Introduction aux méthodes statistiques

Marc Hoffmann

Février 2009

Table des matières

Présentation du document

Ces notes de cours présentent une introduction classique aux méthodes statistiques. Le terme « statistique(s) » reste souvent assez vague en mathématiques appliquées : il concerne aussi bien le traitement des bases de données que l'utilisation de techniques numériques en modélisation stochastique (image, économétrie et finance, physique, biologie) ; dans ce cours, il désigne plutôt une *problématique* – au sein de la théorie des probabilités – qui consiste en l'étude d'objets mathématiques bien définis : les expériences statistiques.

Nous nous plaçons dans un cadre volontairement un peu abstrait, où l'on dispose d'une notion d'expérience statistique associée à une observation dans un modèle stochastique. Le but est de dégager des méthodes quantitatives basées sur des principes relativement généraux, qui permettent de « retrouver » les paramètres d'un modèle et de « prendre des décisions » à partir d'observations issues de ce modèle. Nous voulons quantifier l'erreur de reconstruction ou de décision dans un contexte (relativement) universel, de sorte que des problèmes issus de disciplines différentes puissent être traités de la même manière, en principe. Bien entendu, chaque discipline scientifique a sa spécificité, mais nous insisterons sur des méthodes communes – par exemple le principe de maximum de vraisemblance ou la méthodes des moindres carrées – qui s'étudient de façon unifiée grâce à la théorie des probabilités.

Nous supposons le lecteur familier avec le cours de MAP 311, et nous faisons référence tout au long de ces notes au polycopié de S. Méléard [4]. On trouvera tous les compléments de probabilités éventuellement nécessaires dans le livre de J. Jacod et P. Protter [3] par exemple.

Le Chapitre 1 rappelle les principaux outils de probabilités, et insiste sur les notions fondamentales utiles en statistique : vecteurs gaussiens (lois dérivées des vecteurs gaussiens) et théorèmes limites (modes de convergence et théorème central-limite). Il permet aussi de fixer les notations utilisées dans ce cours.

Le Chapitre 2 présente la notion formelle d'expérience statistique accompagnée des exemples essentiels que sont les modèles d'échantillonnage ou de densité, et les modèles de régression.

Le Chapitre 3 étudie le modèle d'échantillonnage dans sa plus grande généralité. Nous nous posons une question apparemment naïve : si l'on observe (la réalisation) de n variables aléatoires réelles indépendantes de même loi inconnue, que peut-on dire de cette loi ? Ceci nous permet de poser les jalons des méthodes développées dans les chapitres suivants : estimation, régions et intervalles de confiance, tests, lorsque le nombre d'observations n est fixé ou bien dans la limite $n \to \infty$. Le modèle est très simple d'un point de vue probabiliste (les observations sont indépendantes et identiquement distribuées), mais très ardu d'un point de vue statistique, puisque l'on ne fait pas d'hypothèse sur la loi inconnue, et nous verrons très vite les limites de cette généralité.

Les Chapitres 4 et 5 sont consacrés aux méthodes classiques de construction d'estimateurs pour les modèles paramétriques, lorsque la loi inconnue est décrite par un paramètre de dimension finie. On se place dans les modèles de densité et régression, et on construit les estimateurs par moments, les Z- et M- estimateurs, l'estimateur du maximum de vraisemblance et l'estimateur des moindres carrés.

Le Chapitre 6 développe – dans le modèle de densité par souci de simplicité – différentes notions de comparaison d'estimateurs et la recherche d'un estimateur optimal associé à une expérience statistique. C'est un problème ancien qui remonte au programme de Fisher des années 1920, et qui n'a pas de solution totalement satisfaisante : un estimateur optimal dans un sens naïf n'existe pas, il faut faire des concessions. Si l'on suppose suffisamment de régularité (dans ce cours, nous ne rechercherons pas les hypothèses minimales), on peut néanmoins réaliser un programme d'optimalité asymptotique que nous présenterons brièvement, reposant sur le principe du maximum de vraisemblance. Il est associé à une quantité intrinsèque au modèle, l'information de Fisher, que nous étudierons en tant que telle.

Curieusement, la notion de modèle régulier en statistique est limitative : nous verrons sur des exemples que l'on estime souvent « mieux » des paramètres dans des modèles irréguliers. Mais un traitement systématique est plus difficile.

Les Chapitres 7 et 8 sont consacrés aux tests statistiques – dans un cadre non-asymptotique, puis asymptotique – et leur lien canonique avec les intervalles et régions de confiance. Si l'on accepte un certain principe (dit de Neyman) qui hiérarchise les erreurs de décision que l'on commet lorsque l'on fait un test, alors on peut dans certains cas donner une solution optimale au problème de test. On abordera les tests classiques paramétriques (Neyman-Pearson, Wald) et le test d'adéquation du χ^2, incontournable en pratique.

Les paragraphes suivis d'une étoile* pourront être omis en première lecture.

Les exercices à la fin de certains chapitres sont souvent des compléments techniques de certains aspects du cours et sont en général moins fondamentaux que les exercices proposés en P.C.

Faute de place et de temps, certains thèmes essentiels ne sont pas abordés : l'approche bayésienne, la statistique computationnelle (algorithmique statistique, bootstrap), l'estimation non-paramétrique et ses applications en débruitage de signal ou d'image, l'apprentissage et la classification, parmi bien d'autres. Nous donnons à la fin du cours quelques indications et références bibliographiques. .

Il existe par ailleurs de nombreux ouvrages qui traitent de méthodes statistiques au niveau où nous nous plaçons. Ces livres font toujours un compromis (au prix de sacrifices) entre rigueur mathématique et clarté des idées : citons deux livres emblématiques dont nous nous sommes largement inspirés : « All of Statistics » de L. Wasserman [8] qui présente beaucoup d'idées sans preuve rigoureuse et « Statistical Mathematics » de A.A. Borovkov [1], qui développe de façon systématique la théorie et qui reste un grand classique du genre. Aussi, de nombreux polycopiés sur le sujet circulent[1].

Finalement, un cours de statistique, même mathématique, ne se passe pas de `données` ou de simulations. L'accès à des quantités astronomiques de `données` est devenu facile aujourd'hui : par exemple (`www.stat.cmu.edu/~larry/all-of-statistics`) qui fournit les `données` traitées dans les exemples du livre la page de L. Wasserman [8]. Pour des `données` financières, économiques ou démographiques, (`www.economy.com/freelunch/`) ou le site de l'INSEE (`www.insee.fr`).

[1]Citons le polycopié d'A. Tsybakov de son cours à Paris 6 auquel nous avons fait de nombreux emprunts.

Première partie

Modélisation statistique

Chapitre 1

Outils de probabilités

Nous considérons des variables aléatoires à valeurs réelles ou vectorielles, discrètes ou de loi absolument continue. On envisagera (superficiellement) des cas plus complexes de mélanges de lois discrètes et continues.

1.1 Loi d'une variable aléatoire réelle

On désigne par $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilités. Les points $\omega \in \Omega$ s'interprètent comme les résultats d'une expérience aléatoire. Les objets d'intérêt sont les événements, c'est-à-dire les éléments de la tribu $\mathcal{A}$. Une variable aléatoire réelle est une application mesurable

$$X : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathcal{B}),$$

où $\mathcal{B}$ est la tribu borélienne sur $\mathbb{R}$.

Définition 1.1. *La fonction de répartition de la variable aléatoire réelle X est l'application $F : \mathbb{R} \to [0,1]$ définie par*

$$F(x) = \mathbb{P}\big[X \leq x\big] = \mathbb{P}\big[\omega \in \Omega,\ X(\omega) \leq x\big], \quad x \in \mathbb{R}.$$

La fonction F est croissante, continue à droite, tend vers 0 en $-\infty$ et vers 1 en $+\infty$. Pour tout réel x,

$$\mathbb{P}\big[X = x\big] = F(x) - F(x-).$$

La loi d'une variable aléatoire désigne d'habitude la mesure image de $\mathbb{P}$ par X sur $(\mathbb{R}, \mathcal{B})$, notée $\mathbb{P}^X$ et définie par

$$\mathbb{P}^X(A) = \mathbb{P}[X \in A], \quad A \in \mathcal{B}(\mathbb{R}).$$

Puisque la fonction de répartition F caractérise $\mathbb{P}^X$ (voir Méléard [4], Proposition 4.2.3 p. 71), on peut parler indifféremment de F ou de $\mathbb{P}^X$ pour désigner la loi de X.

Définition 1.2. *On appelle loi ou distribution de X la donnée de F.*

1.1.1 Variables discrètes

Une variable aléatoire réelle X est discrète si elle prend un ensemble de valeurs au plus dénombrable $\{x_i, i \in \mathbb{N}\} \subset \mathbb{R}$. La donnée des $\big\{\big(x_i, \mathbb{P}[X_i = x_i]\big), i \in \mathbb{N}\big\}$ détermine entièrement F (et donc caractérise la loi de X).

Remarque 1.1. Si les x_i sont isolés (par exemple si X est à valeurs dans $\mathbb{N}$ ou $\mathbb{Z}$), la fonction de répartition F de X est constante par morceaux, et les points de discontinuité de F sont les points x_i. De plus,

$$\mathbb{P}\big[X = x_i\big] = F(x_i) - F(x_i-), \quad i \in \mathbb{N}.$$

Exemple 1.1.

1. Une variable aléatoire X suit la loi de Bernoulli de paramètre $p \in [0,1]$ si

$$\mathbb{P}\left[X = 1\right] = p = 1 - \mathbb{P}\left[X = 0\right].$$

   Dans ce cas

$$F(x) = p1_{[0,1)}(x) + 1_{[1,+\infty)}(x), \quad x \in \mathbb{R}.$$

2. Une variable aléatoire X suit la loi binômiale de paramètres (n,p) avec $p \in [0,1]$ et $n \in \mathbb{N} \setminus \{0\}$ si

$$\mathbb{P}\left[X = k\right] = C_n^k p^k (1-p)^{n-k}, \quad k = 0, \ldots, n.$$

   Dans ce cas[1]

$$F(x) = \sum_{k \le x} C_n^k p^k (1-p)^{n-k}, \quad x \in \mathbb{R}.$$

3. Une variable aléatoire X suit la loi Poisson de paramètre $\lambda > 0$, si

$$\mathbb{P}\left[X = k\right] = e^{-\lambda} \tfrac{\lambda^k}{k!}, \quad k \in \mathbb{N}.$$

   Dans ce cas,

$$F(x) = e^{-\lambda} \sum_{k \le x} \frac{\lambda^k}{k!}, \quad x \in \mathbb{R}.$$

1.1.2 Variables de loi absolument continue

Une variable aléatoire réelle X est de loi absolument continue (ou à densité) si sa fonction de répartition s'écrit

$$F(x) = \int_{(-\infty,x]} f(t)dt, \quad x \in \mathbb{R}$$

[1] avec la convention $\sum_{\emptyset} = 0$.

où dt désigne la mesure de Lebesgue sur[2] $\mathbb{R}$. La fonction f, définie à un ensemble négligeable près est une densité de probabilité :

$$f \geq 0 \quad \text{et} \quad \int_{\mathbb{R}} f(t)dt = 1.$$

Dans ce cas, la fonction de répartition F de X est différentiable presque-partout et on a

$$F'(x) = f(x) \quad \text{presque-partout.}$$

Si elle existe, la densité d'une variable aléatoire détermine entièrement sa fonction de répartition F, et donc caractérise sa loi. La loi d'une variable absolument continue est diffuse : pour tout $x \in \mathbb{R}$, on a $\mathbb{P}\left[X = x\right] = 0$.

