbb{R} = -\infty$ et $\inf \emptyset = +\infty$). On constate que $F_0^{(-1)}(s) \leq t \Leftrightarrow s \leq F_0(t)$ et, puisque F_0 est continue car P_0 est sans atomes, $F_0 \circ F_0^{(-1)}(s) = s$ pour tout $s \in]0,1[$. Par suite (exercice), pour chaque $i = 1, \ldots, n$ et $s \in]0,1[$,

$$\mathbb{P}(F_0(X_i) \leq s) = \mathbb{P}\big(X_i \leq F_0^{(-1)}(s)\big) = F_0 \circ F_0^{(-1)}(s) = s,$$

les cas $s = 0$ et $s = 1$ donnant les valeurs 0 et 1. Ainsi, chaque variable aléatoire $U_i = F_0(X_i)$ suit la loi $\mathcal{U}([0,1])$. Comme F_0 est croissante, on obtient en prenant la notation $(U_{(1)}, \ldots, U_{(n)})$ de la remarque précédente pour désigner les statistiques d'ordre de $(U_1, \ldots, U_n)$:

$$(U_{(1)}, \ldots, U_{(n)}) = \big(F_0(X_{(1)}), \ldots, F_0(X_{(n)})\big).$$

Finalement, la représentation

$$\hat{K}_n(\mathbb{X}, P_0) = \max_{1 \leq k \leq n} \max \left\{ \left| \frac{k}{n} - F_0(X_{(k)}) \right|, \left| \frac{k-1}{n} - F_0(X_{(k)}) \right| \right\}$$

de la remarque montre que la statistique de Kolmogorov-Smirnov est libre. □

Nous pouvons utiliser cette propriété de liberté et calculer les quantiles de la loi commune, par exemple en simulant des réalisations de $\hat{K}_n(\mathbb{X}, P_0)$ lorsque P_0 est la loi uniforme sur $[0,1]$. Une fois connus ces quantiles universels $\zeta_{n,\alpha}$ pour $\alpha \in]0,1[$, le test de Kolmogorov-Smirnov de niveau α rejette H_0 si $\hat{K}_n(\mathbb{X}, P_0)$ prend des valeurs anormalement grandes, soit

$$T(\mathbb{X}) = \mathbb{1}_{[\hat{K}_n(\mathbb{X}, P_0) \geq \zeta_{n,1-\alpha}]}.$$

Si le Théorème 11 fournit une statistique de test pour le problème de test de H_0 contre H_1, il ne donne cependant aucune information concernant sa puissance. En revanche, une information de cette nature peut être fournie dans le cadre d'un test asymptotique car, si $\mathbb{X} = (X_1, \ldots, X_n) \sim P^{\otimes n}$ avec $P \in \mathscr{P}$ et $P \neq P_0$,

$$\sqrt{n}\, \hat{K}_n(\mathbb{X}, P_0) \xrightarrow{\mathbb{P}} +\infty.$$

Cette propriété s'obtient en invoquant la loi des grands nombres et la minoration

$$\sqrt{n}\,\hat{K}_n(\mathbb{X}, P_0) \geq \sqrt{n}\;|F_{\mathbb{X},n}(x) - F_0(x)|$$

avec $x \in \mathbb{R}$ tel que $F(x) \neq F_0(x)$, F désignant la fonction de répartition associée à la loi P. Par ailleurs, le théorème de Kolmogorov-Smirnov, que nous admettrons, stipule que lorsque $\mathbb{X} = (X_1, \ldots, X_n) \sim P_0^{\otimes n}$,

$$\sqrt{n}\,\hat{K}_n(\mathbb{X}, P_0) \xrightarrow{\mathcal{L}} \mu_{\mathrm{KS}},$$

où μ_{KS} est la loi de Kolmogorov-Smirnov, de fonction de répartition continue définie par

$$\mathbb{R}_+ \ni t \mapsto 1 - 2\sum_{\ell \geq 1} (-1)^{\ell+1} e^{-2\ell^2 t^2}.$$

Dans le cadre du problème de test de H_0 contre H_1, le test de Kolmogorov-Smirnov est le test asymptotique de région de rejet

$$R(\alpha) = \left\{ x \in \mathbb{R}^n : \sqrt{n}\,\hat{K}_n(x, P_0) \geq k_{1-\alpha} \right\},$$

avec $k_{1-\alpha}$ le quantile d'ordre $(1-\alpha)$ de la loi μ_{KS}, i.e.

$$2\sum_{\ell \geq 1} (-1)^{\ell+1} e^{-2\ell^2 k^2} = \alpha.$$

La forme de cette région de rejet est due au fait que H_0 est rejetée lorsque la statistique de test $\hat{K}_n(\mathbb{X}, P_0)$ prend des valeurs anormalement grandes. Ce test asymptotique est convergent et de niveau α, d'après les deux résultats de convergence mentionnés ci-dessus.

Chapitre 8

Apprentissage et classification supervisée

8.1 Objectifs

On considère un couple (X, Y) de variables aléatoires à valeurs dans $\mathbb{R}^d$ (pour X) et $\{0, 1\}$ (pour Y). La variable Y est appelée *label*, *classe* ou *étiquette*. Le problème de la classification supervisée (binaire) consiste à prédire au mieux Y à partir de X, c'est-à-dire à construire une fonction borélienne $g : \mathbb{R}^d \to \{0, 1\}$ (appelée *règle de décision* ou *de classification*) qui, à un x donné (réalisation de X) associe son label supposé 0 ou 1. Pour prendre un exemple, on peut penser à X comme une variable aléatoire représentant la fréquence d'un certain nombres de mots-clés dans un email, et à Y comme la variable associée exprimant le fait que l'email est sain (label 0) ou bien spam (label 1).

Remarques.

1. Dans ce cours, le label est supposé binaire pour simplifier. La théorie s'étend sans trop de difficultés au cas multi-labels, pour lequel Y prend ses valeurs dans un ensemble fini $\{1, \dots, M\}$.
2. Le terme "apprentissage statistique", dont les contours sont larges, se confond parfois avec la classification supervisée. Il peut aussi englober la régression bornée, où Y prend ses valeurs dans un ensemble compact, et la classification non supervisée, que nous aborderons dans le Chapitre 11.

La loi du couple (X, Y) est entièrement caractérisée par le couple (μ, r), où μ est la loi de X et r la fonction de régression de Y sur X. Plus précisément, pour tout $A \in \mathscr{B}(\mathbb{R}^d)$, $\mu(A) = \mathbb{P}(X \in A)$, et

$$r(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x),$$

la dernière égalité provenant du fait que Y prend ses valeurs dans $\{0, 1\}$.

Attention ! Dans ce modèle, Y n'est pas nécessairement lié à X de manière fonctionnelle, i.e. rien ne dit qu'il existe une fonction φ telle que $Y = \varphi(X)$. Pour s'en convaincre, il suffit de penser à l'exemple des emails, au sein duquel le mot "Viagra" peut être associé à un label spam ou non

Bien entendu, n'importe quelle fonction borélienne $g : \mathbb{R}^d \to \mathbb{R}$ fournit une règle de décision, et il est donc nécessaire d'adjoindre un critère de qualité à chaque décision. On remarque pour cela qu'une erreur de classification se produit lorsque $g(X) \neq Y$ et l'on définit naturellement la probabilité d'erreur d'une règle g par

$$L(g) = \mathbb{P}(g(X) \neq Y).$$

La quantité $L(g)$, qui mesure la pertinence de la règle g, permet donc de hiérarchiser les fonctions de décision agissant sur le couple (X, Y). Il est alors légitime de se poser la question de l'existence éventuelle d'une règle meilleure que les autres. Ce champion existe et s'appelle la *règle de Bayes* :

$$g^\star(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x) \\ 0 & \text{sinon.} \end{cases}$$

(Les égalités sont rompues en faveur de 0 par convention.) De façon équivalente,

$$g^\star(x) = \begin{cases} 1 & \text{si } r(x) > 1/2 \\ 0 & \text{sinon.} \end{cases}$$

Le lemme qui suit nous assure que $g^\star$ mérite bien son statut de champion :

Lemme 2. *Quelle que soit la règle de décision $g : \mathbb{R}^d \to \{0, 1\}$, on a*

$$L(g^\star) \leq L(g).$$

Démonstration. Soit $g : \mathbb{R}^d \to \{0, 1\}$ une fonction borélienne arbitraire. Comme

$$\mathbb{P}(g(X) \neq Y) = 1 - \mathbb{P}\left(g(X) = Y\right),$$

on a

$$\begin{aligned}\mathbb{P}(g(X) \neq Y) - \mathbb{P}(g^\star(X) \neq Y) &= \mathbb{P}\left(g^\star(X) = Y\right) - \mathbb{P}\left(g(X) = Y\right) \\ &= \mathbb{E}\left(\mathbb{P}(g^\star(X) = Y|X) - \mathbb{P}(g(X) = Y|X)\right) \\ &\geq 0,\end{aligned}$$

puisque

$$\begin{aligned}\mathbb{P}(g^\star(X) = Y|X) &= \mathbb{P}(g^\star(X) = 1, Y = 1|X) + \mathbb{P}(g^\star(X) = 0, Y = 0|X) \\ &= \mathbb{1}_{[g^\star(X)=1]}\mathbb{P}(Y = 1|X) + \mathbb{1}_{[g^\star(X)=0]}\mathbb{P}(Y = 0|X) \\ &= \max\left(\mathbb{P}(Y = 0|X), \mathbb{P}(Y = 1|X)\right),\end{aligned}$$

par définition de $g^\star$. Le résultat est donc démontré. □

$L^\star$ est appelée *probabilité d'erreur de Bayes* (*erreur de Bayes* ou *risque de Bayes*). On note en particulier que

$$L^\star = \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}(g(X) \neq Y),$$

où l'infimum est évalué sur toutes les fonctions de décision. Il est également instructif de remarquer que $L^\star = 0$ si et seulement si $Y = g^\star(X)$ $\mathbb{P}$-p.s., i.e. si et seulement si Y est une fonction borélienne de X. Dans le jargon de la classification supervisée, les probabilités $\mathbb{P}(Y = 0|X = x)$ et $\mathbb{P}(Y = 1|X = x)$ sont dites *probabilités a posteriori*.

Observons enfin que

$$\begin{aligned}L(g) &= 1 - \mathbb{P}(g(X) = Y) \\ &= 1 - \mathbb{E}\left(\mathbb{P}(g(X) = Y|X)\right) \\ &= 1 - \mathbb{E}\left[\mathbb{1}_{[g(X)=1]}r(X) + \mathbb{1}_{[g(X)=0]}\left(1 - r(X)\right)\right].\end{aligned}$$

En conséquence,

$$L^\star = 1 - \mathbb{E}\left[\mathbb{1}_{[r(X)>1/2]}r(X) + \mathbb{1}_{[r(X)\leq 1/2]}\left(1 - r(X)\right)\right].$$

Ceci montre que

$$L^\star = \mathbb{E}\left[\min\left(r(X), 1 - r(X)\right)\right] = \frac{1}{2} - \frac{1}{2}\mathbb{E}|2r(X) - 1|,$$

et nous fournit donc des écritures alternatives pour $L^\star$.

Le problème : La règle de classification optimale $g^\star$ dépend de la loi du couple (X, Y). Si cette loi est connue (ce qui est rarement le cas), le cours est terminé. Si elle ne l'est pas, $g^\star$ et $L^\star$ sont inaccessibles et il faut alors faire appel à un échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, de même loi que (X, Y), pour espérer récupérer de l'information sur ces deux quantités. C'est là que les choses deviennent intéressantes.

8.2 L'apprentissage

On suppose donc à partir de maintenant que l'on a accès à un n-échantillon i.i.d. (également appelé dans ce contexte *base de données* ou *base d'apprentissage*) formé de n couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ de variables aléatoires indépendantes entre elles, de même loi que (X, Y) et indépendantes de ce dernier couple. Pour abréger, on note $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$. C'est à partir de cet échantillon que l'on va s'attacher à construire une règle de décision $g_n(x) := g_n(x, \mathscr{D}_n)$ à valeurs dans $\{0, 1\}$ dont les performances se rapprochent de celles de la règle de Bayes $g^\star$. C'est le mécanisme d'apprentissage.

La qualité d'une règle g_n est mesurée par la probabilité d'erreur (conditionnelle)

$$L(g_n) = \mathbb{P}(g_n(X) \neq Y | \mathscr{D}_n).$$

Il convient de remarquer que, tout comme g_n, $L(g_n)$ est aléatoire par l'intermédiaire de $\mathscr{D}_n$. Le conditionnement par $\mathscr{D}_n$ dans la probabilité d'erreur permet de distinguer l'aléatoire provenant de l'échantillon de celui issu du couple générique (X, Y). On notera au passage que $\mathbb{E}L(g_n) = \mathbb{P}(g_n(X) \neq Y)$. (Pourquoi ?)

A partir de là, il est raisonnable de s'interroger sur le comportement de la probabilité d'erreur lorsque la taille de l'échantillon tend vers l'infini. On est en particulier en droit d'attendre d'une "bonne règle" que sa probabilité d'erreur se rapproche de $L^\star$ lorsque n croît. Comme $L(g_n)$ est aléatoire (contrairement à $L^\star$), il convient de bien préciser le sens des convergences. C'est l'objet de la définition qui suit.

Définition 29. *Une règle de classification g_n est convergente si $\mathbb{E}L(g_n) \to L^\star$. Elle est fortement convergente si $L_n \to L^\star$ $\mathbb{P}$-p.s.*

Comme $L(g_n) \geq L^\star$, on notera que la propriété $\mathbb{E}L(g_n) \to L^\star$ est équivalente à $L(g_n) \to L^\star$ dans L^1. On pourra aussi montrer, à titre d'exercise, que la convergence équivaut à la convergence en probabilité de $L(g_n)$ vers $L^\star$. On en déduit en particulier que si g_n est fortement convergente, elle est aussi convergente.

8.3 Minimisation du risque empirique

La minimisation du risque empirique fait partie des grands paradigmes de l'apprentissage statistique. Le principe général est le suivant. Donnons-nous un n-échantillon i.i.d. $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (et indépendant de) (X, Y), et une famille $\mathscr{G}$ de règles de décision candidates. On se pose alors le problème de choisir dans $\mathscr{G}$, en utilisant $\mathscr{D}_n$, une règle particulière $g_n^\star$ telle que $L(g_n^\star) = \mathbb{P}(g_n^\star(X) \neq Y | \mathscr{D}_n)$ soit proche de $\inf_{g \in \mathscr{G}} \mathbb{P}(g(X) \neq Y)$. En d'autres termes, on cherche à utiliser au mieux la base de données afin de sélectionner la meilleure technique de prévision possible au sein d'une collection $\mathscr{G}$ de règles fixée a priori. Il peut par exemple s'agir des règles linéaires (qui décident 0 ou 1 selon que l'on tombe d'un côté ou de l'autre d'un hyperplan), de règles polynomiales (qui décident 0 ou 1 en fonction du signe d'un polynôme), mais bien d'autres exemples sont possibles.

Afin d'atteindre cet objectif, une façon naturelle de procéder consiste à sélectionner dans $\mathscr{G}$ une règle $g_n^\star$ qui minimise l'*erreur empirique*

$$\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[g(X_i) \neq Y_i]}$$

parmi tous les éléments de $\mathscr{G}$, soit donc

$$g_n^\star \in \arg\min_{g \in \mathscr{G}} L_n(g).$$

En rappelant que $L(g_n^\star) = \mathbb{P}(g_n^\star(X) \neq Y | \mathscr{D}_n)$, on espère donc naturellement que $L(g_n^\star) \approx \inf_{g \in \mathscr{G}} L(g)$. Remarquons d'emblée que

$$L(g_n^\star) - L^\star = \left[L(g_n^\star) - \inf_{g \in \mathscr{G}} L(g) \right] + \left[\inf_{g \in \mathscr{G}} L(g) - L^\star \right].$$

Cette égalité, simple mais fondamentale, montre que l'erreur commise par $L(g_n^\star)$ en tant qu'estimateur de $L^\star$ se décompose en deux termes, respectivement appelés *erreur d'estimation* et *erreur d'approximation*. L'erreur d'estimation est aléatoire et reflète l'écart, en termes de probabilités, entre la règle sélectionnée et le champion local dans $\mathscr{G}$. L'erreur d'approximation est déterministe et mesure la proximité, toujours en termes de probabilités, entre la famille $\mathscr{G}$ et la règle optimale de Bayes.

Il est facile de voir que les deux termes d'erreur varient en sens inverse avec la taille de la classe $\mathscr{G}$, qui doit donc être suffisamment grande pour que l'erreur d'approximation soit petite, mais aussi suffisamment petite pour que l'erreur d'estimation soit contrôlée ! Pour s'en convaincre, il suffit d'envisager la situation extrême où $\mathscr{G}$ est constituée de toutes les fonctions mesurables de $\mathbb{R}^d$ dans $\{0,1\}$. Dans ce cas, l'erreur d'approximation est nulle, mais l'erreur d'estimation peut être importante, comme le montre le choix de la règle

$$g_n^\star(x) = \begin{cases} Y_i & \text{si } x = X_i,\ 1 \le i \le n \\ 0 & \text{sinon,} \end{cases}$$

dont le risque empirique est nul ! Ce phénomène indésirable, qui traduit une accroche trop importante aux données, est appelé *sur-apprentissage* ("overfitting" en anglais) et nous donnerons dans la suite des conditions précises sur $\mathscr{G}$ permettant de l'éviter. A partir de maintenant, nous supposons donc la classe $\mathscr{G}$ fixée une fois pour toutes et cherchons à contrôler le terme d'estimation.

Lemme 3. *On a, d'une part,*

(**i**) $L(g_n^\star) - \inf_{g\in\mathscr{G}} L(g) \le 2\sup_{g\in\mathscr{G}} |\hat{L}_n(g) - L(g)|$
et, d'autre part,

(**ii**) $|\hat{L}_n(g_n^\star) - L(g_n^\star)| \le \sup_{g\in\mathscr{G}} |\hat{L}_n(g) - L(g)|$.

Démonstration. On écrit

$$L(g_n^\star) - \inf_{g\in\mathscr{G}} L(g) \le \big|L(g_n^\star) - \hat{L}_n(g_n^\star)\big| + \big|\hat{L}_n(g_n^\star) - \inf_{g\in\mathscr{G}} L(g)\big|.$$

Clairement,

$$\big|L(g_n^\star) - \hat{L}_n(g_n^\star)\big| \le \sup_{g\in\mathscr{G}} \big|\hat{L}_n(g) - L(g)\big|,$$

et

$$\left|\hat{L}_n(g_n^\star) - \inf_{g\in\mathscr{G}} L(g)\right| = \left|\inf_{g\in\mathscr{G}} \hat{L}_n(g) - \inf_{g\in\mathscr{G}} L(g)\right| \leq \sup_{g\in\mathscr{G}} \left|\hat{L}_n(g) - L(g)\right|.$$

Cela prouve la première assertion. La preuve de la seconde est immédiate. □

Le Lemme 3 montre qu'en contrôlant la quantité $\sup_{g\in\mathscr{G}} |\hat{L}_n(g) - L(g)|$, on fait coup double, puisque l'on maîtrise non seulement la sous-optimalité de $g_n^\star$ dans $\mathscr{G}$ mais aussi l'erreur $|\hat{L}_n(g_n^\star) - L(g_n^\star)|$ commise lorsque $\hat{L}_n(g_n^\star)$ est utilisée pour estimer $L(g_n^\star)$, la véritable probabilité d'erreur de la règle sélectionnée. Il est donc désormais légitime de faire porter nos efforts sur l'analyse du terme $\sup_{g\in\mathscr{G}} |\hat{L}_n(g) - L(g)|$. Nous commençons simplement, en examinant le cas où la classe de fonctions $\mathscr{G}$ a un cardinal fini.

8.4 Cas d'une classe de cardinal fini

L'inégalité de Hoeffding (Théorème 2) montre que si Z désigne une variable aléatoire de loi binomiale $\mathscr{B}(n,p)$, alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{Z}{n} - p\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

En remarquant que pour une règle de décision fixée g, la quantité

$$n\hat{L}_n(g) = \sum_{i=1}^{n} \mathbb{1}_{[g(X_i)\neq Y_i]}$$

suit une loi $\mathscr{B}(n, L(g))$, on en conclut que

$$\mathbb{P}\left(\left|\hat{L}_n(g) - L(g)\right|\right) \leq 2e^{-2n\varepsilon^2},$$

ce qui conduit au premier résultat fondamental suivant :

Théorème 12. *Supposons que la classe $\mathscr{G}$ soit de cardinal fini majoré par N. Alors, pour tout $\varepsilon > 0$,*

$$\mathbb{P}\Big(\sup_{g\in\mathscr{G}} \left|\hat{L}_n(g) - L(g)\right| \geq \varepsilon\Big) \leq 2Ne^{-2n\varepsilon^2}. \tag{8.1}$$

Il faut noter que cette inégalité est déjà remarquable car la majoration de la probabilité est universelle, au sens où elle ne dépend pas de la loi du couple (X, Y). On en déduit en particulier, en utilisant le lemme de Borel-Cantelli, que

$$\sup_{g \in \mathcal{G}} \left| \hat{L}_n(g) - L(g) \right| \to 0 \ \mathbb{P}\text{-p.s.}$$

et donc, d'après le Lemme 3, que

$$L(g_n^\star) - \inf_{g \in \mathcal{G}} L(g) \to 0 \ \mathbb{P}\text{-p.s.}$$

Ce résultat signifie que pourvu que la classe $\mathcal{G}$ soit de cardinal fini, l'erreur d'estimation tend p.s. vers 0 lorsque n tend vers l'infini ; en d'autres termes, l'apprentissage est asymptotiquement optimal. Tout ceci s'étend sans difficulté au contrôle de l'espérance $\mathbb{E}(\sup_{g \in \mathcal{G}} |\hat{L}_n(g) - L(g)|)$, via le petit lemme technique suivant.

Lemme 4. *Soit Z une variable aléatoire à valeurs dans $\mathbb{R}_+$. Supposons qu'il existe une constante $C > 0$ telle que, pour tout $\varepsilon > 0$,*

$$\mathbb{P}(Z \geq \varepsilon) \leq C e^{-2n\varepsilon^2}.$$

Alors

$$\mathbb{E}Z \leq \sqrt{\frac{\ln(Ce)}{2n}}.$$

Démonstration. En partant de l'identité

$$\mathbb{E}Z^2 = \int_0^{+\infty} \mathbb{P}(Z^2 > \varepsilon) \mathrm{d}\varepsilon,$$

on a, pour tout $u \geq 0$,

$$\begin{aligned} \mathbb{E}Z^2 &= \int_0^u \mathbb{P}(Z^2 > \varepsilon) \mathrm{d}\varepsilon + \int_u^{+\infty} \mathbb{P}(Z^2 > \varepsilon) \mathrm{d}\varepsilon \\ &\leq u + C \int_u^{+\infty} e^{-2n\varepsilon} \mathrm{d}\varepsilon \\ &= u + \frac{C}{2n} e^{-2nu}. \end{aligned}$$

Avec le choix $u^\star = \frac{\ln C}{2n}$ (qui minimise la borne de droite), on en déduit que $\mathbb{E}Z^2 \leq \frac{\ln(Ce)}{2n}$, d'où le résultat par l'inégalité de Cauchy-Schwarz. □

Le lemme précédent, couplé à l'inégalité (8.1), montre que

$$\mathbb{E}\Big(\sup_{g\in\mathscr{G}}\big|\hat{L}_n(g)-L(g)\big|\Big)\leq\sqrt{\frac{\ln(2eN)}{2n}}.$$

Le Lemme 3 nous permet alors de conclure que

$$\mathbb{E}L(g_n^\star)-\inf_{g\in\mathscr{G}}L(g)\leq 2\sqrt{\frac{\ln(2eN)}{2n}},$$

ce qui montre que, pour une classe $\mathscr{G}$ de cardinal fini, l'espérance de l'erreur d'estimation reste sous contrôle (avec une borne plus ou moins grande selon sa taille N) et tend vers 0 à la vitesse $1/\sqrt{n}$ lorsque n tend vers l'infini.

Néanmoins, lorsque $\mathscr{G}$ n'est pas de cardinal fini (comme c'est le cas dans la plupart des problèmes intéressants), l'approche que nous venons de présenter ne fonctionne plus, et il faut trouver de nouveaux outils pour appréhender la "taille" de $\mathscr{G}$. C'est l'objet du chapitre suivant, qui présente la théorie de Vapnik-Chervonenkis.

Chapitre 9

Théorie de Vapnik-Chervonenkis

9.1 Passage du $\sup_{g\in\mathscr{G}}$ au $\sup_{A\in\mathscr{A}}$

Etant donné un n-échantillon i.i.d. $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (et indépendant de) (X, Y), et une famille $\mathscr{G}$ de règles de décision candidates, le chapitre précédent a montré le rôle essentiel joué par le terme $\sup_{g\in\mathscr{G}} |\hat{L}_n(g) - L(g)|$, qu'il faut donc apprendre à contrôler avec la plus grande généralité possible.

Désignons par ν la loi du couple (X, Y) et par ν_n la mesure empirique associée à $\mathscr{D}_n$, i.e., pour tout $A \in \mathscr{B}(\mathbb{R}^d \times \{0, 1\})$,

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[(X_i,Y_i)\in A]}.$$

A une règle de décision quelconque $g \in \mathscr{G}$, nous pouvons associer le borélien

$$A_g = \Big\{(x, y) \in \mathbb{R}^d \times \{0, 1\} : g(x) \neq y\Big\}.$$

En utilisant cette notation, il est alors facile de voir que, d'une part,

$$L(g) = \mathbb{P}(g(X) \neq Y) = \nu(A_g)$$

et, d'autre part,

$$\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[g(X_i)\neq Y_i]} = \nu_n(A_g).$$

On constate ainsi que

$$\left\{ \sup_{g\in\mathscr{G}} \left|\hat{L}_n(g) - L(g)\right| \right\} = \left\{ \sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \right\},$$

où, par définition, $\mathscr{A} = \{A_g : g \in \mathscr{G}\}$. Ce jeu d'écriture nous montre donc que pour analyser le comportement probabiliste du terme $\sup_{g\in\mathscr{G}} |\hat{L}_n(g) - L(g)|$, il faut avant tout comprendre comment se comporte la déviation maximale de la mesure empirique par rapport à la vraie mesure sur une classe d'ensembles mesurables $\mathscr{A}$ donnée. On peut d'ores et déjà observer que, pour un ensemble A fixé,

$$|\nu_n(A) - \nu(A)| \to 0 \ \mathbb{P}\text{-p.s.}$$

d'après la loi des grands nombres. D'autre part, si le cardinal de $\mathscr{A}$ est fini et majoré par N, un raisonnement similaire à celui du Théorème 12 nous apprend que, pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \geq \varepsilon \Big) \leq 2Ne^{-2n\varepsilon^2}, \tag{9.1}$$

d'où l'on déduit (lemme de Borel-Cantelli) que, pour toute loi ν,

$$\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \to 0 \ \mathbb{P}\text{-p.s.}$$

En revanche, si la classe $\mathscr{A}$ est trop massive, des catastrophes peuvent se produire. On s'en convaincra facilement en remarquant que si $\mathscr{A}$ désigne l'ensemble de tous les boréliens de $\mathbb{R}^d \times \{0,1\}$, alors on peut trouver des lois ν (un exemple ?) telles que

$$\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| = 1 \ \mathbb{P}\text{-p.s.}$$

La conclusion de tout ceci c'est qu'il faut arriver, d'une manière ou d'une autre, à contrôler la "taille" de la classe d'ensembles $\mathscr{A}$. Pour atteindre cet objectif, il convient au préalable d'introduire quelques outils combinatoires nouveaux.

9.2 Théorème de Vapnik-Chervonenkis

Soit $\mathscr{A}$ une famille de sous-ensembles de $\mathbb{R}^p$, de cardinal (pas nécessairement fini) strictement supérieur à 1 (cette hypothèse sera implicite dans

la suite). Etant donné n points $z_1, \ldots, z_n$ de $\mathbb{R}^p$, on définit la quantité $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n)$ par

$$\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n) = \big| \{\{z_1, \ldots, z_n\} \cap A : A \in \mathscr{A}\} \big|.$$

En d'autres termes, $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n)$ représente le nombre de sous-ensembles de $\{z_1, \ldots, z_n\}$ que l'on peut obtenir en intersectant ces n points par les ensembles de $\mathscr{A}$. Bien entendu, on a toujours $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n) \leq 2^n$, et lorsque $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n) = 2^n$, on dit que la classe $\mathscr{A}$ *pulvérise* l'ensemble $\{z_1, \ldots, z_n\}$. Afin de ne pas être gêné par le choix arbitraire de $z_1, \ldots, z_n$, on pose

$$\mathbf{S}_{\mathscr{A}}(n) = \max_{(z_1,\ldots,z_n) \in \mathbb{R}^{pn}} \mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n)$$

et on appelle cet indice le *coefficient de pulvérisation* de n points par la classe $\mathscr{A}$.

Clairement, $\mathbf{S}_{\mathscr{A}}(n) \leq 2^n$. D'autre part, $\mathbf{S}_{\mathscr{A}}(1) = 2$ (pourquoi ?), et si l'on a $\mathbf{S}_{\mathscr{A}}(k) < 2^k$ pour un certain entier $k > 1$ alors $\mathbf{S}_{\mathscr{A}}(n) < 2^n$ pour tout $n \geq k$ (pourquoi ?). Il est donc naturel de s'interroger sur l'existence d'un plus grand entier n tel que $\mathbf{S}_{\mathscr{A}}(n) = 2^n$. C'est l'objet de la définition suivante.