Exemple 1.2.

1. Une variable aléatoire X suit la loi uniforme sur $[a, b]$, avec $a < b$, si elle admet pour densité
$$f(t) = \frac{1}{b-a} 1_{[a,b]}(t).$$
Dans ce cas
$$F(x) = \begin{cases} 0 & \text{si} \quad x < a \\ \dfrac{x-a}{b-a} & \text{si} \quad x \in [a,b] \\ 1 & \text{si} \quad x > b. \end{cases}$$

2. Une variable aléatoire suit la loi exponentielle de paramètre $\lambda > 0$, si elle admet pour densité
$$f(t) = \lambda e^{-\lambda t} 1_{[0,+\infty)}(t).$$
Dans ce cas,
$$F(x) = \begin{cases} 0 & \text{si } \ x < 0 \\ 1 - e^{-\lambda x} & \text{sinon.} \end{cases}$$

3. Une variable aléatoire suit la loi normale de moyenne $\mu \in \mathbb{R}$ et de variance $\sigma^2 > 0$, notée $\mathcal{N}(\mu, \sigma^2)$ si elle admet pour densité
$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right).$$
Dans ce cas,
$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \ \ x \in \mathbb{R},$$
où
$$\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

[2]comprendre ici et dans toute la suite « la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B})$ ». Idem pour la mesure de Lebesgue sur $\mathbb{R}^n$, c'est-à-dire sur $(\mathbb{R}^n, \mathcal{B}^n)$, où $\mathcal{B}^n$ est la tribu des boréliens de $\mathbb{R}^n$.

1.1.3 Formules d'intégration

Si X est une variable aléatoire réelle de de loi F (ou $\mathbb{P}^X$), on a, pour toute fonction test[3] φ

$$\mathbb{E}\big[\varphi(X)\big] = \int_\Omega \varphi\big(X(\omega)\big)\mathbb{P}(d\omega) = \int_{\mathbb{R}} \varphi(x)\,\mathbb{P}^X(dx) \tag{1.1}$$

(voir Méléard s[4], Proposition 4.5.1 p. 85), dès que la fonction $\omega \rightsquigarrow \varphi\big(X(\omega)\big)$ est intégrable par rapport à la mesure $\mathbb{P}(d\omega)$. On écrit aussi

$$\int_{\mathbb{R}} \varphi(x)\mathbb{P}^X(dx) = \int_{\mathbb{R}} \varphi(x)dF(x).$$

Remarque 1.2. La mesure $\mathbb{P}^X(dx)$, définie sur $\mathbb{R}$ peut être construite à partir de la fonction de répartition F. Pour cela, on pose

$$\mathbb{P}^X\big[(a,b]\big] = F(b) - F(a), \quad \text{pour tous } \ a < b \ \text{ réels},$$

et ce qui définit $\mathbb{P}^X$ sur un sous-ensemble de $\mathcal{B}$. Le prolongement à $\mathcal{B}$ en entier se fait à l'aide du théorème de la classe monotone (voir par exemple Jacod et Protter, [3]).

Cas discret

Si X est discrète, prenant ses valeurs dans un ensemble $\{x_i,\, i \in \mathbb{N}\} \subset \mathbb{R}$ de points isolés, F est constante par morceaux, et ses discontinuités ont lieu aux points x_i où ses sauts sont d'amplitude $\mathbb{P}[X = x_i] > 0$, et

$$\int_{\mathbb{R}} \varphi(x)dF(x) = \sum_{i\in\mathbb{N}} \varphi(x_i)\mathbb{P}[X = x_i].$$

Cas continu

Si X est (de loi) absolument continue de densité f, on a

$$\int_{\mathbb{R}} \varphi(x)dF(x) = \int_{\mathbb{R}} \varphi(x)f(x)dx,$$

ce qui est cohérent du point de vue des notations avec la propriété $F'(x) = f(x)$ presque-partout.

[3] dans toute la suite, une fonction test désignera une fonction borélienne positive (ou intégrable, ou bornée) de sorte que les formules d'intégration associées soient bien définies.

Mélange de lois discrètes et continues

Une variable aléatoire réelle n'est par exclusivement discrète ou (de loi) absolument continue.

Exemple 1.3. Soit X une variable aléatoire réelle de loi $\mathcal{N}(0,1)$. La variable

$$Y = X 1_{X \geq 0}$$

n'est ni discrète, ni continue : elle n'est pas discrète puisqu'elle peut prendre toutes les valeurs positives, mais elle n'est pas (de loi) absolument continue puisque

$$\mathbb{P}[Y = 0] = \tfrac{1}{2} \neq 0.$$

La fonction de répartition de X s'écrit

$$F(x) = \tfrac{1}{2}\, 1_{x \geq 0} + \Big(\int_0^x \exp(-t^2/2) \frac{dt}{\sqrt{2\pi}} \Big)\, 1_{x \geq 0},$$

et on a[4] pour toute fonction test φ,

$$\mathbb{E}\big[\varphi(X)\big] = \int_{\mathbb{R}} \varphi(x) dF(x) = \tfrac{1}{2}\varphi(0) + \int_0^{+\infty} \varphi(t) \exp(-t^2/2) \frac{dt}{\sqrt{2\pi}}.$$

Remarque 1.3. La loi d'une variable aléatoire peut être discrète, absolument continue, ou bien encore être avoir une partie discrète et une partie absolument continue, comme dans les exemples ci-dessus. Attention : ceci n'épuise pas toutes les possibilités !

1.2 Paramètres de position

Etant donnée une variable aléatoire réelle, on cherche une description de sa loi à l'aide d'indicateurs déterministes les plus simples possibles. On utilise souvent en première approximation quatre indicateurs (s'ils existent) basés sur les quatre premiers moments (à normalisation affine près) qui sont la moyenne, la variance, le coefficient d'asymétrie – ou skewness – et le coefficient d'aplatissement – ou kurtosis –.

Un autre type d'approximation se base sur les quantiles de la loi considérée, qui mesurent dans un certain sens la dispersion de la loi. Plus difficiles à manipuler, ils présentent l'avantage d'être toujours définis.

[4] On peut aussi écrire la loi de X de la façon suivante

$$\mathbb{P}^X(dx) = \tfrac{1}{2}\delta_0(dx) + \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2} 1_{x \geq 0} dx,$$

où $\delta_0(dx)$ désigne la mesure de Dirac au point 0 et dx désigne la mesure de Lebesgue sur $\mathbb{R}$. Le contexte dictera le choix des notations.

1.2.1 Expérance-variance

Une variable aléatoire réelle X admet un moment d'ordre $p \in \mathbb{N} \setminus 0$ si

$$\mathbb{E}\big[|X|^p\big] = \int_\Omega |X(\omega)|^p \, \mathbb{P}(d\omega) < +\infty.$$

Dans ce cas, son moment d'ordre p est

$$\mathbb{E}\big[X^p\big] = \int_\Omega X(\omega)^p \, \mathbb{P}(d\omega)$$

Définition 1.3. *La moyenne ou espérance μ_X, si elle existe, est le moment d'ordre 1 de X :*

$$\mu_X = \mathbb{E}\big[X\big]$$

La variance $\mathrm{Var}[X]$ *(encore notée σ_X^2) de X, si elle existe, est le moment d'ordre 2 recentré de X :*

$$\sigma_X^2 = \mathrm{Var}[X] = \mathbb{E}\big[(X-\mu_X)^2\big] = \int_{\mathbb{R}} (x-\mu_X)^2 dF(x).$$

La racine carrée de la variance $\sigma_X = (\mathrm{Var}[X])^{1/2}$ s'appelle l'écart-type de X.

Le calcul effectif des moments se fait en utilisant la loi de X : par exemple

$$\mathbb{E}\big[X^p\big] = \int_{\mathbb{R}} x^p \, dF(x) = \begin{cases} \sum_{i\in\mathbb{N}} x_i^p \, \mathbb{P}[X = x_i] & \text{si } X \text{ est discrète} \\ \\ \int_{\mathbb{R}} x^p \, f(x) dx & \text{si } X \text{ est continue.} \end{cases}$$

La moyenne μ_X fournit la meilleure prédiction de X par une constante au sens suivant.

Proposition 1.1. *Si X admet un moment d'ordre 2, alors*

$$\mathbb{E}\big[(X-\mu_X)^2\big] = \min_{c\in\mathbb{R}} \mathbb{E}\big[(X-c)^2\big].$$

Démonstration. On a, pour tout réel c, $\mathbb{E}\big[(X-c)^2\big] = \big(\mathbb{E}\big[X\big] - c\big)^2 + \mathrm{Var}[X]$. □

Le couple espérance-variance fournit un indicateur très simple pour contrôler les fluctuations de X autour de sa moyenne μ_X via l'inégalité de Tchebychev :

$$\mathbb{P}\big[|X-\mu_X| \geq t\big] \leq \frac{\sigma_X^2}{t^2}, \;\; t > 0. \tag{1.2}$$

Famille de dilatation-translation associée à une loi

Si X a un moment d'ordre 2, écrivons a la décomposition $X = m_X + \sigma_X \xi$ où ξ est centrée-réduite, c'est-à-dire

$$\mathbb{E}[\xi] = 0, \text{ et } \mathrm{Var}[\xi] = \mathbb{E}[\xi^2] = 1.$$

Alors, avec des notations évidentes,

$$F_X(x) = F_\xi\left(\frac{x - m_X}{\sigma_X}\right), \quad x \in \mathbb{R}$$

et si X est (de loi) absolument continue, sa densité s'écrit

$$f_X(x) = \frac{1}{\sigma_X} f_\xi\left(\frac{x - m_X}{\sigma_X}\right), \quad x \in \mathbb{R}\,.$$

Plus généralement, étant donné une loi F, on peut considérer la famille de lois définies par

$$F_{\mu,\sigma}(x) = F\left(\frac{x - \mu}{\sigma}\right), \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

Les paramètres μ et σ jouent respectivement les rôles de localisation (ou translation, ou position) et de dilatation (ou d'échelle).