Définition 30. *Soit $\mathscr{A}$ une famille de sous-ensembles de $\mathbb{R}^p$. On appelle dimension de Vapnik-Chervonenkis de $\mathscr{A}$, notée $V_{\mathscr{A}}$, le plus grand entier $n_0 \geq 1$ tel que $\mathbf{S}_{\mathscr{A}}(n_0) = 2^{n_0}$. Si $\mathbf{S}_{\mathscr{A}}(n) = 2^n$ pour tout $n \geq 1$, on pose $V_{\mathscr{A}} = +\infty$.*

La dimension de Vapnik-Chervonenkis mesure, en un certain sens, la "taille" (la "dimension") de la famille $\mathscr{A}$ et généralise ainsi la notion de cardinal. Il s'agit d'un concept combinatoire important qui, comme nous le verrons dans la suite, joue un rôle clé dans la théorie de l'apprentissage statistique. Examinons auparavant quelques exemples (les preuves sont de difficultés variées et laissées au lecteur).

Exemples.

1. Supposons $|\mathscr{A}| < \infty$. Bien entendu, $\mathbf{S}_{\mathscr{A}}(n) \leq |\mathscr{A}|$. D'autre part, par définition de $V_{\mathscr{A}}$, $\mathbf{S}_{\mathscr{A}}(V_{\mathscr{A}}) = 2^{V_{\mathscr{A}}}$, d'où l'on déduit que
$$V_{\mathscr{A}} \leq \ln_2 |\mathscr{A}|.$$
2. En dimension $p = 1$. Si $\mathscr{A} = \{]-\infty, a] : a \in \mathbb{R}\}$, alors $V_{\mathscr{A}} = 1$. Si $\mathscr{A} = \{[a, b] : (a, b) \in \mathbb{R}^2\}$, alors $V_{\mathscr{A}} = 2$. On remarquera que dans le premier cas $\mathbf{S}_{\mathscr{A}}(n) = n + 1$, tandis que dans le second $\mathbf{S}_{\mathscr{A}}(n) = \frac{n(n+1)}{2} + 1$.

3. En dimension p quelconque. Si
$$\mathscr{A} = \left\{]-\infty, a_1] \times \cdots \times]-\infty, a_p] : (a_1, \ldots, a_p) \in \mathbb{R}^d \right\},$$
alors $V_{\mathscr{A}} = d$. Si $\mathscr{A} = \{\text{rectangles de } \mathbb{R}^p\}$, alors $V_{\mathscr{A}} = 2d$. En revanche, pour $\mathscr{A} = \{\text{convexes de } \mathbb{R}^p\}$, on a $V_{\mathscr{A}} = +\infty$.
4. (**Important.**) Soit $\mathscr{G}$ un espace vectoriel de fonctions de $\mathbb{R}^p \to \mathbb{R}$, de dimension finie $\dim \mathscr{G}$. Alors, si
$$\mathscr{A} = \big\{ \{x \in \mathbb{R}^p : g(x) \geq 0\} : g \in \mathscr{G} \big\},$$
on a $V_{\mathscr{A}} \leq \dim \mathscr{G}$. En particulier, si $\mathscr{A}$ désigne la famille des 1/2-espaces linéaires, i.e. les sous-ensembles de $\mathbb{R}^p$ de la forme $\{x \in \mathbb{R}^p : a^\top x + b \geq 0 : a \in \mathbb{R}^p, b \in \mathbb{R}\}$, il vient $V_{\mathscr{A}} \leq p + 1$.

Nous sommes désormais équipés pour énoncer le théorème fondamental suivant, appelé *théorème de Vapnik-Chervonenkis*.

Théorème 13 (Vapnik-Chervonenkis)**.** *Soient $Z_1, \ldots, Z_n$ des variables aléatoires indépendantes, de même loi ν sur $\mathbb{R}^p$, et soit ν_n la mesure empirique correspondante. Alors, pour toute famille borélienne $\mathscr{A} \subset \mathbb{R}^p$ et pour tout $\varepsilon > 0$, on a*
$$\mathbb{P}\left(\sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \right) \leq 8 \mathbf{S}_{\mathscr{A}}(n) e^{-n\varepsilon^2/32}.$$

Avant de prouver ce théorème, il convient de souligner quelques points essentiels.

1. La borne est universelle, dans le sens où elle ne dépend pas de la loi particulière ν.
2. Ce résultat généralise l'inégalité (9.1) qui n'était valable que pour une classe $\mathscr{A}$ de cardinal fini. Grosso modo, le cardinal de $\mathscr{A}$ est remplacé par le coefficient de pulvérisation.
3. D'après le lemme de Borel-Cantelli, il s'ensuit que
$$\sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| \to 0 \ \mathbb{P}\text{-p.s.}$$
dès que la série de terme général $\mathbf{S}_{\mathscr{A}}(n) e^{-n\varepsilon^2/32}$ est sommable. C'est par exemple le cas si $|\mathscr{A}| < \infty$ ou si $\mathbf{S}_{\mathscr{A}}(n)$ est un polynôme en n. En revanche, il est impossible de conclure si $\mathbf{S}_{\mathscr{A}}(n) = 2^n$ pour tout n (ou, c'est équivalent, si $V_{\mathscr{A}} = +\infty$).

4. La preuve du Théorème 13 n'est pas compliquée et repose sur quelques arguments clés que l'on rencontre fréquemment en théorie de l'apprentissage. En un mot, le principe consiste à faire sortir le supremum de la parenthèse pour le placer devant la probabilité. Ce grand saut est effectué en jouant sur les propriétés combinatoires de la classe $\mathscr{A}$ telles que décrites par $\mathbf{S}_{\mathscr{A}}(n)$.

Démonstration du Théorème 13. Dans toute la preuve, on suppose $\varepsilon > 0$ fixé et on choisit n assez grand de sorte que $n\varepsilon^2 \geq 2$. Dans le cas contraire, il est facile de voir que le résultat annoncé est correct car la borne du théorème est alors plus grande que 1. La preuve s'organise en 4 étapes.

Etape 1 : Symmétrisation. En sus du n-échantillon i.i.d. original $Z_1, \ldots, Z_n$, on considère un second échantillon i.i.d. $Z'_1, \ldots, Z'_n$ de la loi ν, indépendant du premier. On note ν_n la mesure empirique relative à $Z_1, \ldots, Z_n$ et ν'_n celle relative à $Z'_1, \ldots, Z'_n$. La première étape consiste à montrer que

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \Big) \leq 2\,\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \varepsilon/2 \Big).$$

Choisissons pour cela un ensemble aléatoire $A^\star \in \mathscr{A}$ (dépendant de l'échantillon initial $Z_1, \ldots, Z_n$) tel que $|\nu_n(A^\star) - \nu(A^\star)| > \varepsilon$ (si un tel ensemble n'existe pas, on prend $A^\star = \mathbb{R}^p$). On a alors

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \varepsilon/2 \Big) \\
&\quad = \mathbb{E}\bigg(\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \varepsilon/2 \,\Big|\, Z_1, \ldots, Z_n \Big) \bigg) \\
&\quad \geq \mathbb{E}\bigg(\mathbb{P}\Big(\big|\nu_n(A^\star) - \nu'_n(A^\star)\big| > \varepsilon/2 \,\Big|\, Z_1, \ldots, Z_n \Big) \bigg) \\
&\quad = \mathbb{P}\Big(\big|\nu_n(A^\star) - \nu'_n(A^\star)\big| > \varepsilon/2 \Big).
\end{aligned}$$

On en déduit, en utilisant l'inégalité triangulaire, que

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \varepsilon/2 \Big) \\
&\quad \geq \mathbb{P}\Big(|\nu_n(A^\star) - \nu(A^\star)| > \varepsilon, \big|\nu'_n(A^\star) - \nu(A^\star)\big| < \varepsilon/2 \Big) \\
&\quad = \mathbb{E}\bigg(\mathbb{1}_{[|\nu_n(A^\star)-\nu(A^\star)|>\varepsilon]} \mathbb{P}\Big(\big|\nu'_n(A^\star) - \nu(A^\star)\big| < \varepsilon/2 \,\Big|\, Z_1, \ldots, Z_n \Big) \bigg).
\end{aligned}$$

L'inégalité de Bienaymé-Tchebytchev montre que

$$\begin{aligned}\mathbb{P}\Big(\big|\nu'_n(A^\star) - \nu(A^\star)\big| < \varepsilon/2 \,\Big|\, Z_1, \dots, Z_n \Big) \\ \geq 1 - \frac{\mathbb{E}\Big(\big(\nu'_n(A^\star) - \nu(A^\star)\big)^2 \,\Big|\, Z_1, \dots, Z_n \Big)}{\varepsilon^2/4}.\end{aligned}$$

En observant que, conditionnellement à $Z_1, \dots, Z_n$, $n\nu'_n(A^\star)$ suit une loi $\mathscr{B}(n, \nu(A^\star))$, on en déduit en particulier que

$$\begin{aligned}\mathbb{P}\Big(\big|\nu'_n(A^\star) - \nu(A^\star)\big| < \varepsilon/2 \,\Big|\, Z_1, \dots, Z_n \Big) &\geq 1 - \frac{\mathbb{V}\big(\nu'_n(A^\star) \big| Z_1, \dots, Z_n\big)}{\varepsilon^2/4} \\ &= 1 - \frac{\nu(A^\star)(1-\nu(A^\star))}{n\varepsilon^2/4} \\ &\geq 1 - \frac{1}{n\varepsilon^2} \\ &\quad (\text{car } \sup_{u\in[0,1]} u(1-u) = 1/4).\end{aligned}$$

Ainsi,

$$\begin{aligned}\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \varepsilon/2 \Big) &\geq \mathbb{E}\left(\mathbb{1}_{[|\nu_n(A^\star) - \nu(A^\star)| > \varepsilon]} \Big(1 - \frac{1}{n\varepsilon^2}\Big) \right) \\ &\geq \frac{1}{2}\,\mathbb{P}\Big(|\nu_n(A^\star) - \nu(A^\star)| > \varepsilon \Big) \\ &\quad (\text{car } n\varepsilon^2 \geq 2) \\ &= \frac{1}{2}\,\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \Big).\end{aligned}$$

On en conclut bien que

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \Big) \leq 2\,\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \varepsilon/2 \Big).$$

Etape 2 : Signes aléatoires. On se donne maintenant n variables aléatoires $\sigma_1, \dots, \sigma_n$, indépendantes et chacune de loi de Rademacher, vérifiant $\mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = +1) = 1/2$. On suppose en outre que les variables $\sigma_1, \dots, \sigma_n$ sont indépendantes de $Z_1, \dots, Z_n, Z'_1, \dots, Z'_n$. Il est alors facile de voir (exercice) que

$$n \sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| = \sup_{A\in\mathscr{A}} \left| \sum_{i=1}^n \big(\mathbb{1}_A(Z_i) - \mathbb{1}_A(Z'_i)\big) \right|$$

a même loi que

$$\sup_{A\in\mathscr{A}} \Big| \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_A(Z_i) - \mathbb{1}_A(Z_i') \right) \Big|.$$

Dès lors, en utilisant le résultat de la première étape, nous pouvons écrire que

$$\begin{aligned} &\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \Big) \\ &\quad \le 2\,\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_A(Z_i) - \mathbb{1}_A(Z_i') \right) \Big| > \varepsilon/2 \Big), \end{aligned}$$

et donc

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \Big) \le 4\,\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i) \Big| > \varepsilon/4 \Big).$$

Etape 3 : Le saut du sup. En poursuivant le calcul précédent, on a

$$\begin{aligned} &\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \Big) \\ &\quad \le 4\mathbb{E}\left(\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i) \Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n \Big) \right). \end{aligned}$$

Or,

$$\begin{aligned} &\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i) \Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n \Big) \\ &\quad \le \mathbb{P}\Big(\exists A \in \mathscr{A} : \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i) \Big| > \varepsilon/4 \Big| Z_1, \ldots, Z_n \Big). \end{aligned}$$

Une fois fixés les points $z_1, \ldots, z_n$, le vecteur $(\mathbb{1}_A(z_1), \ldots, \mathbb{1}_A(z_n))$ prend $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n)$ valeurs distinctes lorsque A varie dans $\mathscr{A}$, soit donc un maximum de $\mathbf{S}_{\mathscr{A}}(n)$ valeurs. Du coup,

$$\begin{aligned} &\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i) \Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n \Big) \\ &\quad \le \mathbb{P}\Big(\exists A \in \mathscr{A}_0 : \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i) \Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n \Big), \end{aligned}$$

où $\mathscr{A}_0$ est un ensemble fini (fonction de $Z_1, \ldots, Z_n$) de cardinal au plus $\mathbf{S}_{\mathscr{A}}(n)$. Il s'ensuit que

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \frac{1}{n}\Big| \sum_{i=1}^n \sigma_i \mathbb{1}_A(Z_i)\Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n\Big) \\
&\quad \le \sum_{A\in\mathscr{A}_0} \mathbb{P}\Big(\frac{1}{n}\Big| \sum_{i=1}^n \sigma_i \mathbb{1}_A(Z_i)\Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n\Big) \\
&\quad \le \mathbf{S}_{\mathscr{A}}(n) \sup_{A\in\mathscr{A}} \mathbb{P}\Big(\frac{1}{n}\Big| \sum_{i=1}^n \sigma_i \mathbb{1}_A(Z_i)\Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n\Big).
\end{aligned}$$

On notera ici le saut du sup de l'intérieur vers l'extérieur de la probabilité, accompli grâce à l'introduction des signes aléatoires et du coefficient de pulvérisation. Ainsi,

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon\Big) \\
&\quad \le 4\mathbf{S}_{\mathscr{A}}(n)\mathbb{E}\Big(\sup_{A\in\mathscr{A}} \mathbb{P}\Big(\frac{1}{n}\Big| \sum_{i=1}^n \sigma_i \mathbb{1}_A(Z_i)\Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n\Big)\Big). \qquad (9.2)
\end{aligned}$$

Etape 4 : Inégalité de Hoeffding et conclusion. Conditionnellement à $Z_1, \ldots, Z_n$, la variable aléatoire $\sum_{i=1}^n \sigma_i \mathbb{1}_A(Z_i)$ est la somme de n variables aléatoires indépendantes, centrées et comprises entre -1 et 1. Ainsi, d'après l'inégalité de Hoeffding,

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{n}\Big| \sum_{i=1}^n \sigma_i \mathbb{1}_A(Z_i)\Big| > \varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n\Big) \\
&\quad = \mathbb{P}\Big(\Big| \sum_{i=1}^n \sigma_i \mathbb{1}_A(Z_i)\Big| > n\varepsilon/4 \,\Big|\, Z_1, \ldots, Z_n\Big) \\
&\quad \le 2e^{-n\varepsilon^2/32}.
\end{aligned}$$

On conclut alors en utilisant (9.2) que

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon\Big) \le 8\mathbf{S}_{\mathscr{A}}(n) e^{-n\varepsilon^2/32},$$

ce qui est bien le résultat annoncé. □

Application importante. Plaçons-nous sur la droite réelle et considérons un n-échantillon $Z_1, \ldots, Z_n$ de variables aléatoires i.i.d., de loi commune ν. En prenant $\mathscr{A} = \{]-\infty, z] : z \in \mathbb{R}\}$, il est facile de voir que, pour tout $A =]-\infty, z] \in \mathscr{A}$, on a $\nu(A) = F(z)$ et $\nu_n(A) = F_n(z)$, où F (respectivement, F_n) est la fonction de répartition associée à la loi ν (respectivement, la fonction de répartition empirique associée à $Z_1, \ldots, Z_n$). D'autre part, une analyse rapide indique que $\mathbf{S}_{\mathscr{A}}(n) = n + 1$. Ainsi, en utilisant le théorème de Vapnik-Chervonenkis, on montre que

$$\begin{aligned}\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon\Big) &= \mathbb{P}\Big(\sup_{z\in\mathbb{R}} |F_n(z) - F(z)| > \varepsilon\Big) \\ &\leq 8(n+1)e^{-n\varepsilon^2/32}.\end{aligned}$$

Le lemme de Borel-Cantelli implique alors que

$$\sup_{z\in\mathbb{R}} |F_n(z) - F(z)| \to 0 \ \mathbb{P}\text{-p.s.},$$

c'est-à-dire que la fonction de répartition empirique converge presque sûrement vers la fonction de répartition, au sens de la convergence uniforme des fonctions. Ce résultat remarquable porte le nom de *théorème de Glivenko-Cantelli*.

Avant de tirer les conséquences du Théorème 13 pour la théorie de l'apprentissage, il convient de préciser quelques propriétés élémentaires de la dimension de Vapnik-Chervonenkis.

9.3 Aspects combinatoires

Nous admettrons le résultat combinatoire suivant, connu sous le nom de *lemme de Sauer* :

Théorème 14 (Lemme de Sauer)**.** *Soit $\mathscr{A}$ une famille d'ensembles admettant une dimension de Vapnik-Chervonenkis finie $V_{\mathscr{A}}$. Alors, pour tout $n \geq 1$,*

$$\mathbf{S}_{\mathscr{A}}(n) \leq \sum_{i=1}^{V_{\mathscr{A}}} \binom{n}{i}.$$

Dans la suite, c'est surtout le corollaire suivant qui nous sera utile :

Corollaire 1. *Soit $\mathscr{A}$ une famille d'ensembles admettant une dimension de Vapnik-Chervonenkis finie $V_{\mathscr{A}}$. Alors, pour tout $n \geq 1$,*

$$\mathbf{S}_{\mathscr{A}}(n) \leq (n+1)^{V_{\mathscr{A}}}.$$

Démonstration. On a

$$\begin{aligned}(n+1)^{V_{\mathscr{A}}} = \sum_{i=0}^{V_{\mathscr{A}}} \binom{V_{\mathscr{A}}}{i} n^i &= \sum_{i=0}^{V_{\mathscr{A}}} \frac{n^i V_{\mathscr{A}}!}{i!(V_{\mathscr{A}}-i)!} \\ &\geq \sum_{i=0}^{V_{\mathscr{A}}} \frac{n^i}{i!} \\ &\geq \sum_{i=0}^{V_{\mathscr{A}}} \binom{n}{i} \geq \mathbf{S}_{\mathscr{A}}(n).\end{aligned}$$

□

On déduit en particulier du corollaire précédent qu'un coefficient de pulvérisation tombe forcément dans l'une des deux catégories suivantes :

- ▷ Ou bien $\mathbf{S}_{\mathscr{A}}(n) = 2^n$ pour tout $n \geq 1$ (dans ce cas, $V_{\mathscr{A}} = +\infty$).
- ▷ Ou bien $\mathbf{S}_{\mathscr{A}}(n) \leq (n+1)^{V_{\mathscr{A}}}$ (dans ce cas, $V_{\mathscr{A}} < \infty$).

On ne peut donc **jamais** avoir des situations intermédiaires, comme par exemple $\mathbf{S}_{\mathscr{A}}(n) \sim 2^{\sqrt{n}}$.

Enfin, en combinant le théorème de Vapnik-Chervonenkis, le Lemme technique 4 et le Corollaire 1, on conclut que pour toute famille d'ensembles mesurables $\mathscr{A} \subset \mathbb{R}^p$ admettant une dimension de Vapnik-Chervonenkis finie $V_{\mathscr{A}}$,

$$\begin{aligned}\mathbb{E}\Big(\sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| \Big) &\leq 8\sqrt{\frac{\ln\left(8e\mathbf{S}_{\mathscr{A}}(n)\right)}{2n}} \\ &\leq 8\sqrt{\frac{V_{\mathscr{A}} \ln(n+1) + 4}{2n}} \\ &= \mathrm{O}\left(\sqrt{\frac{V_{\mathscr{A}} \log n}{n}}\right).\end{aligned}$$

Il est à noter qu'il est possible de se débarasser du terme logarithmique en utilisant des techniques dites *de chaînage*, dont la présentation dépasserait le cadre de ce cours.

9.4 Application à la minimisation du risque empirique

Nous pouvons à présent peaufiner les bornes sur l'erreur d'estimation dans le problème de classification supervisée. Rappelons, pour mémoire, que l'on considère un n-échantillon i.i.d. $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (et indépendant de) (X, Y), et une famille $\mathscr{G}$ de règles de décision candidates. En désignant par $g_n^\star$ un minimiseur du risque empirique dans $\mathscr{G}$, nous savons que, d'une part,

$$L(g_n^\star) - \inf_{g \in \mathscr{G}} L(g) \leq 2 \sup_{g \in \mathscr{G}} \left| \hat{L}_n(g) - L(g) \right|$$

et, d'autre part,

$$\left\{ \sup_{g \in \mathscr{G}} \left| \hat{L}_n(g) - L(g) \right| \right\} = \left\{ \sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| \right\},$$

où, par définition, $\mathscr{A} = \{A_g : g \in \mathscr{G}\}$, avec

$$A_g = \Big\{ (x, y) \in \mathbb{R}^d \times \{0, 1\} : g(x) \neq y \Big\}.$$

Il est alors clair, de par le théorème de Vapnik-Chervonenkis, que le coefficient de pulvérisation $\mathbf{S}_{\mathscr{A}}(n)$ va jouer un rôle fondamental dans le contrôle du terme $\sup_{g \in \mathscr{G}} |\hat{L}_n(g) - L(g)|$. Néanmoins, la classe $\mathscr{A}$, composée de sous-ensembles de $\mathbb{R}^d \times \{0, 1\}$, revêt une structure un peu complexe qui ne se prête pas bien à l'analyse combinatoire. Fort heureusement, les choses se simplifient grâce à la proposition suivante, dont la preuve est laissée à titre d'exercice.

Proposition 5. *Soit $\mathscr{A} = \{A_g : g \in \mathscr{G}\}$ et $\bar{\mathscr{A}} = \{\{x \in \mathbb{R}^d : g(x) = 1\} : g \in \mathscr{G}\}$. Alors, pour tout $n \geq 1$, $\mathbf{S}_{\bar{\mathscr{A}}}(n) = \mathbf{S}_{\mathscr{A}}(n)$. En particulier, $V_{\bar{\mathscr{A}}} = V_{\mathscr{A}}$.*

Nous sommes désormais en mesure d'énoncer le principal résultat de ce chapitre, dont la preuve découle du Lemme 3, du théorème de Vapnik-Chervonenkis (et du Lemme technique 4 pour la seconde assertion).

Théorème 15. *On a, pour tout $n \geq 1$,*

$$\mathbb{P}\Big(\left| L(g_n^\star) - \inf_{g \in \mathscr{G}} L(g) \right| > \varepsilon \Big) \leq 8 \mathbf{S}_{\bar{\mathscr{A}}}(n) e^{-n\varepsilon^2/128}.$$

En outre,

$$\mathbb{E}L(g_n^\star) - \inf_{g\in\mathscr{G}} L(g) \leq 16\sqrt{\frac{\ln\left(8e\mathbf{S}_{\mathscr{A}}(n)\right)}{2n}}.$$

D'après le lemme de Borel-Cantelli, il suit de ce résultat que

$$L(g_n^\star) - \inf_{g\in\mathscr{G}} L(g) \to 0 \ \mathbb{P}\text{-p.s.}$$

dès que la série de terme général $\mathbf{S}_{\mathscr{A}}(n)e^{-n\varepsilon^2/128}$ est sommable. Or, d'après le Corollaire 1, c'est précisément le cas dès que $V_{\mathscr{A}}$ (ou $V_{\mathscr{A}}$) est finie puisqu'alors $\mathbf{S}_{\mathscr{A}}(n)$ a une croissance au plus polynomiale en n. On retiendra donc de tout ceci que la condition $V_{\mathscr{A}} < \infty$ est suffisante pour assurer la convergence presque sûre du terme d'estimation vers 0 et que, dans ce cas,

$$\mathbb{E}L(g_n^\star) - \inf_{g\in\mathscr{G}} L(g) = \mathrm{O}\left(\sqrt{\frac{V_{\mathscr{A}}\ln n}{n}}\right).$$

Exemples.

1. **Classification linéaire.** En notant $x = (x^{(1)}, \ldots, x^{(d)})$, on considère des règles très simples, de la forme

$$g(x) = \begin{cases} 1 & \text{si } \sum_{j=1}^d a_j x^{(j)} + a_0 > 0 \\ 0 & \text{sinon,} \end{cases}$$

où $(a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1}$ est un paramètre vectoriel. Dans ce cas,

$$\mathscr{A} \subset \left\{ \{x \in \mathbb{R}^d : a^\top x + a_0 \geq 0\} : a \in \mathbb{R}^d, a_0 \in \mathbb{R} \right\}$$

et, d'après les propriétés de la dimension de Vapnik-Chervonenkis vues plus haut, on a $V_{\mathscr{A}} \leq d+1$. Ainsi,

$$\mathbb{P}\left(\left|L(g_n^\star) - \inf_{g\in\mathscr{G}} L(g)\right| > \varepsilon\right) \leq 8(n+1)^{d+1}e^{-n\varepsilon^2/128}$$

et

$$L(g_n^\star) - \inf_{g\in\mathscr{G}} L(g) \to 0 \ \mathbb{P}\text{-p.s.}$$

2. **Classification par des boules fermées.** La classe $\mathscr{G}$ est composée de toutes les indicatrices des boules fermées de $\mathbb{R}^d$. Ainsi,

$$\mathscr{A} = \left\{ \left\{ x \in \mathbb{R}^d : \sum_{j=1}^{d} |x^{(j)} - a_j|^2 \leq a_0 \right\} : (a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1} \right\}.$$

   En remarquant que

$$a_0 - \sum_{j=1}^{d} |x^{(j)} - a_j|^2 = a_0 - \sum_{j=1}^{d} \left(x^{(j)}\right)^2 + 2 \sum_{j=1}^{d} x^{(j)} a_j - \sum_{j=1}^{d} a_j^2,$$

   on voit que $\mathscr{A}$ est inclus dans une famille d'ensembles de la forme $= \{\{x \in \mathbb{R}^d : f(x) \geq 0\} : f \in \mathscr{F}\}$, où $\mathscr{F}$ est un espace vectoriel de dimension $d + 2$. On conclut comme précédemment que

$$L(g_n^\star) - \inf_{g \in \mathscr{G}} L(g) \to 0 \ \mathbb{P}\text{-p.s.}$$

3. **Classification par des convexes.** On prend pour $\mathscr{A}$ l'ensemble de tous les convexes de $\mathbb{R}^d$, famille pour laquelle nous avons déjà vu que $V_{\mathscr{A}} = +\infty$. Cette classe d'ensembles est trop massive pour que l'erreur d'estimation puisse être raisonnablement contrôlée par la théorie de Vapnik-Chervonenkis.
4. **Classification linéaire généralisée.** On se place dans $\mathbb{R}^d$ et on se donne $\Psi_1, \ldots, \Psi_{d^\star}$ un nombre fixe de fonctions mesurables de $\mathbb{R}^d \to \mathbb{R}$. Les règles de classification considérées sont alors de la forme

$$g(x) = \begin{cases} 1 & \text{si } \sum_{j=1}^{d^\star} a_j \Psi_j(x) + a_0 > 0 \\ 0 & \text{sinon,} \end{cases}$$

   où $(a_0, a_1, \ldots, a_{d^\star}) \in \mathbb{R}^{d^\star}$ est un paramètre vectoriel. Lorsque $\Psi_j(x) = x^{(j)}$, on retrouve la famille des règles linéaires. Néanmoins, bien d'autres choix sont possibles. En prenant par exemple pour les Ψ_j les applications coordonnées et produits de ces coordonnées, on voit que $\mathscr{A}$ est incluse dans une famille d'ensembles du type

$$\left\{ a_0 + \sum_{j=1}^{d} a_j x^{(j)} + \sum_{j=1}^{d} b_j \left(x^{(j)}\right)^2 + \sum_{1 \leq j_1 < j_2 \leq d} c_{j_1} c_{j_2} x^{(j_1)} x^{(j_2)} \geq 0 \right\}.$$

   Ainsi, en posant $d^\star = 1 + 2d + \frac{d(d+1)}{2}$, $V_{\mathscr{A}} \leq d^\star$, et donc

$$L(g_n^\star) - \inf_{g \in \mathscr{G}} L(g) \to 0 \ \mathbb{P}\text{-p.s.}$$

Chapitre 10

Théorème de Stone et plus proches voisins

10.1 Classification et régression

Les chapitres précédents ont mis en lumière le rôle essentiel joué par le principe de minimisation du risque empirique et le théorème de Vapnik-Chervonenkis dans le contrôle de l'erreur d'estimation. Il s'avère cependant que les familles de règles de décision admettant une dimension de Vapnik-Chervonenkis finie sont presque toujours trop petites et ne permettent pas d'approcher correctement le risque de Bayes. On peut par exemple montrer que pour n'importe quelle famille de règles $\mathcal{G}$ dont la classe de boréliens associée $\mathscr{A}$ admet une dimension de Vapnik-Chervonenkis finie, et pour tout $\varepsilon \in]0, 1/2[$, il existe un couple de variables aléatoires (X, Y) tel que

$$\inf_{g \in \mathcal{G}} L(g) - L^\star > 1/2 - \varepsilon.$$

Il existe cependant d'autres façons de procéder. Ainsi, en partant de l'identité

$$g^\star(x) = \begin{cases} 1 & \text{si } r(x) > 1/2 \\ 0 & \text{sinon,} \end{cases} \tag{10.1}$$

une stratégie concurrente de la minimisation du risque empirique consiste à utiliser l'échantillon $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ pour estimer la fonction de régression $r(x) = \mathbb{E}(Y|X = x)$, et remplacer r par son estimateur r_n

dans (10.1). La règle de classification résultante, dite *règle plug-in*, s'écrit donc naturellement

$$g_n(x) = \begin{cases} 1 & \text{si } r_n(x) > 1/2 \\ 0 & \text{sinon.} \end{cases}$$

Le théorème qui suit précise le lien entre g_n et r_n, en termes d'erreur. On rappelle que μ désigne la loi de la variable aléatoire X.