Remarque 1.4. Pour définir une famille de translations-dilatations associée à une loi F, il n'est pas nécessaire que cette loi admette un moment d'ordre 1 ou 2.

1.2.2 Coefficients d'asymétrie et d'aplatissement

Le coefficient d'asymétrie (skewness) et le coefficient d'aplatissement (kurtosis) correspondent, à normalisation par la moyenne et la variance près, au moment d'ordre 3 et 4.

Asymétrie (skewness)

Définition 1.4. *La loi de X est symétrique par rapport à $\mu \in \mathbb{R}$ si*

$$\forall x \in \mathbb{R}, \quad F(\mu + x) = 1 - F(\mu - x)$$

où F est la fonction de répartition de X.

Dans le cas absolument continu, si f est la densité de X, cela entraîne

$$f(\mu + x) = f(\mu - x) \quad \text{presque-partout.}$$

On dit qu'une loi est symétrique si elle est symétrique par rapport à 0.

Si X admet un moment d'ordre 3, on introduit une mesure « d'éloignement » aux distributions symétriques de la manière suivante

Définition 1.5. *Le coefficient d'asymétrie (skewness) d'une variable aléatoire réelle X telle que $\mathbb{E}\left[|X|^3\right] < +\infty$ est*

$$\alpha\big[X\big] = \frac{\mathbb{E}\left[\big(X - \mathbb{E}[X]\big)^3\right]}{\sigma_X^3}.$$

Le coefficient d'asymétrie est une mesure grossière de symétrie : si la loi de X est symétrique, alors $\alpha\big[X\big] = 0$. Mais avoir $\alpha\big[X\big] = 0$ ne signifie pas que la loi de X est symétrique.

Remarque 1.5. Le coefficient $\alpha\big[X\big]$ est invariant par dilatation-translation : pour tout $\mu \in \mathbb{R}$ et pour tout $\sigma > 0$, on a

$$\alpha\big[\mu + \sigma X\big] = \alpha\big[X\big].$$

Aplatissement (kurtosis)

Définition 1.6. *Le coefficient d'aplatissement (kurtosis) d'une variable aléatoire réelle X telle que $\mathbb{E}\left[X^4\right] < +\infty$ est*

$$\kappa\big[X\big] = \frac{\mathbb{E}\left[\big(X - \mathbb{E}[X]\big)^4\right]}{\sigma_X^4} - 3.$$

Le coefficient d'aplatissement est une mesure grossière de l'écartement de la loi de X à la loi gaussienne en terme de queues de distribution, c'est-à-dire du comportement de

$$\mathbb{P}\left[|X| \geq x\right] \text{ au voisinage de } x \to +\infty.$$

Si $X \sim \mathcal{N}(0,1)$, on a $\kappa(X) = 0$. Lorsque $\kappa\big[X\big] < 0$ on dit que les queues de distribution de la loi de X sont plus légères que les queues gaussiennes, alors qu'elles sont plus lourdes lorsque $\kappa\big[X\big] > 0$. Par l'inégalité de Cauchy-Schwarz, on a toujours $\kappa\big[X\big] \geq -2$.

Remarque 1.6. Comme pour le coefficient d'asymétrie, le coefficient d'aplatissement est invariant par dilatation-translation : pour tout $\mu \in \mathbb{R}$ et pour tout $\sigma > 0$, on a

$$\kappa\big[\mu + \sigma X\big] = \kappa\big[X\big].$$

1.2.3 Quantiles

Si X est une variable aléatoire réelle dont la fonction de répartition F est continue et strictement croissante, le quantile d'ordre p, $0 < p < 1$, de la loi F est défini comme l'unique solution q_p de l'équation

$$F(q_p) = p. \tag{1.3}$$

On a, par construction, la propriété caractéristique

$$\mathbb{P}\big[X \leq q_p\big] = p.$$

Si F n'est pas strictement croissante ou n'est pas continue, il se peut que (1.3) n'ait pas de solution ou bien ait une infinité de solutions. On peut alors modifier la définition (1.3) de la façon suivante.

Définition 1.7. *Le quantile q_p d'ordre p, $0 < p < 1$ de la loi F est la quantité*

$$q_p = \tfrac{1}{2}\big(\inf\{x,\ F(x) > p\} + \sup\{x,\ F(x) < p\}\big).$$

Si (1.3) admet une solution unique, les deux définitions coïncident. Si (1.3) n'a pas de solution, alors p n'a pas d'antécédent et q_p est un point de saut de F qui vérifie : $F(q_p-) \leq p < F(q_p)$. Si (1.3) a une infinité de solutions, alors l'ensemble de ces solutions est un intervalle borné et q_p est le milieu de cet intervalle.

Définition 1.8. *La médiane de X désigne le quantile d'ordre $1/2$ de la loi F. Les quartiles de X désignent la médiane, $q_{1/4}$ et $q_{3/4}$.*

On a toujours

$$\mathbb{P}\big[X \geq q_{1/2}\big] \geq \tfrac{1}{2}, \ \text{ et } \ \mathbb{P}\big[X \leq q_{1/2}\big] \geq \tfrac{1}{2}.$$

Si F est continue, $F_X(q_{1/2}) = \frac{1}{2}$.

Remarque 1.7. La médiane est un indicateur de localisation d'une loi de probabilité, alors que l'intervalle interquartile $q_{3/4} - q_{1/2}$ est un indicateur d'échelle. Médiane et intervalles interquartiles sont des analogues de la moyenne et de l'écart-type, et sont toujours définis.

La médiane jouit d'une propriété analogue à celle de la moyenne (Proposition 1.1) lorsque l'on remplace le moment d'ordre 2 par la valeur absolue.

Proposition 1.2. *Si X admet un moment d'ordre 1, alors*

$$\mathbb{E}\big[|X - a|\big] = \min_{c \in \mathbb{R}} \mathbb{E}\big[|X - c|\big],$$

pour tout $a \in \mathbb{R}$ vérifiant $\mathbb{P}\big[X \geq a\big] \geq \frac{1}{2}$ et $\mathbb{P}\big[X \leq a\big] \geq \frac{1}{2}$. En particulier

$$\mathbb{E}\big[|X - q_{1/2}|\big] = \min_{c \in \mathbb{R}} \mathbb{E}\big[|X - c|\big].$$

Démonstration. Montrons $\mathbb{E}\left[|X - c|\right] \geq \mathbb{E}\left[|X - a|\right]$ pour tout $c \in \mathbb{R}$. Sans perdre de généralité, on suppose $c > a$. On a alors

$$\begin{array}{lll} |X - c| = & |X - a| + (c - a) & \text{sur} \quad \{X \leq a\}, \\ |X - c| \geq & |X - a| & \text{sur} \quad \{a < X \leq (a + c)/2\}, \\ |X - c| \geq & |X - a| - (c - a) & \text{sur} \quad \{X > (a + c)/2\}. \end{array}$$

En écrivant

$$|X-c| \geq |X-a| + (c-a)1_{\{X\leq a\}} - (c-a)1_{\{X>(a+c)/2\}}$$

et en intégrant cette dernière inégalité, on obtient

$$\mathbb{E}\big[|X-c|\big] \geq \mathbb{E}\big[|X-a|\big] + (c-a)\big(\mathbb{P}\big[X\leq a\big] - \mathbb{P}\big[X>(a+c)/2\big]\big).$$

La propriété de a garantit de plus $\mathbb{P}\big[X\leq a\big] \geq \mathbb{P}\big[X>(a+c)/2\big]$, ce qui permet de conclure, puisque $\mathbb{P}\big[X>a\big] = 1-\mathbb{P}\big[X\leq a\big] \leq 1/2$. □

1.3 Vecteurs gaussiens

1.3.1 Loi normale multivariée

Préliminaires

Si

$$\boldsymbol{X} = (X_1,\ldots,X_n)^T$$

est un vecteur aléatoire de $\mathbb{R}^n$, son espérance est définie composante par composante en prenant les espérances des X_i lorsque cela a un sens.

La variance de $\boldsymbol{X}$ est la matrice

$$\Sigma\big[\,\boldsymbol{X}\,\big] = \mathbb{E}\left[(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])(X-\mathbb{E}[\boldsymbol{X}])^T\right]$$

appelée aussi matrice de variance-covariance de $\boldsymbol{X}$. Elle existe dès lors que

$$\mathbb{E}\left[\|\,\boldsymbol{X}\,\|^2\right] < +\infty,$$

où $\|\,\boldsymbol{x}\,\| = (\boldsymbol{x}^T\,\boldsymbol{x})^{1/2}$ est la norme euclidienne du vecteur $\boldsymbol{x}\in\mathbb{R}^n$. On a les propriétés suivantes :

1. $\Sigma\big[\,\boldsymbol{X}\,\big] = \mathbb{E}\left[\,\boldsymbol{X}^T\,\boldsymbol{X}\,\right] - \mathbb{E}\left[\,\boldsymbol{X}\,\right]\mathbb{E}\left[\,\boldsymbol{X}\,\right]^T$
2. Pour tout $a\in\mathbb{R}^n$, $\mathrm{Var}\big[a^T\,\boldsymbol{X}\,\big] = a^T\Sigma\big[\,\boldsymbol{X}\,\big]a$. En particulier, $\Sigma\big[\,\boldsymbol{X}\,\big]$ est symétrique positive.
3. Si A est une matrice $k\times n$ et $b\in\mathbb{R}^k$, on a $\Sigma\big[A\,\boldsymbol{X}+b\big] = A\Sigma\big[\,\boldsymbol{X}\,\big]A^T$.

Vecteurs gaussiens

Si Id_n désigne la matrice unité $n\times n$, on note

$$\mathcal{N}(0,\mathrm{Id}_n)$$

la loi du vecteur aléatoire

$$\boldsymbol{X} = (\xi_1,\ldots,\xi_n)^T$$

dont toutes les composantes sont des variables aléatoires gaussiennes indépendantes, centrées réduites. On écrit $\boldsymbol{X} \sim \mathcal{N}(0, \mathrm{Id}_n)$.