Théorème 16. *Soit r_n un estimateur de la fonction de régression et g_n la règle de décision plug-in associée. Alors*

$$0 \leq L(g_n) - L^\star \leq 2 \int_{\mathbb{R}^d} |r_n(x) - r(x)| \, \mu(\mathrm{d}x).$$

En particulier, pour tout $p \geq 1$,

$$0 \leq L(g_n) - L^\star \leq 2 \left(\int_{\mathbb{R}^d} |r_n(x) - r(x)|^p \, \mu(\mathrm{d}x) \right)^{1/p},$$

et

$$0 \leq \mathbb{E}L(g_n) - L^\star \leq 2 \, \mathbb{E}^{1/p} \left| r_n(X) - r(X) \right|^p.$$

Démonstration. En procédant comme dans la preuve du Lemme 2, nous pouvons écrire

$$\begin{aligned}
&\mathbb{P}(g_n(X) \neq Y | X, \mathcal{D}_n) \\
&\quad = 1 - \mathbb{P}(g_n(X) = Y | X, \mathcal{D}_n) \\
&\quad = 1 - \left(\mathbb{P}(g_n(X) = 1, Y = 1 | X, \mathcal{D}_n) + \mathbb{P}(g_n(X) = 0, Y = 0 | X, \mathcal{D}_n) \right) \\
&\quad = 1 - \left(\mathbb{1}_{[g_n(X)=1]} \mathbb{P}(Y = 1 | X, \mathcal{D}_n) + \mathbb{1}_{[g_n(X)=0]} \mathbb{P}(Y = 0 | X, \mathcal{D}_n) \right) \\
&\quad = 1 - \left(\mathbb{1}_{[g_n(X)=1]} r(X) + \mathbb{1}_{[g_n(X)=0]} \left(1 - r(X)\right) \right),
\end{aligned}$$

où, dans la dernière inégalité, nous avons utilisé l'indépendance entre le couple (X, Y) et $\mathcal{D}_n$. De façon similaire,

$$\mathbb{P}(g^\star(X) \neq Y | X) = 1 - \left(\mathbb{1}_{[g^\star(X)=1]} r(X) + \mathbb{1}_{[g^\star(X)=0]} \left(1 - r(X)\right) \right).$$

Ainsi,

$$\begin{aligned}
&\mathbb{P}(g_n(X) \neq Y | X, \mathcal{D}_n) - \mathbb{P}(g^\star(X) \neq Y | X) \\
&\quad = r(X) \left(\mathbb{1}_{[g^\star(X)=1]} - \mathbb{1}_{[g_n(X)=1]} \right) + (1 - r(X)) \left(\mathbb{1}_{[g^\star(X)=0]} - \mathbb{1}_{[g_n(X)=0]} \right) \\
&\quad = (2r(X) - 1) \left(\mathbb{1}_{[g^\star(X)=1]} - \mathbb{1}_{[g_n(X)=1]} \right) \\
&\quad = |2r(X) - 1| \mathbb{1}_{[g_n(X) \neq g^\star(X)]}.
\end{aligned}$$

Finalement,

$$\begin{aligned}\mathbb{P}(g_n(X) \neq Y|\mathcal{D}_n) - L^\star &= 2\int_{\mathbb{R}^d} |r(x) - 1/2| \mathbb{1}_{[g_n(x) \neq g^\star(x)]} \mu(\mathrm{d}x) \\ &\leq 2\int_{\mathbb{R}^d} |r_n(x) - r(x)|\, \mu(\mathrm{d}x),\end{aligned}$$

puisque $g_n(x) \neq g^\star(x)$ implique $|r_n(x) - r(x)| \geq |r(x) - 1/2|$. Les autres assertions découlent de l'inégalité de Hölder et de celle de Jensen. □

On retiendra du Théorème 16 que si l'on dispose d'un estimateur r_n de la fonction de régression qui soit tel que

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \, \mu(\mathrm{d}x) \to 0 \tag{10.2}$$

dans L^1 (ou presque sûrement), alors la règle de classification associée g_n est automatiquement convergente (ou fortement convergente). Il nous reste donc à savoir comment construire des estimateurs de la fonction de régression qui possèdent la propriété de convergence (10.2). Le théorème de Stone, que nous présentons dans la prochaine section, répond précisément à cette question.

10.2 Le théorème de Stone

Une façon canonique de construire des estimateurs de la fonction de régression $r(x) = \mathbb{E}(Y|X = x)$ à partir du n-échantillon $\mathscr{D}_n$ consiste à écrire

$$r_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \; x \in \mathbb{R}^d, \tag{10.3}$$

où chaque $W_{ni}(x)$ est une fonction borélienne réelle de x et $X_1, \ldots, X_n$ (pas $Y_1, \ldots, Y_n$). Il est intuitivement clair que les couples (X_i, Y_i) pour lesquels X_i est "proche" de x (en un sens qui reste à préciser) devraient apporter davantage d'information sur $r(x)$ que leurs homologues plus éloignés. En conséquence, les poids devront en règle générale être plus grands autour de x, de telle sorte que $r_n(x)$ ainsi défini se présente comme une moyenne pondérée des Y_i correspondants aux X_i situés dans un voisinage de x. Voilà pourquoi un estimateur r_n de la forme (10.3) est appelé estimateur *de type*

moyenne locale. Bien souvent (mais pas toujours), les $W_{ni}(x)$ sont positifs et normalisés à 1, de telle sorte que $(W_{n1}(x), \ldots, W_{nn}(x))$ est en fait un vecteur de probabilités.

Un exemple typique d'estimateur de type moyenne locale est l'*estimateur à noyau*, qui est obtenu en prenant

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)},$$

où K une fonction positive mesurable sur $\mathbb{R}^d$ (appelée *noyau*), et h est un paramètre strictement positif (appelé *fenêtre*), éventuellement fonction de n. En d'autres termes, pour $x \in \mathbb{R}^d$,

$$r_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}.$$

(Si le dénominateur est nul, on pose $r_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i$.) En particulier, pour le choix de noyau dit *naïf* $K(z) = \mathbb{1}_{[\|z\| \leq 1]}$, on obtient

$$r_n(x) = \frac{\sum_{i=1}^{n} \mathbb{1}_{[\|x-X_i\| \leq h]} Y_i}{\sum_{j=1}^{n} \mathbb{1}_{[\|x-X_j\| \leq h]}},$$

ce qui montre que $r(x)$ est estimé par la moyenne des Y_i tels que la distance euclidienne entre x et X_i ne dépasse pas h. Pour un noyau plus général K, le poids de Y_i dépend de la distance entre x et X_i par l'intermédiaire de la forme du noyau. Les noyaux les plus classiques sont le noyau d'*Epanechnikov* $K(z) = (1 - \|z\|^2)1_{[\|z\| \leq 1]}$ et le noyau *gaussien* $K(z) = e^{-\|z\|^2}$.

Un second exemple important d'estimateur de type moyenne locale nous est fourni par l'*estimateur des plus proches voisins* :

$$r_n(x) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(x), \ x \in \mathbb{R}^d,$$

pour lequel $(v_{n1}, \ldots, v_{nn})$ est un vecteur de poids (déterministes) positifs normalisés à 1, et la suite $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ est la permutation de $(X_1, Y_1), \ldots, (X_n, Y_n)$ correspondante aux distances croissantes

des $\|X_i - x\|$ (en cas d'égalité $\|X_i - x\| = \|X_j - x\|$ avec $i < j$, X_i sera arbitrairement déclaré plus proche de x que X_j). En d'autres termes,

$$\|X_{(1)}(x) - x\| \leq \|X_{(2)}(x) - x\| \leq \cdots \leq \|X_{(n)}(x) - x\|.$$

Pour s'assurer que cet estimateur est bien de la forme (10.3), il suffit de poser

$$W_{ni}(x) = v_{n\Sigma_i},$$

où $(\Sigma_1, \ldots, \Sigma_n)$ est la permutation de $(1, \ldots, n)$ telle que X_i est le Σ_i-ème plus proche voisin de x. Parmi tous les choix possibles $(v_{n1}, \ldots, v_{nn})$, un cas particulier important est obtenu en posant $v_{ni} = 1/k$ pour $1 \leq i \leq k$ et $v_{ni} = 0$ autrement, avec $\{k\} = \{k_n\}$ une suite d'entiers strictement positifs ne dépassant pas n. L'estimateur résultant s'appelle *estimateur des k-plus proches voisins* et s'écrit donc

$$r_n(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(x), \ x \in \mathbb{R}^d.$$

Le principe est naturel : pour estimer la fonction de régression autour de x, on regarde les k observations X_i les plus proches de x et on fait la moyenne des Y_i correspondants.

Conformément à ce que nous avons dit dans la section précédente, on associe naturellement à un estimateur de type moyenne locale la règle de classification plug-in

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n W_{ni}(x) Y_i > 1/2 \\ 0 & \text{sinon} \end{cases}$$

ou, de façon équivalente, si $\sum_{i=1}^n W_{ni}(x) = 1$,

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{[Y_i=1]} > \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{[Y_i=0]} \\ 0 & \text{sinon.} \end{cases}$$

Le théorème ci-après, connu sous le nom de *théorème de Stone*, donne des conditions suffisantes sur les poids $W_{ni}(x)$ garantissant que la règle de classification plug-in g_n soit convergente. Pour simplifier, nous supposerons désormais que les poids $W_{ni}(x)$ sont positifs et normalisés à 1 (i.e., $\sum_{i=1}^n W_{ni}(x) = 1$), ce qui fait de $(W_{n1}(x), \ldots, W_{nn}(x))$ un vecteur de probabilités.

Théorème 17 (Stone)**.** *Supposons que, quelle que soit la loi de X, les poids satisfassent les conditions suivantes :*

1. *Il existe une constante c telle que, pour toute fonction borélienne $f : \mathbb{R}^d \to \mathbb{R}$ telle que $\mathbb{E}|f(X)| < \infty$,*
$$\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X) \, |f(X_i)| \right) \leq c \, \mathbb{E} \, |f(X)| \text{ pour tout } n \geq 1.$$
2. *Pour tout $a > 0$,*
$$\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{[\|X_i - X\| > a]} \right) \to 0.$$
3. *On a*
$$\mathbb{E}\left(\max_{1 \leq i \leq n} W_{ni}(X) \right) \to 0.$$

Alors la règle de classification associée g_n est universellement convergente, i.e.

$$\mathbb{E}L(g_n) \to L^\star$$

quelle que soit la loi du couple (X, Y).

La condition 2 exprime le fait que la contribution des poids à l'extérieur de n'importe quelle boule fermée centrée en X doit être asymptotiquement négligeable. En d'autres termes, seuls les points situés dans un voisinage local de la cible sont importants pour l'évaluation de la moyenne. La condition 3 interdit à un seul point d'avoir une influence disproportionnée sur le calcul de l'estimateur. Enfin, l'hypothèse 1, parfois appelée condition de Stone, est essentiellement de nature technique. Insistons bien sur le fait que le résultat du théorème est universel, au sens où la convergence est valable *quelle que soit la loi* du couple (X, Y) !

Démonstration du Théorème 17. En vertu du Théorème 16, nous savons qu'il suffit de montrer que, quelle que soit la loi de (X, Y), on a

$$\mathbb{E} \, (r_n(X) - r(X))^2 = \int_{\mathbb{R}^d} (r_n(x) - r(x))^2 \, \mu(\mathrm{d}x) \to 0.$$

Posons

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) r(X_i).$$

En utilisant l'inégalité $(a+b)^2 \leq 2(a^2+b^2)$, on a

$$\begin{aligned}\mathbb{E}\left(r_n(X)-r(X)\right)^2 &= \mathbb{E}\left(r_n(X)-\hat{r}_n(X)+\hat{r}_n(X)-r(X)\right)^2 \\ &\leq 2\Big(\mathbb{E}\left(r_n(X)-\hat{r}_n(X)\right)^2+\mathbb{E}\left(\hat{r}_n(X)-r(X)\right)^2\Big). \quad (10.4)\end{aligned}$$

Il suffit donc de montrer que chacun des deux termes ci-dessus tend vers 0 lorsque n tend vers l'infini. Comme les poids $W_{ni}(x)$ sont positifs et normalisés, l'inégalité de Jensen permet d'écrire que

$$\begin{aligned}\mathbb{E}\left(\hat{r}_n(X)-r(X)\right)^2 &= \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\left(r(X_i)-r(X)\right)\Big)^2 \\ &\leq \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\left(r(X_i)-r(X)\right)^2\Big).\end{aligned}$$

Si la fonction r, qui satisfait $0 \leq r \leq 1$, est continue à support borné, alors elle est aussi uniformément continue. Ainsi, pour tout $\varepsilon > 0$, il existe $a > 0$ tel que $\|x-x'\| \leq a$ implique $|r(x)-r(x')|^2 \leq \varepsilon$. Dans ce cas, comme $|r(x)-r(x')| \leq 1$, on obtient

$$\begin{aligned}&\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\left(r(X_i)-r(X)\right)^2\Big) \\ &\quad\leq \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{[\|X_i-X\|>a]}\Big)+\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\varepsilon\Big) \to \varepsilon,\end{aligned}$$

d'après la condition 2. Dans le cas général, par densité, on peut toujours trouver une fonction $r^\star$ continue à support borné telle que

$$\mathbb{E}\left(r(X)-r^\star(X)\right)^2 < \varepsilon.$$

Avec ce choix, et en utilisant le fait que $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, on trouve

$$\begin{aligned}&\mathbb{E}\left(\hat{r}_n(X)-r(X)\right)^2 \\ &\quad\leq \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\left(r(X_i)-r(X)\right)^2\Big) \\ &\quad\leq 3\Big(\sum_{i=1}^n W_{ni}(X)\Big(\left(r(X_i)-r^\star(X_i)\right)^2+\left(r^\star(X_i)-r^\star(X)\right)^2 \\ &\qquad+\left(r^\star(X)-r(X)\right)^2\Big)\Big).\end{aligned}$$

Ainsi, en utilisant la condition 1,

$$\begin{aligned}
&\mathbb{E}\left(\hat{r}_n(X) - r(X)\right)^2 \\
&\quad \leq 3c\,\mathbb{E}\left(r(X) - r^\star(X)\right)^2 + 3\mathbb{E}\left(\sum_{i=1}^n W_{ni}(X)\left(r^\star(X_i) - r^\star(X)\right)^2\right) \\
&\qquad + 3\mathbb{E}\left(r^\star(X) - r(X)\right)^2.
\end{aligned}$$

Au final,

$$\limsup_{n\to+\infty} \mathbb{E}\left(\hat{r}_n(X) - r(X)\right)^2 \leq 3\varepsilon(2+c).$$

Il nous reste à contrôler le premier terme du membre de droite de l'inégalité (10.4). Observons pour cela que, pour $i \neq j$,

$$\begin{aligned}
&\mathbb{E}\left(W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\big(Y_j - r(X_j)\big)\right) \\
&\quad = \mathbb{E}\Big[\mathbb{E}\left(W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\big(Y_j - r(X_j)\big) \mid X, X_1, \ldots, X_n, Y_i\right)\Big] \\
&\quad = \mathbb{E}\Big[W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\mathbb{E}(Y_j - r(X_j) \mid X, X_1, \ldots, X_n, Y_i)\Big] \\
&\quad = \mathbb{E}\Big[W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\mathbb{E}(Y_j - r(X_j) \mid X_j)\Big] \\
&\qquad \text{(par indépendence entre } (X_j, Y_j) \text{ et } X, X_1, X_{j-1}, X_{j+1}, \ldots, X_n, Y_i) \\
&\quad = \mathbb{E}\Big[W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\big(r(X_j) - r(X_j)\big)\Big] \\
&\quad = 0.
\end{aligned}$$

Du coup,

$$\begin{aligned}
\mathbb{E}\left(r_n(X) - \hat{r}_n(X)\right)^2 &= \mathbb{E}\left(\sum_{i=1}^n W_{ni}(X)\left(Y_i - r(X_i)\right)\right)^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left(W_{ni}(X)\left(Y_i - r(X_i)\right)W_{nj}(X)\left(Y_j - r(X_j)\right)\right) \\
&= \sum_{i=1}^n \mathbb{E}\left(W_{ni}^2(X)\left(Y_i - r(X_i)\right)^2\right).
\end{aligned}$$

On en déduit que

$$\begin{aligned}
\mathbb{E}\left(r_n(X) - \hat{r}_n(X)\right)^2 &\leq \mathbb{E}\left(\sum_{i=1}^n W_{ni}^2(X)\right) \leq \mathbb{E}\left(\max_{1\leq i\leq n} W_{ni}(X)\sum_{j=1}^n W_{nj}(X)\right) \\
&= \mathbb{E}\left(\max_{1\leq i\leq n} W_{ni}(X)\right) \to 0,
\end{aligned}$$

d'après 3, ce qui conclut la preuve du théorème. □

10.3 k-plus proches voisins

Dans la suite, k est un entier strictement positif compris entre 1 et n (fonction de n). On rappelle que $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ désigne le réordonnement de l'échantillon original $(X_1, Y_1), \ldots, (X_n, Y_n)$ suivant les distances euclidiennes croissantes des X_i à x.

Nous avons vu dans la section précédente que la règle de classification des k-plus proches voisins a pour expression

$$g_n(x) = \begin{cases} 1 & \text{si } \frac{1}{k}\sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(x)=1]} > \frac{1}{k}\sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(x)=0]} \\ 0 & \text{sinon} \end{cases}$$

ou, de façon équivalente,

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(x)=1]} > \sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(x)=0]} \\ 0 & \text{sinon.} \end{cases}$$

Le prochain théorème, dont la preuve utilise le théorème de Stone, établit la convergence universelle de la règle g_n, pourvu que la dépendence de k en n ne soit pas trop anarchique.

Théorème 18. *Supposons que $k \to +\infty$ et $k/n \to 0$. Alors la règle de classification des k-plus proches voisins est universellement convergente, i.e.*

$$\mathbb{E}L(g_n) \to L^\star$$

quelle que soit la loi du couple (X, Y).

Pour prouver le résultat, il suffit simplement de s'assurer que les conditions $1-3$ du théorème de Stone sont effectivement vérifiées par la règle des k-plus proches voisins. Pour ce faire, nous aurons au préalable besoin de quelques lemmes techniques. Pour simplifier un peu, nous supposerons dans la suite que les égalités entre distances $\|X_i - x\| = \|X_j - x\|$ se produisent avec probabilité zéro (c'est par exemple le cas lorsque $\|X - x\|$ admet une densité par rapport à la mesure de Lebesgue). La preuve du Théorème 18 s'étend au cas général, au prix de quelques petits aménagements

techniques pour gérer les distances ex-aequo. On rappelle que le support de la loi μ est défini comme l'ensemble des $x \in \mathbb{R}^d$ tels que $\mu(B(x,\varepsilon)) > 0$, avec $B(x,\varepsilon)$ la boule fermée de centre x et de rayon ε. Alternativement, il s'agit du plus petit ensemble fermé de μ-mesure 1.

Lemme 5. *Soit x dans le support de μ. Alors, si $k/n \to 0$, on a*

$$\|X_{(k)}(x) - x\| \to 0 \ \mathbb{P}\text{-}p.s.$$

Démonstration. Fixons $\varepsilon > 0$ et observons, puisque x appartient au support de μ, que $\mu(B(x,\varepsilon)) > 0$. Notons également l'égalité suivante entre événements :

$$\left\{ \|X_{(k)}(x) - x\| > \varepsilon \right\} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[X_i \in B(x,\varepsilon)]} < \frac{k}{n} \right\}.$$

Or, d'après la loi forte des grands nombres,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[X_i \in B(x,\varepsilon)]} \to \mu\left(B(x,\varepsilon)\right) \ \mathbb{P}\text{-p.s.}$$

Comme $k/n \to 0$, on en conclut immédiatement que $\|X_{(k)}(x) - x\| \to 0$ $\mathbb{P}$-p.s. □

Lemme 6. *Soit ν une mesure de probabilité sur $\mathbb{R}^d$. Fixons $x' \in \mathbb{R}^d$ et définissons, pour $a \geq 0$,*

$$B_a(x') = \left\{ x \in \mathbb{R}^d : \nu\left(B(x, \|x' - x\|)\right) \leq a \right\}.$$

Alors

$$\nu\left(B_a(x')\right) \leq \gamma_d \, a,$$

où γ_d est une constante strictement positive ne dépendant que de d.

Démonstration. Fixons $x' \in \mathbb{R}^d$ et considérons une famille $\mathscr{C}_1, \ldots, \mathscr{C}_{\gamma_d}$ de cones d'angle $\pi/6$ centrés en x', suffisamment nombreux pour que leur union recouvre $\mathbb{R}^d$. En d'autres termes,

$$\bigcup_{j=1}^{\gamma_d} \mathscr{C}_j = \mathbb{R}^d.$$

Il est facile de voir (exercice) que si $u \in \mathscr{C}_j - \{x'\}$, $u' \in \mathscr{C}_j$ et $\|u - x'\| \leq \|u' - x'\|$, alors $\|u - u'\| < \|u' - x'\|$. Par ailleurs,

$$\nu\left(B_a(x')\right) \leq \sum_{j=1}^{\gamma_d} \nu\left(\mathscr{C}_j \cap B_a(x')\right).$$

Soit alors $x^\star \in \mathscr{C}_j \cap B_a(x')$. En utilisant la propriété géométrique des cones mentionnée ci-dessus, nous pouvons écrire

$$\nu\left(\mathscr{C}_j \cap B(x', \|x^\star - x'\|) \cap B_a(x')\right) \leq \nu\left(B(x^\star, \|x' - x^\star\|)\right) \leq a,$$

la seconde inégalité provenant du fait que $x^\star \in B_a(x')$. Puisque $x^\star$ est arbitraire, on en conclut que

$$\nu\left(\mathscr{C}_j \cap B_a(x')\right) \leq a.$$

□

Une conséquence fondamentale du lemme précédent indique que le nombre de points dans $\{X_1, \ldots, X_n\}$ pour lesquels X figure parmi les k plus proches voisins ne dépasse pas une constante fois k. Dans la suite, l'abréviation k-ppv signifie "k-plus proches voisins".

Corollaire 2. *Si les égalités entre distances se produisent avec probabilité zéro, on a*

$$\sum_{i=1}^{n} \mathbb{1}_{[X \text{ est parmi les } k\text{-ppv de } X_i \text{ dans } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}]} \leq k\gamma_d,$$

$\mathbb{P}$-*p.s.*

Démonstration. On applique le Lemme 6 avec $a = k/n$ et ν la mesure empirique μ_n associée à $X_1, \ldots, X_n$. Avec ce choix, on a

$$B_{k/n}(X) = \left\{x \in \mathbb{R}^d : \mu_n\left(B(x, \|X - x\|)\right) \leq k/n\right\}$$

et, $\mathbb{P}$-p.s.,

$$\begin{aligned} & X_i \in B_{k/n}(X) \\ & \Leftrightarrow \mu_n\left(B(X_i, \|X - X_i\|)\right) \leq k/n \\ & \Leftrightarrow X \text{ est parmi les } k\text{-ppv de } X_i \text{ dans } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}. \end{aligned}$$

(Noter que la seconde équivalence utilise le fait que les égalités entre distances se produisent avec probabilité zéro.) Ainsi, en appliquant le Lemme 6, il vient, $\mathbb{P}$-p.s.,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{1}_{[X \text{ est parmi les } k\text{-ppv de } X_i \text{ dans } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}]} \\
&\quad = \sum_{i=1}^{n} \mathbb{1}_{[X_i \in B_{k/n}(X)]} \\
&\quad = n \times \mu_n\left(B_{k/n}(X)\right) \\
&\quad \leq k\gamma_d.
\end{aligned}$$

□

Lemme 7. *Supposons que les égalités entre distances se produisent avec probabilité zéro. Alors, pour toute fonction borélienne $f : \mathbb{R}^d \to \mathbb{R}$ telle que $\mathbb{E}|f(X)| < \infty$, on a*

$$\sum_{i=1}^{n} \mathbb{E}\left|f(X_{(i)}(X)\right| \leq k\gamma_d \mathbb{E}\left|f(X)\right|,$$

où γ_d est une constante strictement positive ne dépendant que de d.

Démonstration. Prenons f une fonction comme dans l'énoncé. Alors

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left|f(X_{(i)}(X)\right| \\
&= \mathbb{E}\left(\sum_{i=1}^{n} |f(X_i)|\, \mathbb{1}_{[X \text{ est parmi les } k\text{-ppv de } X \text{ dans } \{X_1,\ldots,X_n\}]}\right) \\
&= \mathbb{E}\Bigg(|f(X)| \\
&\times \sum_{i=1}^{n} \mathbb{1}_{[X \text{ est parmi les } k\text{-ppv de } X_i \text{ dans } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}]}\Bigg) \\
&\quad \text{(en échangeant } X \text{ et } X_i\text{)} \\
&\leq \mathbb{E}\left(|f(X)|\, k\gamma_d\right),
\end{aligned}$$

d'après le Corollaire 2. Ceci conclut la preuve du lemme. □

Nous sommes désormais en mesure de démontrer le Théorème 18. Il suffit pour cela de vérifier les conditions du théorème de Stone, avec $W_{ni}(x) = 1/k$ si X_i est parmi les k plus proches voisins de X, et $W_{ni}(x) = 0$ sinon.

La condition 3 est évidente dans la mesure où $k \to +\infty$. Pour la condition 2, on note que

$$\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{[\|X_i - X\|] > \varepsilon} \right) = \mathbb{E}\left(\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{[\|X_{(i)}(x) - X\|] > \varepsilon} \right),$$

de sorte que

$$\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{[\|X_i - X\|] > \varepsilon} \right) \to 0$$

dès lors que, pour tout $\varepsilon > 0$,

$$\mathbb{P}\big(\|X_{(k)}(X) - X\| > \varepsilon \big) \to 0.$$

Or,

$$\mathbb{P}\big(\|X_{(k)}(X) - X\| > \varepsilon \big) = \int_{\mathbb{R}^d} \mathbb{P}\big(\|X_{(k)}(x) - x\| > \varepsilon \big) \mu(\mathrm{d}x).$$

Pour x fixé dans le support de μ, le Lemme 5 indique que la convergence

$$\mathbb{P}\big(\|X_{(k)}(x) - x\| > \varepsilon \big) \to 0$$

a lieu lorsque $k/n \to 0$. Le résultat s'en déduit par convergence dominée, en notant que le support de μ est de μ-mesure 1.

Examinons pour terminer la condition 1. Il s'agit de voir que, pour toute fonction f telle que $\mathbb{E}|f(X)| < \infty$, on a

$$\mathbb{E}\left(\frac{1}{k} \sum_{i=1}^{n} |f(X_i)| \, \mathbb{1}_{[X_i \text{ est parmi les } k\text{-ppv de } X]} \right) \leq c \, \mathbb{E}\,|f(X)|\,,$$

pour une certaine constante c. C'est précisément l'énoncé du Lemme 7.

Chapitre 11

Quantification et clustering

11.1 Principe de la quantification

La quantification est un principe probabiliste dont l'objectif est de compresser l'information contenue dans une variable aléatoire X à valeurs dans $(\mathbb{R}^d, \|\cdot\|)$, où $\|\cdot\|$ désigne la norme euclidienne. On se donne dorénavant une telle variable X, en notant μ sa loi et en supposant que $\mathbb{E}\|X\|^2 < \infty$ ou, ce qui est équivalent, que

$$\int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x) < \infty.$$

Définition 31. *Soit k un entier ≥ 1. Un quantifieur q d'ordre k est une fonction mesurable $q : \mathbb{R}^d \to \mathscr{C} \subset \mathscr{H}$ avec $|\mathscr{C}| \leq k$.*

Un quantifieur q d'ordre k est donc caractérisé par :

- ▷ Un *alphabet* $\mathscr{C} = \{c_1, \ldots, c_k\}$.
- ▷ Une *partition* $\mathscr{P} = \{A_1, \ldots, A_k\}$ de $\mathbb{R}^d$, avec la numérotation imposée par
$$q(x) = c_j \Leftrightarrow x \in A_j.$$

On écrira dans la suite $q = (\mathscr{C}, \mathscr{P})$. Un quantifieur apparaît ainsi comme un outil de compression de l'information. L'étape suivante consiste alors à se doter d'un critère mesurant la pertinence de la compression de la variable aléatoire X (ou de sa loi) au travers de q.

Définition 32. *La distorsion (pour X ou μ) d'un quantifieur $q = (\mathscr{C}, \mathscr{P})$ d'ordre k est définie par*

$$D(\mu, q) = \mathbb{E}\|X - q(X)\|^2 = \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu(\mathrm{d}x).$$

La distorsion minimale à l'ordre k est

$$D_k^\star(\mu) = \inf_q D(\mu, q),$$

où l'infimum est évalué sur tous les quantifieurs d'ordre k.

Bien entendu, plus la distorsion est faible, meilleure est la compression. Par ailleurs, comme on s'en doute, la qualité d'une quantification s'améliore lorsque k grandit. Ce phénomène est précisé dans le lemme ci-dessous.