On a les propriétés suivantes :

1. La moyenne de $\boldsymbol{X}$ est 0 et sa matrice de variance-covariance est Id_n.
2. La loi de $\boldsymbol{X}$ est absolument continue, de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ donnée par

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\,\boldsymbol{x}^T\,\boldsymbol{x}\right), \quad \boldsymbol{x} \in \mathbb{R}^n\,.$$

3. La fonction caractéristique (voir Méléard [4], Définition 6.1. p. 125) de $\boldsymbol{X}$ est donnée par

$$\phi_{\boldsymbol{X}}(a) = \mathbb{E}\left[e^{ia^T\boldsymbol{X}}\right] = \exp\left(-\frac{1}{2}a^T a\right), \quad a \in \mathbb{R}^n\,.$$

Définition 1.9. *Un vecteur aléatoire $\boldsymbol{X}$ à valeurs dans $\mathbb{R}^n$ est gaussien (ou normal) s'il existe A une matrice $n \times n$, et un vecteur $\boldsymbol{\mu} \in \mathbb{R}^n$ tels que*

$$\boldsymbol{X} = \boldsymbol{\mu} + A\,\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(0, Id_n).$$

On a les propriétés suivantes :

1. La moyenne (vectorielle) de $\boldsymbol{X}$ est $\mathbb{E}\left[\,\boldsymbol{X}\,\right] = \boldsymbol{\mu}$.
2. La matrice de covariance de $\boldsymbol{X}$ est $\Sigma_{\boldsymbol{X}} = \mathrm{Var}\left[\,\boldsymbol{X}\,\right] = AA^T$.
3. La fonction caractéristique de $\boldsymbol{X}$ vaut

$$\begin{aligned}
\phi_{\boldsymbol{X}}(a) &= \mathbb{E}\left[e^{ia^T\boldsymbol{X}}\right] \\
&= \mathbb{E}\left[e^{ia^T(\boldsymbol{\mu}+A\boldsymbol{\xi})}\right] \\
&= \exp\left(ia^T\boldsymbol{\mu}\right)\mathbb{E}\left[e^{i(A^Ta)^T\boldsymbol{\xi}}\right] \\
&= \exp\left(ia^T\boldsymbol{\mu} - \tfrac{1}{2}(a^TA)^Ta^TA\right) \\
&= \exp\left(ia^T\boldsymbol{\mu} - \tfrac{1}{2}a^T\Sigma a\right), \quad a \in \mathbb{R}^n\,.
\end{aligned}$$

On a la caractérisation suivante d'un vecteur gaussien :

Proposition 1.3. *Une application $\phi : \mathbb{R}^n \to \mathbb{C}$ est la fonction caractéristique d'un vecteur gaussien si et seulement si il existe $\boldsymbol{\mu} \in \mathbb{R}^n$ et une matrice Σ symétrique positive (dont toutes les valeurs propres sont positives ou nulles) tels que*

$$\phi(a) = \exp\left(ia^T\boldsymbol{\mu} - \tfrac{1}{2}a^T\Sigma a\right), \quad a \in \mathbb{R}^n\,.$$

Démonstration. Le calcul de la fonction caractéristique d'un vecteur gaussien établi plus haut monte que la condition est nécessaire. Pour montrer la condition suffisante, il suffit d'exhiber un vecteur gaussien de $\mathbb{R}^n$ dont ϕ est la fonction caractéristique. Pour cela, on peut poser $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{\xi}$, où $\Sigma^{1/2}$ est une racine carrée de Σ et $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathrm{Id}_n)$. □

En conséquence, la loi d'un vecteur gaussien $\boldsymbol{X}$ est entièrement déterminée par sa moyenne $\boldsymbol{\mu}$ et sa matrice de covariance Σ. On écrira par la suite $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$.

Remarque 1.8. Dans la décomposition $\Sigma = A^T A$ d'une matrice symétrique positive, la matrice A n'est pas unique. On peut prendre pour A une racine carrée de Σ, mais il existe aussi d'autres choix où A n'est pas nécessairement symétrique. Si Λ désigne la matrice diagonale formée à partir des valeurs propres λ_j de Σ, de rang $k \leq n$ alors, on a la décomposition

$$\Sigma = \Gamma\Lambda\Gamma^T = \sum_{j=1}^{n} \boldsymbol{\gamma}_{\bullet,j}\lambda_j\boldsymbol{\gamma}_{\bullet,j}^T = \sum_{i=1}^{k} \boldsymbol{a}_{\bullet,j}\boldsymbol{a}_{\bullet,j}^T = AA^T$$

où les $\boldsymbol{\gamma}_{\bullet,j}$ sont les colonnes de Γ, $\boldsymbol{a}_j = \sqrt{\lambda_j}\boldsymbol{\gamma}_{\bullet,j}$ et A est une matrice $n \times n$ définie par $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k, 0 \ldots, 0)$.

Une caractérisation équivalente de la loi d'un vecteur gaussien est la suivante :

Proposition 1.4. *Un vecteur aléatoire $\boldsymbol{X}$ est gaussien si et seulement si toute combinaison linéaire des composantes de $\boldsymbol{X}$ est une variable aléatoire gaussienne réelle*[5].

Démonstration. Si $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, pour tout $u \in \mathbb{R}$, on a

$$\begin{aligned}\phi_{a^T\boldsymbol{X}}(u) &= \mathbb{E}\left[e^{ia^T\boldsymbol{X}u}\right] \\ &= \phi_{\boldsymbol{X}}(ua) \\ &= \exp\left(iua^T\boldsymbol{\mu} - \tfrac{1}{2}u^2a^T\Sigma a\right),\end{aligned}$$

donc $a^T\boldsymbol{X} \sim \mathcal{N}(a^T\boldsymbol{\mu}, a^T\Sigma a)$. Réciproquement, si pour tout $a \in \mathbb{R}^n$, la variable aléatoire réelle $a^T\boldsymbol{X}$ est gaussienne, alors $\mathbb{E}\left[\|\boldsymbol{X}\|^2\right] < +\infty$ (prendre pour a les projections sur les coordonnées), donc $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{X}\right]$ et $\Sigma = \Sigma\big[X\big]$ existent. Soit $a \in \mathbb{R}^n, m \in \mathbb{R}$ et $s^2 \geq 0$ de sorte que $a^T\boldsymbol{X} \sim \mathcal{N}(m, s^2)$. Nécessairement,

$$m = a^T\boldsymbol{\mu} \quad \text{et} \quad s^2 = a^T\Sigma a,$$

[5]On admet dans cette terminologie qu'une constante est une variable aléatoire gaussienne, de moyenne elle-même et de variance 0.

par linéarité de l'espérance et parce que $\mathrm{Var}\big[a^T \boldsymbol{X}\big] = a^T \Sigma[\boldsymbol{X}] a$ (voir le paragraphe précédent). Donc

$$\begin{aligned}\phi_{a^T \boldsymbol{X}}(u) &= \exp\big(imu - \tfrac{1}{2}s^2u^2\big)\\ &= \exp\big(iua^T\boldsymbol{\mu} - \tfrac{1}{2}u^2 a^T\Sigma a\big)\\ &= \phi_{a^T \boldsymbol{X}}(1)\\ &= \phi_{\boldsymbol{X}}(a).\end{aligned}$$

Puisque le choix de $a \in \mathbb{R}^n$ est arbitraire, on a la conclusion. □

Densité de la loi normale multivariée

Si Σ est définie positive, la loi de $\boldsymbol{X}$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, et la densité du vecteur $\boldsymbol{X}$ est obtenue à partir de la densité de $\boldsymbol{\xi}$ via la représentation $\boldsymbol{X} = \boldsymbol{\mu} + A\boldsymbol{\xi}$ par changement de variable affine (Méléard [4], paragraphe 4.10.2 p. 107) :

$$\begin{aligned}f_{\boldsymbol{X}}(\boldsymbol{x}) &= \det A^{-1} f_{\boldsymbol{\xi}}\big(A^{-1}(\boldsymbol{x} - \mu)\big)\\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right), \quad \boldsymbol{x} \in \mathbb{R}^n.\end{aligned}$$

Loi normale multivariée dégénérée

Si Σ est singulière, soit $\mathrm{Rang}(\Sigma) = k < n$, le vecteur $\boldsymbol{X}$ n'a plus de densité sur $\mathbb{R}^n$. La représentation $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{\xi}$ montre que $\boldsymbol{X}$ se concentre à une transformation affine près sur l'image de $\Sigma^{1/2}$, qui est un sous-espace de dimension k.

Proposition 1.5. *Si $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$, avec $\mathrm{Rang}(\Sigma) = k < n$, alors il existe un sous-espace vectoriel $H \subset \mathbb{R}^n$ de dimension $n - k$ tel que pour tout $a \in H$, la loi de $a^T \boldsymbol{X}$ est dégénérée, c'est-à-dire $a^T \boldsymbol{X}$ est une constante (déterministe).*

Démonstration. On pose $H = \mathrm{Ker}(\Sigma)$. Alors H est de dimension $n - k$ et si $a \in H$, pour tout $u \in \mathbb{R}^n$, on a

$$\begin{aligned}\phi_{a^T bX}(u) &= \mathbb{E}\left[e^{iu\,a^T \boldsymbol{X}}\right]\\ &= \exp\big(iu\,a^T\boldsymbol{\mu} - \tfrac{1}{2}u^2 a^T\Sigma a\big)\\ &= \exp\big(iu\,a^T\boldsymbol{\mu}\big) \quad \text{puisque} \quad \Sigma a = 0.\end{aligned}$$

□

Indépendance de deux vecteurs gaussiens

Si $\boldsymbol{X}$ et $\mathbf{Y}$ sont deux vecteurs aléatoires à valeurs dans $\mathbb{R}^p$ et $\mathbb{R}^q$ respectivement, et tels que $\mathbb{E}\left[\| \boldsymbol{X} \|^2\right] > +\infty$ et $\mathbb{E}\left[\|\mathbf{Y}\|^2\right] < +\infty$, leur matrice de covariance est la matrice $p \times q$ définie par

$$\Sigma\big[\boldsymbol{X}, \mathbf{Y}\big] = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T\right].$$

L'indépendance entre des transformations linéaires d'un vecteur gaussien se lit sur la matrice de covariance :