Lemme 8. *On a $D_k^\star(\mu) \searrow 0$ si $k \to +\infty$.*

Démonstration. Tout d'abord, il est clair que la distorsion minimale décroît à mesure que son ordre augmente. Puis, comme $\mathbb{R}^d$ est un espace polonais, la mesure bornée ν définie pour tout borélien A de $\mathbb{R}^d$ par

$$\nu(A) = \int_A \|x\|^2 \mu(\mathrm{d}x)$$

est tendue, i.e. pour tout $\varepsilon \in]0, 1]$, il existe un compact K tel que $\nu(K) \geq 1 - \varepsilon$. On note $\{c_1, c_2, \ldots\}$ un sous-ensemble dénombrable dense dans $\mathbb{R}^d$. Comme K est compact, on a pour tout k assez grand

$$K \subset B := \bigcup_{j=1}^{k} B(c_j, \sqrt{\varepsilon}).$$

On a donc $\nu(B) \geq 1 - \varepsilon$. Notons maintenant q_{k+1} le quantifieur d'ordre $k+1$ d'alphabet $\{c_1, \ldots, c_k, 0\}$ (en supposant, sans perte de généralité, que $0 \notin \{c_1, c_2, \ldots\}$) et de partition $\{A_1, \ldots, A_k, B^c\}$, avec $A_1 = B(c_1, \sqrt{\varepsilon})$ et, pour $j = 2, \ldots, k$, $A_j = B(c_j, \sqrt{\varepsilon}) \setminus A_{j-1}$. Comme $\|x - c_j\| \leq \sqrt{\varepsilon}$ si $x \in A_j$, on a

$$\begin{aligned} D_{k+1}^\star(\mu) \leq D_{k+1}(\mu, q_{k+1}) &= \int_{\mathbb{R}^d} \|x - q_{k+1}(x)\|^2 \mu(\mathrm{d}x) \\ &= \sum_{j=1}^{k} \int_{A_j} \|x - c_j\|^2 \mu(\mathrm{d}x) + \int_{B^c} \|x\|^2 \mu(\mathrm{d}x) \\ &\leq \varepsilon \mu\left(\bigcup_{j=1}^{k} A_j \right) + \nu(B^c) \leq 2\varepsilon, \end{aligned}$$

ce qui achève la preuve. □

Parmi toutes les façons possibles de compresser l'information, la classe des quantifieurs de type *plus proches voisins* (que nous abrégerons désormais en *quantifieurs PPV*) joue un rôle bien particulier. Dans la suite, on suppose que les quantifieurs sont d'ordre k et on note, pour un alphabet $\mathscr{C} = \{c_1, \dots, c_k\} \subset \mathbb{R}^d$ de taille k, $\mathscr{P}_{\mathrm{V}}(\mathscr{C})$ la partition de Voronoi associée à $\mathscr{C}$, définie par

$$A_1 = \left\{ x \in \mathbb{R}^d : \|x - c_1\| \leq \|x - c_\ell\|, \, \forall \ell = 1, \dots, k \right\}, \text{ et}$$
$$A_j = \left\{ x \in \mathbb{R}^d : \|x - c_j\| \leq \|x - c_\ell\|, \, \forall \ell = 1, \dots, k \right\} \setminus \bigcup_{t=1}^{j-1} A_t,$$

pour $j = 2, \dots, k$.

Définition 33. *Un quantifieur d'ordre k est un quantifieur PPV si sa partition est une partition de Voronoi associée à son alphabet. En d'autres termes, un quantifieur PPV s'écrit $q = (\mathscr{C}, \mathscr{P}_{\mathrm{V}}(\mathscr{C}))$, avec $\mathscr{C} \subset \mathbb{R}^d$ de cardinal inférieur ou égal à k.*

Un quantifieur PPV q est donc entièrement caractérisé par son alphabet (dont les élèments sont appelés *centres* ou *centroïdes*), via la règle

$$\|x - q(x)\| = \min_{c_j \in \mathscr{C}} \|x - c_j\|,$$

les égalités entre distances sur le bord des cellules étant brisées en faveur des plus petits indices. On notera les propriétés élémentaires suivantes.

Proposition 6. *Soit q_{ppv} un quantifieur PPV d'alphabet $\mathscr{C} = \{c_1, \dots, c_k\}$. Alors*

$$D(\mu, q_{\mathrm{ppv}}) = \mathbb{E} \min_{1 \leq j \leq k} \|X - c_j\|^2 = \int_{\mathbb{R}^d} \min_{1 \leq j \leq k} \|x - c_j\|^2 \mu(\mathrm{d}x).$$

En outre, pour tout quantifieur $q = (\mathscr{C}, \mathscr{P})$, on a $D(\mu, q_{\mathrm{ppv}}) \leq D(\mu, q)$.

Démonstration. Pour la première propriété, en désignant par $\mathscr{P}_{\mathrm{V}}(\mathscr{C}) = \{A_{\mathrm{V},1}, \dots, A_{\mathrm{V},k}\}$ la partition de Voronoi associée à $\mathscr{C}$:

$$\begin{aligned} D(\mu, q_{\mathrm{ppv}}) &= \int_{\mathbb{R}^d} \|x - q_{\mathrm{ppv}}(x)\|^2 \mu(\mathrm{d}x) = \sum_{j=1}^{k} \int_{A_{\mathrm{V},j}} \|x - c_j\|^2 \mu(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d} \min_{1 \leq j \leq k} \|x - c_j\|^2 \mu(\mathrm{d}x). \end{aligned}$$

Puis, pour la seconde propriété, si $\mathscr{P} = \{A_1, \dots, A_k\}$,

$$\begin{aligned} D(\mu, q_{\text{ppv}}) &= \sum_{j=1}^{k} \int_{A_j} \min_{1\le j\le k} \|x - c_j\|^2 \mu(\mathrm{d}x) \\ &\le \sum_{j=1}^{k} \int_{A_j} \|x - c_j\|^2 \mu(\mathrm{d}x) \\ &\le \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu(\mathrm{d}x) = D(\mu, q), \end{aligned}$$

par définition de la distorsion. □

La conséquence fondamentale de cette dernière proposition est que les quantifieurs de distorsion minimale, s'ils existent, sont à rechercher parmi les quantifieurs du type $q_{\text{ppv}} = (\mathbf{c}, \mathscr{P}_{\text{V}}(\mathbf{c}))$ avec $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$ (noter l'abus de notation), de distorsion

$$W(\mu, \mathbf{c}) := \int_{\mathbb{R}^d} \min_{1\le j\le k} \|x - c_j\|^2 \mu(\mathrm{d}x) = D(\mu, q_{\text{ppv}}).$$

Théorème 19. *Il existe un quantifieur de distorsion minimale.*

Esquisse de démonstration. D'après la Proposition 6, on peut restreindre l'étude aux quantifieurs PPV. Il s'agit donc de montrer qu'il existe $\mathbf{c}^\star \in \mathbb{R}^{dk}$ tel que

$$W(\mu, \mathbf{c}^\star) = \inf_{\mathbf{c}\in\mathbb{R}^{dk}} W(\mu, \mathbf{c}).$$

On prouve d'abord (admis) qu'il existe $R > 0$ tel que

$$\inf_{\mathbf{c}\in\mathbb{R}^{dk}} W(\mu, \mathbf{c}) = \inf_{\|\mathbf{c}\|\le R} W(\mu, \mathbf{c}).$$

On établit ensuite que la fonction $\mathbb{R}^{dk} \ni \mathbf{c} \mapsto W(\mu, \mathbf{c})$ est continue. Observons pour cela que la fonction $x \mapsto \min_{1\le j\le k} \|x - c_j\|$ est continue. Dès lors, pour $\mathbf{c}_0 = (c_{1,0}, \dots, c_{k,0}) \in \mathbb{R}^{dk}$, on a

$$\begin{aligned} \lim_{\mathbf{c}\to\mathbf{c}_0} W(\mu, \mathbf{c}) &= \int_{\mathbb{R}^d} \lim_{\mathbf{c}\to\mathbf{c}_0} \min_{1\le j\le k} \|x - c_j\|^2 \mu(\mathrm{d}x) \\ &\quad \text{(d'après le théorème de Lebesgue)} \\ &= \int_{\mathbb{R}^d} \min_{1\le j\le k} \|x - c_{j,0}\|^2 \mu(\mathrm{d}x) \\ &\quad \text{(par continuité)} \\ &= W(\mu, \mathbf{c}_0), \end{aligned}$$

ce qui montre bien que $W(\mu, \cdot)$ est continue.

On déduit de cette dernière propriété et de la compacité de la boule $B(0, R)$ de $\mathbb{R}^{dk}$ qu'il existe $\mathbf{c}^\star \in \mathbb{R}^{dk}$ minimum de $W(\mu, \cdot)$. Le quantifieur $q^\star = (\mathbf{c}^\star, \mathscr{P}_{\text{ppv}}(\mathbf{c}^\star))$ est alors de distorsion minimale car

$$W(\mu, \mathbf{c}^\star) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, c) = \inf_q D(\mu, q) = D^\star(\mu).$$

□

11.2 Quantification empirique et clustering

Dans la pratique, la loi μ de la variable aléatoire X est inconnue et il est donc, par voie de conséquence, impossible de procéder à sa quantification. On dispose cependant bien souvent d'un n-échantillon i.i.d. $X_1, \ldots, X_n$ formé de variables aléatoires indépendantes entre elles, de même loi que X et indépendante de cette dernière. C'est à partir de cet échantillon que l'on va s'attacher à construire un *quantifieur empirique* $q_n(\cdot) = q_n(\cdot, X_1, \ldots, X_n)$ dont les performances se rapprochent si possible de celles du quantifieur optimal.

On suppose toujours que $\mathbb{E}\|X\|^2 < \infty$ et on désigne par μ_n la mesure empirique associée à $X_1, \ldots, X_n$, i.e.

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

Dans ce contexte, la distorsion pour μ du quantifieur empirique q_n (d'ordre k) est naturellement définie par

$$D(\mu, q_n) = \mathbb{E}\big(\|X - q_n(X)\|^2 \mid X_1, \ldots, X_n\big) = \int_{\mathbb{R}^d} \|x - q_n(x)\|^2 \mu(\mathrm{d}x)$$

(noter qu'il s'agit d'une variable aléatoire). La *distorsion empirique* d'un quantifieur q prend la forme

$$D(\mu_n, q) = \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu_n(\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^n \|X_i - q(X_i)\|^2.$$

Dans le cas particulier de $q_{\text{ppv}} = (\mathbf{c}, \mathscr{P}_{\text{V}}(\mathbf{c}))$, avec $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$, on a

$$D(\mu_n, q_{\text{ppv}}) = W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|X_i - c_j\|^2.$$

Pour se doter d'outils qui assurent que la méthode de quantification empirique est performante, on introduit la définition qui suit.

Définition 34. *Soit q_n un quantifieur empirique. On dit qu'il est consistant si $\mathbb{E}D(\mu, q_n) \to D^\star(\mu)$. On dit qu'il est de vitesse $(v_n)_n$ si $\mathbb{E}D(\mu, q_n) - D^\star(\mu) = \mathrm{O}(1/v_n)$, avec $v_n \to +\infty$.*

On aura noté au passage que, puisque $D(\mu, q_n) \geq D^\star(\mu)$, la propriété $\mathbb{E}D(\mu, q_n) \to D^\star(\mu)$ est équivalente à $D(\mu, q_n) \to D^\star(\mu)$ dans L^1.

Le quantifieur empirique $q_n^\star$ le plus naturel est obtenu en minimisant la distorsion empirique sur les quantifieurs PPV. En d'autres termes, on cherche les centres optimaux $\mathbf{c}_n^\star = (c_{n,1}^\star, \ldots, c_{n,k}^\star)$ tels que

$$W(\mu_n, \mathbf{c}_n^\star) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu_n, \mathbf{c}). \tag{11.1}$$

(Observons que $q_n^\star$ existe en vertu du Théorème 19.) On a donc

$$q_n^\star = (\mathbf{c}_n^\star, \mathscr{P}_{\text{V}}(\mathbf{c}_n^\star)).$$

Un quantifieur empirique q_n (d'ordre k) est naturellement associé à une méthode de regroupement (ou *clustering*) des données $X_1, \ldots, X_n$ en k classes, en décidant que l'observation X_i est rangée dans la classe j ($1 \leq j \leq k$) si $q_n(X_i) = j$.

Pour le quantifieur empirique PPV optimal $q_n^\star$, le j-ème cluster est constitué des observations X_i telles que $\|X_i - c_{n,j}^\star\| \leq \|X_i - c_{n,\ell}^\star\|$, $\forall \ell = 1, \ldots, k$.

On parle parfois, en lieu et place de clustering, de *classification (ou apprentissage) non supervisé*, l'adjectif "non supervisé" renvoyant au fait qu'il n'y a pas d'information annexe apportée par des variables réponses Y_i. Le problème consiste ici à regrouper les données $X_1, \ldots, X_n$ "à l'aveugle", de la façon la plus pertinente possible et sans information annexe.

Algorithme des k-means. En pratique, l'approche (11.1) par minimisation de la distorsion empirique est difficile à mettre en œuvre, surtout en grande dimension (problème NP-complet). On a alors recours à une technique approchée, appelée *algorithme des k-means*. Cette procédure est fondée sur la remarque suivante. Pour une partition $\mathscr{P} = \{A_1, \dots, A_k\}$ et $\mathscr{C} = \{c_1, \dots, c_k\} \subset \mathbb{R}^d$, on définit $q = (\mathscr{C}, \mathscr{P})$ et $q_n = (\mathscr{C}_n, \mathscr{P})$, avec $\mathscr{C}_n = \{c_{n,1}, \dots, c_{n,k}\}$ tel que, pour tout $j = 1, \dots, k$,

$$c_{n,j} = \arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^{n} \|X_i - y\|^2 \mathbb{1}_{[X_i \in A_j]} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{[X_i \in A_j]}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \in A_j]}}. \tag{11.2}$$

Noter que $c_{n,j}$ est, à un facteur près, une espérance conditionnelle pour la mesure empirique. On a alors

$$\begin{aligned} D(\mu_n, q) &= \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \|X_i - c_j\|^2 \mathbb{1}_{[X_i \in A_j]} \\ &\geq \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \|X_i - c_{n,j}\|^2 \mathbb{1}_{[X_i \in A_j]} \\ &= D(\mu_n, q_n). \end{aligned}$$

En d'autres termes, l'opération (11.2) permet de faire décroître la distorsion du quantifieur q, en conservant sa partition mais en modifiant les centres.