Proposition 1.6. *Si $\boldsymbol{X}$ est un vecteur gaussien de $\mathbb{R}^n$ et si A et B sont deux matrices $n \times p$ et $n \times q$, alors les vecteurs $A\,\boldsymbol{X}$ et $B\,\boldsymbol{X}$ sont indépendants si et seulement si*

$$\Sigma\big[A\,\boldsymbol{X}, B\,\boldsymbol{X}\big] = 0.$$

Démonstration. On concatène $A\,\boldsymbol{X}$ et $B\,\boldsymbol{X}$ en un vecteur $\mathbf{Y} = (A\,\boldsymbol{X}, B\,\boldsymbol{X})^T$ de $\mathbb{R}^{p+q}$ qui est gaussien comme transformation linéaire du vecteur gaussien $\boldsymbol{X}$. On a

$$\Sigma_{\mathbf{Y}} = \begin{pmatrix} \Sigma_{A\,\boldsymbol{X}} & \Sigma\big[A\,\boldsymbol{X}, B\,\boldsymbol{X}\big] \\ \Sigma\big[A\,\boldsymbol{X}, B\,\boldsymbol{X}\big] & \Sigma_{B\,\boldsymbol{X}} \end{pmatrix} = \begin{pmatrix} \Sigma_{A\,\boldsymbol{X}} & 0 \\ 0 & \Sigma_{B\,\boldsymbol{X}} \end{pmatrix}$$

si $\Sigma\big[A\,\boldsymbol{X}, B\,\boldsymbol{X}\big] = 0$. Il vient, pour $u = (a, b) \in \mathbb{R}^p \times \mathbb{R}^q$,

$$\begin{aligned} \phi_{\mathbf{Y}}(u) &= \phi_{\mathbf{Y}}(a, b) \\ &= \exp\left(ia^T\,\mathbb{E}[A\,\boldsymbol{X}] + b^T\,\mathbb{E}[B\,\boldsymbol{X}] - \tfrac{1}{2}(a^T, b^T)\Sigma_{\mathbf{Y}}(a, b)^T\right) \\ &= \exp\left(ia^T\,\mathbb{E}[A\,\boldsymbol{X}] - \tfrac{1}{2}a^T\Sigma_{A\,\boldsymbol{X}}a + ib^T\,\mathbb{E}[B\,\boldsymbol{X}] - \tfrac{1}{2}b^T\Sigma_{B\,\boldsymbol{X}}b\right) \\ &= \phi_{\boldsymbol{X}}(a)\phi_{\boldsymbol{X}}(b). \end{aligned}$$

Réciproquement, si $A\,\boldsymbol{X}$ et $B\,\boldsymbol{X}$ sont indépendants, on a $\Sigma\big[A\,\boldsymbol{X}, B\,\boldsymbol{X}\big] = 0$ par le même calcul. □

1.3.2 Dérivées des lois gaussiennes

Il s'agit de trois familles de lois très classiques en statistique – et utilisées pour la construction de tests et d'intervalles de confiance – obtenues comme transformation de lois gaussiennes : loi du χ^2, loi de Student et loi de Fisher-Snedecor.

Loi du χ^2 à n degrés de liberté

Définition 1.10. *Une variable aléatoire réelle Y suit la loi du χ^2 à n degrés de liberté si elle peut s'écrire*

$$Y = \sum_{i=1}^{n} X_i^2,$$

où les variables $X_1, \dots, X_n$ sont indépendantes, de même loi $\mathcal{N}(0,1)$.

On écrit $Y \sim \chi^2(n)$. Autrement dit, si $\boldsymbol{X} \sim \mathcal{N}(0, \mathrm{Id}_n)$, alors $\| \boldsymbol{X}^2 \| \sim \chi^2(n)$. On a les propriétés suivantes :

1. La densité de la loi du $\chi^2(n)$ est donnée par
$$y \rightsquigarrow c(n) y^{n/2-1} e^{-y/2}, \quad y \in \mathbb{R}_+ \setminus \{0\}$$
avec $c(n) = 2^{-n/2}\Gamma(n/2)^{-1}$ et $\Gamma(x) = \int_0^{+\infty} u^{x-1}e^{-u/2}du$.
2. Si $Y \sim \chi^2(n)$, on a $\mathbb{E}\left[Y\right] = n$ et $\mathbb{E}\left[Y^2\right] = 2n$.
3. Si $Y \sim \chi^2(n)$, sa transformée de Laplace[6] est donnée par
$$\mathcal{L}(u) = \mathbb{E}\left[e^{-uY}\right] = \left(\frac{1}{1-2u}\right)^n.$$

On utilise souvent le résultat suivant :

Proposition 1.7. *Soit $\boldsymbol{X}$ un vecteur aléatoire de $\mathbb{R}^n$ tel que $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, où Σ est définie positive. Alors*
$$(\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2(n).$$

Démonstration. On a
$$(\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = \|\Sigma^{-1/2}\boldsymbol{X} - \boldsymbol{\mu}\|^2.$$
On conclut en utilisant : $\Sigma^{-1/2}\boldsymbol{X} - \boldsymbol{\mu} \sim \mathcal{N}(0, \mathrm{Id}_n)$. □

Loi t de Student

Définition 1.11. *Une variable aléatoire réelle T suit la loi de Student à n degré de libertés si*
$$T = \frac{\xi}{\sqrt{Y/n}},$$
où $\xi \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2(n)$ sont indépendantes.

On écrit $T \sim \mathfrak{t}(n)$. On a les propriétés suivantes

1. La densité de la loi $\mathfrak{t}(n)$ est donnée par
$$y \rightsquigarrow c(n)\left(1 + \frac{y^2}{n}\right)^{-(n+1)/2}, \quad y \in \mathbb{R}$$
avec
$$c(n) = \frac{1}{\sqrt{n}B(1/2, n/2)}, \quad \text{et} \quad B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q).$$

[6] puisque Y est à valeurs dans $\mathbb{R}_+$, on utilise sa transformée de Laplace plutôt que sa fonction caractéristique.

2. La loi $\mathfrak{t}(n)$ est symétrique.
3. La loi $\mathfrak{t}(1)$ est la loi de Cauchy.
4. Lorsque n est grand, Y/n est proche de 1 par la loi des grands nombres et la loi $\mathfrak{t}(n)$ se « rapproche » de la loi $\mathcal{N}(0,1)$.

La loi $\mathfrak{t}$ de Student intervient en statistique comme une approximation de la loi $\mathcal{N}(0,1)$, lorsque la variance 1 est approchée par une loi du χ^2 à n degrés de liberté renormalisée.

Remarque 1.9. Par cette approximation même, la loi $\mathfrak{t}(n)$ est plus « dispersée » que la loi $\mathcal{N}(0,1)$: si $T \sim \mathfrak{t}(n)$ et $\xi \sim \mathcal{N}(0,1)$, on a, par exemple,

$$\kappa\big[T\big] > \kappa\big[X\big],$$

où $\kappa[\bullet]$ est le coefficient d'aplatissement (la kurtosis) défini dans la Section 1.2. Le cas extrême est $n = 1$ où la kurtosis n'est même pas définie (il faut prendre au moins $n = 6$).

Loi de Fisher-Snedecor

Définition 1.12. *Une variable aléatoire Y suit la loi de Fisher-Snedecor de degrés de libertés (p,q) si*

$$Y = \frac{U/p}{V/q},$$

où $U \sim \chi^2(p)$ et $V \sim \chi^2(q)$ sont indépendantes.

On écrit $Y \sim F_{p,q}$ et on a les propriétés suivantes :

1. La densité de la loi $F_{p,q}$ est donnée par

$$y \rightsquigarrow c(p,q)\frac{y^{p/2-1}}{(q+py)^{(p+q)/2}}, \quad y \in \mathbb{R}_+ \setminus \{0\},$$

où

$$c(p,q) = \frac{p^{p/2}q^{q/2}}{B(p/2,q/2)}.$$

2. Lorsque q est grand, la loi $F(p,q)$ se rapproche de la loi du $\chi^2(p)$. C'est le même raisonnement que pour la loi de Student.

1.3.3 Cochran

Il s'agit d'un résultat d'algèbre linéaire que l'on utilise pour déduire des propriétés de transformations linéaires de vecteurs gaussiens.

Théorème 1.1 (Cochran)**.** *Soit $\boldsymbol{X} \sim \mathcal{N}(0, \mathrm{Id}_n)$ et $A_1, \ldots, A_J$ des matrices $n \times n$ telles que $\sum_{j=1}^{J} \mathrm{Rang}(A_j) \leq n$ et vérifiant*

(i) les A_j sont symétriques,
(ii) $A_j A_k = 0$ si $j \neq k$ et $A_j^2 = A_j$.

Alors

1. *Les vecteurs aléatoires $(A_j \boldsymbol{X}, j = 1, \ldots, J)$ sont mutuellement indépendants, et $A_j \boldsymbol{X} \sim \mathcal{N}(0, A_j)$.*
2. *Les variables aléatoires $(\|A_j \boldsymbol{X}\|^2, j = 1, \ldots, J)$ sont mutuellement indépendantes et $\|A_j \boldsymbol{X}\|^2 \sim \chi^2\big(\mathrm{Rang(A_j)}\big)$.*

Démonstration. On a, pour tout $u \in \mathbb{R}^n$ et $j = 1, \ldots, J$

$$\begin{aligned}
\mathbb{E}\left[e^{iu^T A_j \boldsymbol{X}}\right] &= \mathbb{E}\left[e^{i(A_j^T u)^T \boldsymbol{X}}\right] \\
&= \exp\big(-\tfrac{1}{2}(A_j^T u)^T A_j^T u\big) \\
&= \exp\big(-\tfrac{1}{2} u^T A_j^2 u\big) \quad \text{par } (i) \\
&= \exp\big(-\tfrac{1}{2} u^T A_j u\big) \quad \text{par } (ii).
\end{aligned}$$

On a donc $A_j \boldsymbol{X} \sim \mathcal{N}(0, A_j)$. Soient $u_1, \ldots, u_J \in \mathbb{R}^n$. On a

$$\begin{aligned}
\mathbb{E}\left[e^{i\sum_{j=1}^J u_j^T A_j \boldsymbol{X}}\right] &= \mathbb{E}\left[e^{i(\sum_{j=1}^J A_j^T u_j)^T \boldsymbol{X}}\right] \\
&= \exp\big[-\tfrac{1}{2}\big(\sum_{j=1}^J A_j^T u_j\big)^T\big(\sum_{j=1}^J A_j^T u_j\big)\big] \\
&= \exp\big[-\tfrac{1}{2}\big(\sum_{j=1}^J A_j^T u_j\big)^T\big(\sum_{j=1}^J A_j u_j\big)\big] \quad \text{par } (i) \\
&= \exp\big(-\tfrac{1}{2}\sum_{j,j'=1}^J u_j^T A_j A_{j'} u_{j'}\big) \\
&= \exp\big(-\tfrac{1}{2}\sum_{j=1}^J u_j^T A_j A_j u_j\big) \quad \text{par } (ii) \\
&= \prod_{j=1}^J \exp\big(-\tfrac{1}{2}(A_j^T u_j)^T A_j^T u_j\big) \quad \text{par } (i) \\
&= \prod_{j=1}^J \mathbb{E}\left[e^{iu_j^T A_j \boldsymbol{X}}\right]
\end{aligned}$$

ce qui entraîne l'indépendance (Méléard [4], Proposition 6.1.4 p. 130) des $A_j \boldsymbol{X}$. Pour montrer le point 2 du théorème, on écrit, pour j fixé,

$$A_j = \Gamma \Lambda \Gamma^T$$

où Γ est une matrice orthogonale et $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$ est la matrice diagonale des valeurs propres de A_j. Il vient

$$\|A_j \boldsymbol{X}\|^2 = \boldsymbol{X}^T A_j^T A_j \boldsymbol{X} = \boldsymbol{X}^T A_j \boldsymbol{X} = (\Gamma^T \boldsymbol{X})^T \Lambda \Gamma^T \boldsymbol{X}\,. \tag{1.4}$$

par (i) et (ii). Posons $\mathbf{Y} = \Gamma^T \boldsymbol{X}$. On a $\mathbf{Y} \sim \mathcal{N}(0, \mathrm{I}d_n)$ car Γ est orthogonale. En réécrivant (1.4) à l'aide de $\mathbf{Y}$, on en déduit

$$\|A_j \boldsymbol{X}\|^2 = \mathbf{Y}^T \Lambda \mathbf{Y} = \sum_{i=1}^{n} \lambda_i Y_i^2 \sim \chi^2\big(\mathrm{Rang}(A_i)\big)$$

puisque A_i est un projecteur, donc $\lambda_i = 0$ ou 1 et le nombre de λ_i non nuls est le rang de A_i. L'indépendance des $\|A_j \boldsymbol{X}\|^2$ est une conséquence immédiate de celle des $A_j \boldsymbol{X}$ prouvée précédemment. □

1.4 Convergences et théorèmes limites

1.4.1 Modes de convergences

On considère une suite $(\xi_n)_n$ de variables aléatoires réelles ξ_n définies sur un espace de probabilité commun $(\Omega, \mathcal{A}, \mathbb{P})$.

Définition 1.13. *La suite $(\xi_n)_n$ ou plus simplement ξ_n converge vers ξ en probabilité (notation : $\xi_n \xrightarrow{\mathbb{P}} \xi$) si pour tout $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left[|\xi_n - \xi| \geq \varepsilon\right] = 0.$$

La suite ξ_n converge vers ξ presque-sûrement (notation : $\xi_n \xrightarrow{\text{p.s.}} \xi$) si

$$\mathbb{P}\left[\limsup_{n\to\infty} |\xi_n - \xi| > 0\right] = 0.$$

La suite ξ_n converge vers ξ dans $\mathcal{L}^p$ (notation : $\xi_n \xrightarrow{\mathcal{L}^p} \xi$), avec $0 < p < \infty$, si

$$\lim_{n\to\infty} \mathbb{E}\left[|\xi_n - \xi|^p\right] = 0.$$

On a les propriétés suivantes :

1. La convergence presque-sûre ou la convergence dans $\mathcal{L}^p$ entraînent la convergence en probabilité.
2. La convergence presque-sûre et la convergence dans $\mathcal{L}^p$ ne sont pas comparables.
3. Si $\xi_n \xrightarrow{\mathbb{P}} \xi$, elle admet une sous suite qui converge presque-sûrement.
4. Si $\xi_n \xrightarrow{\mathbb{P}} \xi$ et si $|\xi_n| \leq \eta$, avec $\mathbb{E}\left[\eta^p\right] < +\infty$ pour un $p > 0$, alors alors $\xi_n \xrightarrow{\mathcal{L}^p} \xi$.

5. Si f est continue et $\xi_n \xrightarrow{\mathbb{P}} \xi$, alors

$$f(\xi_n) \xrightarrow{\mathbb{P}} f(\xi).$$

Pour parler de convergence presque-sûre, il est nécessaire que les variables ξ_n et leur limite soient définies simultanément sur le même espace de probabilité. [7]

Remarque 1.10. La convergence en probabilité est sans doute la notion la plus adaptée à la problématique statistique. Elle traduit la propriété suivante : pour tout niveau de risque $\alpha > 0$ et pour toute précision $\varepsilon > 0$, il existe un rang $n(\varepsilon, \alpha)$ à partir duquel on peut « affirmer » que ξ_n approche ξ avec une erreur inférieure à ε. La probabilité que cette affirmation soit fausse est inférieure à α :

$$\text{pour } \; n \geq n(\varepsilon, \alpha), \quad \mathbb{P}\left[|\xi_n - \xi| \leq \varepsilon\right] \geq 1 - \alpha.$$

Cependant, pour contrôler précisément le comportement asymptotiques de suites de variables aléatoires, on aura besoin d'un mode de convergence plus faible : la convergence en loi.

Définition 1.14. *La suite ξ_n converge vers ξ en loi (notation $\xi_n \xrightarrow{d} \xi$) si pour toute fonction φ continue bornée, on a*

$$\mathbb{E}\left[\varphi(X_n)\right] \to \mathbb{E}\left[\varphi(\xi)\right] \quad \textit{lorsque} \quad n \to \infty.$$

[7] **Remarque** (qu'on omettra en première lecture) : Ce n'est pas forcément le cas pour la convergence dans $\mathcal{L}^p$ ou en probabilité. Dans les chapitres qui suivront, on travaillera souvent avec une suite de variables aléatoires réelles

$$X_1, \ldots, X_n$$

indépendantes, et identiquement distribuées de loi $\mathbb{Q}$ sur $(\mathbb{R}, \mathcal{B})$. On utilisera la construction suivante : pour chaque n, on pose

$$\Omega_n = \mathbb{R}^n, \; \mathcal{A}^n = \mathcal{B}^n, \; \mathbb{P}_n = \mathbb{Q} \otimes \ldots \otimes \mathbb{Q} \quad n - \text{fois}.$$

On peut ainsi définir $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ sur $(\Omega_n, \mathcal{A}^n)$ et la loi $\mathbb{P}^{\boldsymbol{X}}$ du vecteur $\boldsymbol{X}$ coïncide avec $\mathbb{P}_n$. Si on considère une suite de variable aléatoires de la forme $\xi_n = \phi_n(X_1, \ldots, X_n)$, où $\phi_n : \mathbb{R}^n \to \mathbb{R}$ est une application donnée, chaque ξ_n est définie sur un espace différent $(\Omega_n, \mathcal{A}^n, \mathbb{P}_n)$. Si la « limite » de ξ_n est une constante $c \in \mathbb{R}$ déterministe, ce qui sera souvent le cas, alors on peut parfaitement parler de convergence en probabilité et dans $\mathcal{L}^p$ en posant

$$\xi_n \xrightarrow{\mathbb{P}_n} c \quad \text{si} \quad \forall \varepsilon > 0, \; \lim_{n \to \infty} \mathbb{P}_n \left[|\xi_n - c| \geq \varepsilon\right] = 0$$

et

$$\xi_n \xrightarrow{\mathcal{L}(\mathbb{P}_n)} c \quad \text{si} \quad \lim_{n \to \infty} \mathbb{E}_n \left[|\xi_n - c|^p\right] = 0.$$

Puisque $\mathbb{P}_n$ est entièrement déterminée par $\mathbb{Q}$, on écrira, sans qu'il y ait de confusion possible,

$$\xi_n \xrightarrow{\mathbb{Q}} c \quad \text{ou} \quad \xi_n \xrightarrow{\mathcal{L}^p(\mathbb{Q})} c.$$

Par contre, on ne peut plus parler de convergence presque-sûre. Toutefois, en travaillant un peu, on peut se placer sur un produit infini et donner de même un sens à la convergence presque-sûre. *A posteriori* il n'y a pas d'ambiguité d'écriture. Nous ne reviendrons plus sur ces questions techniques.

Remarque 1.11. On peut remplacer dans la définition la suite réelle ξ_n par une suite de vecteurs aléatoires $\boldsymbol{\xi}_n$ de $\mathbb{R}^d$ avec $d \geq 1$ et ξ par un vecteur aléatoire $\boldsymbol{\xi}$ de $\mathbb{R}^d$.

La convergence en loi est une notion plus faible que la convergence en probabilité. Elle ne fait intervenir que la suite des lois $\mathbb{P}^{\xi_n}$ et $\mathbb{P}^{\xi}$. En particulier, on n'a pas besoin que les variables ξ_n ou la limite ξ soient définies sur le même espace de probabilité.

On a les propriétés suivantes

1. $\xi_n \xrightarrow{d} \xi$ si et seulement si pour tout $u \in \mathbb{R}$,

$$\phi_{\xi_n}(u) \to \phi_{\xi}(u) \quad \text{lorsque } n \to \infty.$$

   Cette propriété caractérise la convergence en loi[8] (Théorème de Lévy).

2. (Astuce de Wold). La suite de vecteurs $\boldsymbol{\xi}_n$ de $\mathbb{R}^d$ converge vers $\boldsymbol{\xi}$ en loi si et seulement si $a^T\boldsymbol{\xi} \xrightarrow{d} a^T\boldsymbol{\xi}$ pour tout $a \in \mathbb{R}^d$.

3. Dans la Définition 1.14, on peut remplacer f continue bornée par

$$f(x) = 1_{(-\infty, x_0]}(x), \quad x \in \mathbb{R}$$

   en tous les points $x_0 \in \mathbb{R}$ tels que $\mathbb{P}\left[\xi = x_0\right] = 0$. Autrement dit $\xi_n \xrightarrow{d} \xi$ si et seulement si

$$\mathbb{P}\left[\xi_n \leq x\right] \to \mathbb{P}\left[\xi \leq x\right], \quad \text{lorsque } n \to \infty.$$

   en tout point x où la fonction de répartition de ξ est continue.

4. Si $\xi_n \xrightarrow{d} \xi$ et $g : \mathbb{R} \to \mathbb{R}$ est continue, alors[9] $g(\xi_n) \xrightarrow{d} g(\xi)$.

Voici un résultat technique que nous utiliserons constamment dans ce cours :

Proposition 1.8 (Slutsky)**.** *Si $\xi_n \xrightarrow{d} \xi$ et $\eta_n \xrightarrow{\mathbb{P}} c$ où c est une constante (déterministe), alors*

$$(\xi_n, \eta_n) \xrightarrow{d} (\xi, c).$$

En particulier, si $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est continue, alors $h(\xi_n, \eta_n) \xrightarrow{d} h(\xi, c)$. Ceci entraîne alors $\xi_n + \eta_n \xrightarrow{d} \xi + c$, $\eta_n \xi_n \xrightarrow{d} c\,\xi$, et ainsi de suite.

Démonstration. Soient $u, v \in \mathbb{R}$. On écrit

$$\begin{aligned} &\mathbb{E}\left[e^{i(u\xi_n + v\eta_n)}\right] - \mathbb{E}\left[e^{iu\xi}\right]e^{ivc} \\ &= \mathbb{E}\left[e^{iu\xi_n}\left(e^{iv\eta_n} - e^{ivc}\right)\right] + \left(\mathbb{E}\left[e^{iu\xi_n}\right] - \mathbb{E}\left[e^{iu\xi}\right]\right)e^{ivc}. \end{aligned}$$

[8] On peut remplacer ξ_n et ξ par des vecteurs de $\mathbb{R}^d$ avec $d \geq 1$, en prenant $u \in \mathbb{R}^d$.
[9] On peut remplacer ξ_n et ξ par des vecteurs de $\mathbb{R}^d$ avec $d \geq 1$ et $g : \mathbb{R}^d \to \mathbb{R}$ continue.