Ce qui précède conduit naturellement à une technique de minimisation effective permettant d'approcher les centres optimaux $\mathbf{c}_n^\star$ définis par (11.1). Pour ce faire, on commence par choisir un jeu de k centres (plus ou moins au hasard) $\mathscr{C}^{(1)} = \{c_1^{(1)}, \dots, c_k^{(1)}\}$, en notant $\mathscr{P}_V^{(1)} = \{A_1^{(1)}, \dots, A_k^{(1)}\}$ la partition de Voronoi associée. On répète ensuite le processus en passant de $\mathscr{C}^{(\ell)} = \{c_1^{(\ell)}, \dots, c_k^{(\ell)}\}$ à $\mathscr{C}^{(\ell+1)} = \{c_1^{(\ell+1)}, \dots, c_k^{(\ell+1)}\}$ via l'itération (dite *de Lloyd*) :

$$c_j^{(\ell+1)} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{[X_i \in A_j^{(\ell)}]}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \in A_j^{(\ell)}]}}, \; 1 \leq j \leq k,$$

avec $\{A_1^{(\ell)}, \dots, A_k^{(\ell)}\}$ la partition de Voronoi à l'étape ℓ, associée à $\mathscr{C}^{(\ell)}$. On notera que l'affectation de X_i à sa cellule de Voronoi s'effectue très facilement via la recherche de son plus proche voisin dans $c_1^{(\ell)}, \dots, c_k^{(\ell)}$. Même s'il

est assuré que la distorsion décroît entre deux itérations et que l'algorithme s'arrête au bout d'un nombre d'itérations fini, rien ne garantit cependant que les centres ainsi définis soient proches des centres optimaux $\mathbf{c}_n^\star$. Il s'agit d'une méthode approchée que l'on manipulera donc avec prudence...

11.3 Consistance et vitesse

L'outil indispensable pour établir la consistance du quantifieur empirique PPV optimal défini via (11.1) est la *distance de Wasserstein*.

Définition 35. *Soient ν_1 et ν_2 des probabilités d'ordre 2 sur $\mathbb{R}^d$. La distance de Wasserstein ρ_W entre ν_1 et ν_2 est définie par*

$$\rho_W(\nu_1, \nu_2) = \inf_{X\sim\nu_1, Y\sim\nu_2} \sqrt{\mathbb{E}\|X - Y\|^2}.$$

Il s'agit d'une distance usuelle sur les mesures de probabilité. Mentionnons sans preuve deux de ses propriétés fondamentales :

Propriétés.

1. Pour ν_1, ν_2 des probabilités d'ordre 2 sur $\mathbb{R}^d$, il existe un couple de variables aléatoires (X_0, Y_0) telles que $X_0 \sim \nu_1$, $Y_0 \sim \nu_2$, et
$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$
2. Soient $(\nu_n)_n$ et ν des probabilités d'ordre 2 sur $\mathbb{R}^d$. On a $\rho_W(\nu_n, \nu) \to 0$ si
$$\nu_n \Rightarrow \nu \quad \text{et} \quad \int_{\mathbb{R}^d} \|x\|^2 \nu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \|x\|^2 \nu(\mathrm{d}x).$$

Le lien entre l'étude qui nous intéresse et la distance de Wasserstein est établi dans la proposition qui suit.

Proposition 7. *Soient ν_1 et ν_2 des probabilités d'ordre 2 sur $\mathbb{R}^d$. Si q est un quantifieur PPV, alors*

$$\left| D(\nu_1, q)^{1/2} - D(\nu_2, q)^{1/2} \right| \leq \rho_W(\nu_1, \nu_2).$$

Démonstration. Soit (X_0, Y_0) tel que $X_0 \sim \nu_1$, $Y_0 \sim \nu_2$, et

$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$

Si $q = (\mathbf{c}, \mathscr{P}_V(\mathbf{c}))$, alors :

$$\begin{aligned}
D(\nu_1, q)^{1/2} = W(\nu_1, \mathbf{c})^{1/2} &= \sqrt{\mathbb{E} \min_{1\le j\le k} \|X_0 - c_j\|^2} \\
&= \sqrt{\mathbb{E}\big(\min_{1\le j\le k} \|X_0 - c_j\|\big)^2} \\
&\le \sqrt{\mathbb{E}\Big(\min_{1\le j\le k} \big(\|X_0 - Y_0\| + \|Y_0 - c_j\|\big) \Big)^2} \\
&\le \sqrt{\mathbb{E}\Big(\|X_0 - Y_0\| + \min_{1\le j\le k} \|Y_0 - c_j\| \Big)^2} \\
&\le \sqrt{\mathbb{E}\|X_0 - Y_0\|^2} + \sqrt{\mathbb{E} \min_{1\le j\le k} \|Y_0 - c_j\|^2} \\
&\quad \text{(en utilisant l'inégalité de Cauchy-Schwarz)} \\
&= \rho_W(\nu_1, \nu_2) + D(\nu_2, q)^{1/2},
\end{aligned}$$

d'où la proposition. □

On considère à partir de maintenant le quantifieur empirique optimal PPV $q_n^\star$, défini via (11.1) par ses centres $\mathbf{c}_n^\star = (c_{n,1}^\star, \dots, c_{n,k}^\star)$:

$$q_n^\star = (\mathbf{c}_n^\star, \mathscr{P}_V(\mathbf{c}_n^\star)).$$

Théorème 20. *Le quantifieur $q_n^\star$ est consistant, i.e. $\mathbb{E}D(\mu, q_n^\star) \to D^\star(\mu)$.*

Démonstration. Si $q^\star$ est un quantifieur optimal PPV pour μ, on a, d'après la Proposition 7,

$$\begin{aligned}
&D(\mu, q_n^\star)^{1/2} - D^\star(\mu)^{1/2} \\
&\quad = \Big[D(\mu, q_n^\star)^{1/2} - D(\mu_n, q_n^\star)^{1/2}\Big] + \Big[D(\mu_n, q_n^\star)^{1/2} - D(\mu, q^\star)^{1/2}\Big] \\
&\quad \le \Big[D(\mu, q_n^\star)^{1/2} - D(\mu_n, q_n^\star)^{1/2}\Big] + \Big[D(\mu_n, q^\star)^{1/2} - D(\mu, q^\star)^{1/2}\Big] \\
&\quad \le 2\,\rho_W(\mu, \mu_n).
\end{aligned}$$

Or, $\rho_W(\mu_n, \mu) \to 0$ $\mathbb{P}$-p.s. car $\mathbb{P}(\mu_n \Rightarrow \mu) = 1$ (théorème de Varadarajan) et $\mathbb{P}$-p.s.

$$\int_{\mathbb{R}^d} \|x\|^2 \mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x)$$

(loi forte des grands nombres). On en conclut que $D(\mu, q_n^\star) \to D^\star(\mu)$ $\mathbb{P}$-p.s, i.e. $q_n^\star$ est consistant. □

Analysons maintenant la vitesse de convergence de $q_n^\star$. Pour ce faire, nous supposerons qu'il existe une constante $R \geq 0$ telle que $\|X\| \leq R$ $\mathbb{P}$-p.s. Cette hypothèse est parfois appelée *contrainte de pic* dans le vocabulaire de la quantification.

Théorème 21. *Si $\|X\| \leq R$ $\mathbb{P}$-p.s., alors*

$$\mathbb{E}D(\mu, q_n^\star) - D^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

Enonçons tout d'abord, sans preuve, un outil fondamental dans l'étude de la mesure empirique :

Lemme 9 (Principe de contraction)**.** *Soit $\sigma_1, \ldots, \sigma_n$ des variables aléatoires i.i.d. de loi de Rademacher, indépendantes de $X_1, \ldots, X_n$, et soit $\mathscr{F}$ un ensemble de fonctions réelles définies sur $\mathbb{R}^d$. Si $|\mathscr{F}| = \{|f| : f \in \mathscr{F}\}$, on a*

$$\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i |f(X_i)| \right| \leq \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(X_i) \right|.$$

Remarques préliminaires.

1. Si $\|X\| \leq R$ $\mathbb{P}$-p.s, alors les centres optimaux sont dans $B_R = B(0, R)$. En effet, si $\|c\| > R$ et p est la projection orthogonale sur B_R, alors, par définition de la projection orthogonale, on a, $\forall x \in B_R$,

$$\begin{aligned} \|x - c\|^2 &= \|x - p(c)\|^2 + \|p(c) - c\|^2 - 2\langle x - p(c), c - p(c)\rangle \\ &\geq \|x - p(c)\|^2. \end{aligned}$$

   On a donc une distorsion plus petite pour des centres dans B_R.

2. Si $X \sim \mu$, on a

$$\begin{aligned} W(\mu, \mathbf{c}) &= \mathbb{E} \min_{1 \leq j \leq k} \|X - c_j\|^2 \\ &= \mathbb{E}\|X\|^2 + \mathbb{E} \min_{1 \leq j \leq k} \left(-2\langle X, c_j \rangle + \|c_j\|^2 \right). \end{aligned}$$

Ces deux observations nous conduisent à la conclusion suivante : plutôt que de minimiser $W(\mu, \cdot)$ sur $\mathbb{R}^{dk}$, il suffit donc de minimiser, sur B_R^k,

$$\bar{W}(\mu, \mathbf{c}) = \mathbb{E} \min_{1 \le j \le k} f_{c_j}(X), \quad \text{si } f_c(x) = -2\langle x, c \rangle + \|c\|^2.$$

La même observation est valable en remplaçant μ_n au lieu de μ.

Démonstration du Théorème 21. On a

$$\begin{aligned}
&D(\mu, q_n^\star) - D^\star(\mu) \\
&\quad = W(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c} \in B_R^k} W(\mu, \mathbf{c}) \\
&\quad = \bar{W}(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu, \mathbf{c}) \\
&\quad = \left[\bar{W}(\mu, \mathbf{c}_n^\star) - \bar{W}(\mu_n, \mathbf{c}_n^\star)\right] + \left[\inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu_n, \mathbf{c}) - \inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu, \mathbf{c})\right] \\
&\quad \le 2 \sup_{\mathbf{c} \in B_R^k} \left|\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})\right| \\
&\quad = 2 \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \left| \sum_{i=1}^n \left(\min_{1 \le j \le k} f_{c_j}(X_i) - \mathbb{E} \min_{1 \le j \le k} f_{c_j}(X) \right) \right|.
\end{aligned}$$

En utilisant un n-échantillon indépendant annexe $X_1', \ldots, X_n'$ et un argument de symétrisation similaire à celui employé dans la preuve du théorème de Vapnik-Chervonenkis (Théorème 13), il vient

$$\begin{aligned}
&\mathbb{E} D(\mu, q_n^\star) - D^\star(\mu) \\
&\le 2\mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \left| \sum_{i=1}^n \left(\min_{1 \le j \le k} f_{c_j}(X_i) - \mathbb{E} \min_{1 \le j \le k} f_{c_j}(X) \right) \right| \\
&= 2\mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \left| \mathbb{E}\left[\sum_{i=1}^n \left(\min_{1 \le j \le k} f_{c_j}(X_i) - \min_{1 \le j \le k} f_{c_j}(X_i') \right) \mid X_1, \ldots, X_n \right] \right| \\
&\le 2\mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \mathbb{E}\left[\left| \sum_{i=1}^n \left(\min_{1 \le j \le k} f_{c_j}(X_i) - \min_{1 \le j \le k} f_{c_j}(X_i') \right) \right| \mid X_1, \ldots, X_n \right] \\
&\quad \text{(par l'inégalité de Jensen).}
\end{aligned}$$

Ainsi, en observant que $\sup \mathbb{E}(\cdot) \leq \mathbb{E} \sup(\cdot)$,

$$\begin{aligned}\mathbb{E}D(\mu, q_n^\star) - D^\star(\mu) &\leq 2\mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \left| \sum_{i=1}^n \big(\min_{1 \leq j \leq k} f_{c_j}(X_i) - \min_{1 \leq j \leq k} f_{c_j}(X_i') \big) \right| \\ &\leq 4\mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \min_{1 \leq j \leq k} f_{c_j}(X_i) \right|.\end{aligned}$$

Pour le traitement de ce dernier terme, nous allons procéder par itération sur k, en nous appuyant sur le principe de contraction. On note

$$S_k = \mathbb{E} \sup_{(c_1, \ldots, c_k) \in B_R^k} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \min_{1 \leq j \leq k} f_{c_j}(X_i) \right|.$$

Cas $k = 1$. Comme $\|X\| \leq R$:

$$\begin{aligned}S_1 &= \mathbb{E} \sup_{c \in B_R} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \big(-2\langle X_i, c \rangle + \|c\|^2 \big) \right| \\ &\leq 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \langle X_i, c \rangle \right| + \mathbb{E} \sup_{c \in B_R} \frac{\|c\|^2}{n} \left| \sum_{i=1}^n \sigma_i \right| \\ &\leq 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \langle X_i, c \rangle \right| + \frac{R^2}{n} \mathbb{E} \left| \sum_{i=1}^n \sigma_i \right| \\ &\leq 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \langle X_i, c \rangle \right| + \frac{R^2}{\sqrt{n}} \\ &\qquad \text{(par l'inégalité de Cauchy-Schwarz).}\end{aligned}$$

Ainsi,

$$\begin{aligned}S_1 &\leq 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \left| \left\langle \sum_{i=1}^n \sigma_i X_i, c \right\rangle \right| + \frac{R^2}{\sqrt{n}} \\ &\leq \frac{2R}{n} \mathbb{E} \left\| \sum_{i=1}^n \sigma_i X_i \right\| + \frac{R^2}{\sqrt{n}} \\ &\leq 2R \sqrt{\frac{\mathbb{E}\|X\|^2}{n}} + \frac{R^2}{\sqrt{n}} \\ &\qquad \text{(par l'inégalité de Cauchy-Schwarz)} \\ &\leq \frac{3R^2}{\sqrt{n}}.\end{aligned}$$

Cas $k = 2$**.** Comme $\min(a,b) = \frac{a+b}{2} - \frac{|a-b|}{2}$, on a

$$\begin{aligned} S_2 &= \mathbb{E} \sup_{(c_1,c_2)\in B_R^2} \frac{1}{2n} \left| \sum_{i=1}^n \sigma_i \big(f_{c_1}(X_i) + f_{c_2}(X_i) - |f_{c_1}(X_i) - f_{c_2}(X_i)| \big) \right| \\ &\leq S_1 + \mathbb{E} \sup_{(c_1,c_2)\in B_R^2} \frac{1}{2n} \left| \sum_{i=1}^n \sigma_i \big| f_{c_1}(X_i) - f_{c_2}(X_i) \big| \right|. \end{aligned}$$

En appliquant le principe de contraction, on obtient

$$\begin{aligned} S_2 &\leq S_1 + \mathbb{E} \sup_{(c_1,c_2)\in B_R^2} \frac{1}{2n} \left| \sum_{i=1}^n \sigma_i \big(f_{c_1}(X_i) - f_{c_2}(X_i) \big) \right| \\ &\leq 2S_1. \end{aligned}$$

Cas $k = 3$**.** Comme $S_2 \leq 2S_1$,

$$\begin{aligned} S_3 &\leq \frac{S_1 + S_2}{2} + \frac{S_1 + S_2}{2} \\ &\leq 3S_1. \end{aligned}$$

En itérant le procédé, on trouve

$$S_k \leq kS_1 \leq \frac{3kR^2}{\sqrt{n}}.$$

Finalement,

$$\mathbb{E}D(\mu, q_n^\star) - D^\star(\mu) \leq 4S_k \leq \frac{12kR^2}{\sqrt{n}},$$

d'où le théorème. □