La convergence $\xi_n \xrightarrow{d} \xi$ entraîne immédiatement la convergence vers 0 du second terme du membre de droite de l'égalité.

Concernant le premier terme, pour $\varepsilon > 0$, on introduit l'événement $\{|\eta_n - c| \geq \varepsilon\}$. On a alors

$$\begin{aligned}
&\left| \mathbb{E}\left[e^{iu\xi_n}\left(e^{iv\eta_n} - e^{ivc}\right)\right]\right| \\
&= \left| \mathbb{E}\left[e^{iu\xi_n}\left(e^{iv\eta_n} - e^{ivc}\right)1_{|\eta_n - c| \geq \varepsilon}\right] + \mathbb{E}\left[e^{iu\xi_n}\left(e^{iv\eta_n} - e^{ivc}\right)1_{|\eta_n - c| < \varepsilon}\right]\right| \\
&\leq 2\,\mathbb{P}\left[|\eta_n - c| \geq \varepsilon\right] + |v|\varepsilon,
\end{aligned}$$

où l'on a utilisé $|e^{iv\eta_n} - e^{ivc}| \leq |v||\eta_n - c|$. On conclut en utilisant $\eta_n \xrightarrow{\mathbb{P}} c$ puis en faisant tendre ε vers 0. □

1.4.2 Lois des grands nombres et théorème central-limite

L'outil probabiliste essentiel de ce cours est le contrôle de la somme de variables aléatoire indépendantes (et souvent équidistribuées).

Notations

Si $X_1, \ldots, X_n$ est une suite de variables aléatoires réelles, on notera toujours

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

leur moyenne empirique. Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathbb{Q}$, on écrira

$$X_1 \ldots X_n \sim_{\text{i.i.d.}} \mathbb{Q}.$$

Dans ce contexte – et lorsqu'il n'y aura pas d'ambiguité – on introduira parfois la notation X pour désigner une variable de même loi que les X_i.

Lois des grands nombres

Proposition 1.9. *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi, telles que* $\mathrm{Var}\left[X\right] = \sigma^2 < +\infty$. *On note $\mu = \mathbb{E}\left[X\right]$. Alors*

$$\mathbb{E}\left[\overline{X}_n\right] = \mu \quad \textit{et} \quad \mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}.$$

Démonstration. On utilise simplement la linéarité de l'espérance et la propriété

$$\mathrm{Var}\big[\sum_{i=1}^{n} X_i\big] = \sum_{i=1}^{n} \mathrm{Var}\big[X_i\big]$$

qui est vérifiée si les X_i sont indépendantes. □

Remarque 1.12. *La Proposition 1.9 implique la convergence* $\overline{X}_n \xrightarrow{\mathcal{L}^2} \mu$ *et donc aussi* $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$.

Théorème 1.2 (Loi forte des grands nombres)**.** *Soient* $X_1, \ldots, X_n$ *des variables aléatoires indépendantes de même loi, telles que* $\mathbb{E}\big[|X|\big] < +\infty$. *On note* $\mu = \mathbb{E}\big[X\big]$. *Alors*

$$\overline{X}_n \xrightarrow{\text{p.s.}} \mu \quad \textit{lorsque} \quad n \to \infty.$$

Théorème central limite

Le théorème central limite donne la vitesse de convergence dans la loi des grands nombres. La Proposition 1.9 suggère que la bonne normalisation est $\sqrt{n}$: en effet, on a

$$\mathbb{E}\Big[\Big(\sqrt{n}\big(\overline{X}_n - \mu\big)\Big)^2\Big] = n\,\mathbb{E}\big[\big(\overline{X}_n - \mathbb{E}[\overline{X}_n]\big)^2\big] = n\mathrm{Var}\big[\,\overline{X}_n\,\big] = \sigma^2,$$

qui reste bornée lorsque $n \to \infty$. On cherche donc le comportement de l'erreur normalisée

$$\sqrt{n}\big(\overline{X}_n - \mu\big), \quad \text{lorsque} \quad n \to \infty.$$

Malheureusement, si la convergence existe, elle ne peut pas avoir lieu en probabilité[10] et il faut affaiblir le mode de convergence.

Théorème 1.3 (Théorème central limite)**.** *Soient* $X_1, \ldots, X_n$ *des variables aléatoires indépendantes de même loi, telles que* $\mathbb{E}\big[X^2\big] < +\infty$ *et* $\sigma^2 = \mathrm{Var}\big[X\big] > 0$. *On note* $\mu = \mathbb{E}\big[X\big]$. *Alors*

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1).$$

On dira que la suite ξ_n est asymptotiquement normale s'il existe deux constantes $\mu \in \mathbb{R}$ et $\sigma > 0$ telles que

$$\sqrt{n}(\xi_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

En particulier, le théorème central limite implique que la moyenne empirique est asymptotiquement normale. Le résultat suivant montre que si ξ_n est asymptotiquement normale, alors $g(\xi_n)$ l'est aussi à condition que $g : \mathbb{R} \to \mathbb{R}$ soit suffisamment régulière.

Cet outil technique essentiel porte en statistique le nom de « méthode delta ».

[10] voir l'Exercice 1.2.

Proposition 1.10 (méthode delta)**.** *Si ξ_n est asymptotiquement normale et $g : \mathbb{R} \to \mathbb{R}$ est continûment différentiable, alors $g(\xi_n)$ l'est aussi et*

$$\sqrt{n}\big(g(\xi_n) - g(\mu)\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2 g'(\mu)^2\big).$$

Démonstration. La fonction

$$h(x) = \begin{cases} \frac{g(x)-g(\mu)}{x-\mu} & \text{si} \quad x \neq \mu \\ g'(\mu) & \text{si} \quad x = \mu \end{cases}$$

est continue. La normalité asymptotique de ξ_n entraîne en particulier la convergence $\xi_n \xrightarrow{\mathbb{P}} \mu$, et donc aussi

$$h(\xi_n) \xrightarrow{\mathbb{P}} h(\mu) = g'(\mu).$$

Or $\sqrt{n}\big(g(\xi_n) - g(\mu)\big) = h(\xi_n)\eta_n$, avec $\eta_n = \sqrt{n}(\xi_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. La Proposition 1.8 (Slutsky) permet de conclure

$$h(\xi_n)\eta_n \xrightarrow{d} g'(\mu)\,\mathcal{N}(0, \sigma^2) \stackrel{d}{=} \mathcal{N}\big(\sigma^2 g'(\mu)^2\big),$$

le symbole $\stackrel{d}{=}$ signifiant « égalité en loi ». □

Version multidimensionnelle du théorème central limite

Théorème 1.4. *Soient $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ une suite de vecteurs aléatoires de $\mathbb{R}^d$ indépendants et de même loi, tels que $\mathbb{E}\left[\,\|\,\boldsymbol{X}\,\|^2\right] < +\infty$. On note $\boldsymbol{\mu} = \mathbb{E}\left[\,\boldsymbol{X}\,\right]$ et Σ la matrice de variance-covariance $d \times d$ de $\boldsymbol{X}$. On a*

$$\sqrt{n}\big(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\big) \xrightarrow{d} \mathcal{N}\big(0, \Sigma\big).$$

La « méthode delta » a elle aussi une version multidimensionnelle. Si $g : \mathbb{R}^d \to \mathbb{R}^k$ est continûment différentiable, elle s'écrit

$$g(x) = \big(g_1(x), \ldots, g_k(x)\big), \quad g_i : \mathbb{R}^d \to \mathbb{R},$$

et on note $J_g(x)$ la matrice de la différentielle de g au point $x \in \mathbb{R}^d$:

$$J_g(x) = \begin{pmatrix} \partial_1 g_1(x) & \ldots & \partial_d g_1(x) \\ \vdots & & \vdots \\ \partial_1 g_k(x) & \ldots & \partial_d g_k(x) \end{pmatrix}.$$

Proposition 1.11. *Soient $\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_n}$ une suite de vecteurs aléatoires de $\mathbb{R}^d$ asymptotiquement normale, au sens où :*

$$\sqrt{n}\big(\boldsymbol{\xi}_n - \boldsymbol{\mu}\big) \xrightarrow{d} \mathcal{N}\big(0, \Sigma\big)$$

où $\boldsymbol{\mu} \in \mathbb{R}^d$ et Σ est une matrice $d \times d$ symétrique positive. Alors, si $g : \mathbb{R}^d \to \mathbb{R}^k$ est continûment différentiable, on a

$$\sqrt{n}\big(g(\boldsymbol{\xi}_n) - g(\boldsymbol{\mu})\big) \xrightarrow{d} \mathcal{N}\big(0, J_g(\mu)\Sigma\, J_g(\mu)^T\big).$$

1.5 Exercices

Exercice 1.1. Soient X_n et Y_n deux suites de variables aléatoires réelles telles que $X_n \xrightarrow{\mathbb{P}} 0$ et $\sup_n \mathbb{E}\left[|Y_n|\right] < \infty$. Montrer que $X_n Y_n \xrightarrow{\mathbb{P}} 0$.

Exercice 1.2. Soit X_n une suite de variables aléatoires indépendantes centrées réduites. Par le théorème central limite, on a

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(0,1).$$

Le but de cet exercice est de montrer que S_n ne peut pas converger en probabilité.

- Décomposer la variable S_{2n} en fonction de S_n et d'une variable aléatoire indépendante de la précédente.
- Calculer la fonction caractéristique de $S_{2n} - S_n$ et montrer que cette différence converge en loi.
- En raisonnant par l'absurde, en déduire que S_n ne converge pas en probabilité.

Exercice 1.3. On pose

$$f(x) = \frac{|x|}{1+|x|}.$$

- Montrer que la suite de variables aléatoires X_n converge en probabilité vers X si et seulement si
$$\lim_{n\to\infty} \mathbb{E}\big[f(X_n - X)\big] = 0.$$
- Montrer que l'on peut remplacer f par $g(x) = \min\{|x|, 1\}$, et plus généralement par toute fonction f positive, continue, bornée, croissante sur $\mathbb{R}\backslash\{0\}$ vérifiant $f(0) = 0$ et $f(x) > 0$ si $x > 0$.
- En déduire que si X_n converge vers X en probabilité, il existe une sous-suite qui converge presque-sûrement. (Il existe une autre preuve facile de ce résultat à l'aide du lemme de Borel-Cantelli).

Chapitre 2

Expérience statistique

Une expérience statistique est la description mathématique d'une variable ou d'un vecteur aléatoire (l'observation) et d'un ensemble de lois de probabilité (le modèle) susceptibles d'avoir engendré cette observation.

A une expérience statistique est toujours associée une problématique : reconstruction d'un paramètre du modèle (estimation), décision sur les propriétés du modèle (tests).

2.1 Modélisation statistique*

2.1.1 Exemples introductifs

Exemple 1 : Sondage

Une élection entre deux candidats A et B a lieu : on effectue un sondage à la sortie des urnes. On interroge n votants, n étant considéré comme petit devant le nombre total de votants, et on récolte les nombres n_A et n_B de voix pour A et B respectivement ($n_A + n_B = n$, en ne tenant pas compte des votes blancs ou nuls pour simplifier).

> *Problématique statistique : peut-on affirmer que A ou B a gagné au vu de n_A et n_B seulement ? Si l'on décide d'annoncer A (ou B) vainqueur, comment quantifier l'erreur de décision ?*

La réponse va de toute évidence dépendre de n et du rapport n_A/n_B. Ce problème semble intimement lié avec l'expérience suivante : on lance une pièce de monnaie n fois et on compte les nombres n_P et n_F de piles et faces obtenus.

> *Problématique statistique : la pièce est-elle truquée ? Si $n = 100$ et $n_P = 19$, $n_F = 81$, on ne va pas vraiment hésiter. Mais qu'en est-il si $n = 20$, $n_P = 12$ et $n_F = 8$?*

Intuitivement, dans ces deux expériences statistiques, le problème de décision sera d'autant plus difficile à résoudre que la pièce est « peu truquée », ou bien que les deux candidats sont proches dans le coeur des électeurs d'un part, et si l'on a récolté peu de lancers ou de réponses (n petit) d'autre part.

Exemple 2 : Reconstruction d'un signal bruité

On transmet une signal périodique $\big(f(t), t \in [0,T]\big)$ échantillonné à une certaine fréquence N. Chaque donnée $f(k/N)$, $k = 1, \dots, NT$, est corrompue lors de la transmission par une erreur e_k, de sorte que l'on capte

$$Y_k = f(k/N) + e_k, \quad k = 1, \dots, NT.$$

On a $n = NT$ observations. On postule que les erreurs sont indépendantes les unes des autres, nulles en moyenne, et leur « ordre de grandeur » sans préciser plus pour le moment est $\sigma > 0$.

> *Problématique statistique : comment reconstruire f, c'est-à-dire comment construire une fonction $t \rightsquigarrow \widehat{f}\big(t;\ (Y_k)\big)$ ne dépendant que des observations Y_k – on dira un estimateur de f – de sorte que $\widehat{f}$ soit « proche » de f ?*

Intuitivement, la difficulté du problème va dépendre de N et du rapport entre la taille de f et le niveau de bruit σ, et bien sûr de la complexité du signal[1]. Voici une autre question très proche

> *Problématique statistique : comment décider si le canal transmet effectivement un signal (afin de déclencher une alarme, par exemple). Autrement dit, peut-on décider en vue des Y_k si $f = 0$ ou $f \neq 0$? Avec quelle probabilité de se tromper ?*

On peut imaginer un signal en dimension 2 : par exemple, une image définie sur le carré unité $[0,1] \times [0,1]$ pour une certaine discrétisation en pixels auxquels sont associés des niveaux de gris dans $[1, M] \cap \mathbb{N}$. Das ce cas, on observe

$$Y_{k,\ell} = f(k/N, \ell/N) + \xi_{k,\ell}, \quad 1 \leq k, \ell \leq N,$$

où

$$f : [0,1] \times [0,1] \to [1, M] \cap \mathbb{N}$$

et les $\xi_{l,\ell}$ sont des erreurs, nulles en moyenne et d'ordre de grandeur σ. On a $n = N^2$ observations. On pourra s'intéresser au problème de reconstruction de l'image f ou bien décider si une certaine caractéristique est présente dans l'image ou non.

[1] Un signal constant ou ayant une forme prescrite sera plus facile à reconstruire qu'un signal irrégulier.

Exemple 3 : Evaluation du risque d'un actif financier

On recueille sur le marché les données du prix $(S_t, t \geq 0)$ d'un actif financier sur l'intervalle de temps $[0, T]$, pour une certaine échelle d'échantillonnage Δ : par exemple, une semaine ou un jour, une heure, quelques minutes, etc. On observe les rendements logarithmiques

$$Y_i^{\Delta} = \log \frac{S_{i\Delta}}{S_{(i-1)\Delta}}, \quad i \ = 1, \ldots, n = \lfloor T/\Delta \rfloor.$$

On a $n = \lfloor T/\Delta \rfloor$ observations. Si l'on se place dans la théorie classique de Black-Scholes, la dynamique du prix suit l'équation

$$\frac{dS_t}{S_t} = \mu dt + \sigma dB_t, \tag{2.1}$$

où $(B_t, t \geq 0)$ est un mouvement brownien, $\mu \in \mathbb{R}$ est le drift et $\sigma > 0$ la volatilité de l'actif.

> *Problématique statistique : comment reconstruire*[2] *la volatilité σ à partir des données historiques Y_i^{Δ} ? On peut aussi vouloir estimer le risque $\mu/(\sigma\sqrt{T})$ de l'actif*[3]*.*

La réponse va dépendre de T, σ et μ, mais aussi de Δ, choisi par le statisticien.

Exemple 4 : Biopuces et analyse d'ADN

On dispose d'un procédé de biologie moléculaire, les biopuces (ou microarrays) qui permet – dans un certain sens – de mesurer l'activité de l'expression de gènes d'un individu d'une espèce biologique dans certaines situations[4]. Dans ce cas, on dispose pour chaque individu i d'une suite de localisations (qui correspondent grossièrement à des gènes) et d'une expression correspondante qui prend la forme

$$\boldsymbol{X}_i = (X_1^{(i)}, \ldots, X_J^{(i)}), \ \ i = 1, \ldots, N$$

où $X_j^{(i)} \geq 0$ est le niveau d'expression des gènes parmi les sites $\{1, \ldots, J\}$ pour l'individu i pris dans une population de taille N. On a[5] $n = JN$ observations.

> *Problématique statistique : peut-on localiser les sites i responsables d'un état donné, sachant que les mesures des $X_j^{(i)}$ sont sujettes à des erreurs ? Si l'on se donne deux populations, l'une atteinte d'une maladie soupçonnée d'être*

[2] par exemple, pour la comparer avec la volatilité implicite donnée par des prix d'options.

[3] que l'on désigne aussi comme son ratio de Sharpe.

[4] Par exemple, en laboratoire, on peut mesurer l'intensité de l'expression de certains gènes d'un insecte infecté dans le but de localiser les gènes promoteurs de la réponse immunitaire.

[5] avec le fait notable qu'en pratique $N \ll J$: N est de l'ordre de quelques individus alors que J est de l'ordre de plusieurs milliers.

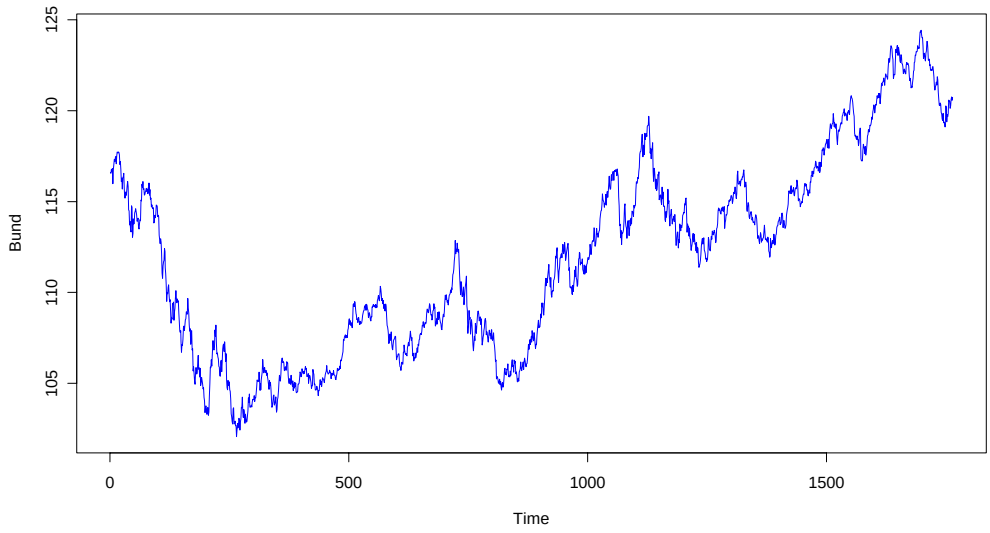

FIG. 2.1 – Exemple 3 : observation des prix du contrat futur FGBL (Obligation 10 ans de l'Etat allemand), entre avril 1999 et décembre 2005. L'échantillonnage est de $\Delta = 1$ jour.

d'origine génétique, l'autre population étant saine, peut-on décider au vu des données $\boldsymbol{X}_i$ pour les deux populations si la maladie en question est d'origine génétique ?

Exemple 5 : Contrôle de qualité, données censurées

On cherche – en laboratoire – à tester la fiabilité d'un appareil industriel. On fait fonctionner en parallèle n appareils jusqu'à ce qu'ils tombent tous en panne. On note

$$X_1, \dots, X_n$$

les instants de panne observés. On dispose donc de n observations.

Problématique statistique : comment reconstruire la loi du temps de panne ? Le temps de panne moyen est-il raisonnable (plus petit qu'un seuil donné) ?

La précision d'estimation sur la loi du temps de panne des X_i sera d'autant meilleure que n est grand.

Si les appareils sont fiables, ce qui est réaliste en pratique, la quantité $\max_{i=1,\dots,n} X_i$ sera souvent hors d'atteinte pour le statisticien. On stoppe l'expérience après un temps terminal T et on observe plutôt

$$X_i^\star = \min\{X_i, T\}, \quad i = 1, \dots n.$$

Problématique statistique : quelle est la perte d'information, quantifiée par T, dans cette seconde expérience plus réaliste ?

Exemple 6 : Influence d'une variable sur une autre

Comment quantifier une assertion comme « la taille d'un individu est fonction de son âge » ? Si on note Y la taille et X l'âge typiques d'un individu, il est irréaliste de postuler l'existence d'une fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $Y = f(X)$.

Toutefois, on peut espérer que la « variabilité » de Y est « essentiellement contenue » dans celle de X dans le sens suivant : si X et Y sont deux variables aléatoires avec Y de carré intégrable écrivons

$$Y = r(X) + \xi, \quad \text{avec} \quad r(X) = \mathbb{E}\left[Y \,|\, X\right],$$

de sorte que $\xi = Y - \mathbb{E}\left[Y \,|\, X\right]$ est un « bruit » centré. Cette décomposition est motivée par la propriété de l'espérance conditionelle qui est la meilleure approximation de Y par une variable X- mesurable, au sens suivant :

$$\mathbb{E}\left[\big(Y - r(X)\big)^2\right] = \min_h \mathbb{E}\left[\big(Y - h(X)\big)^2\right]$$

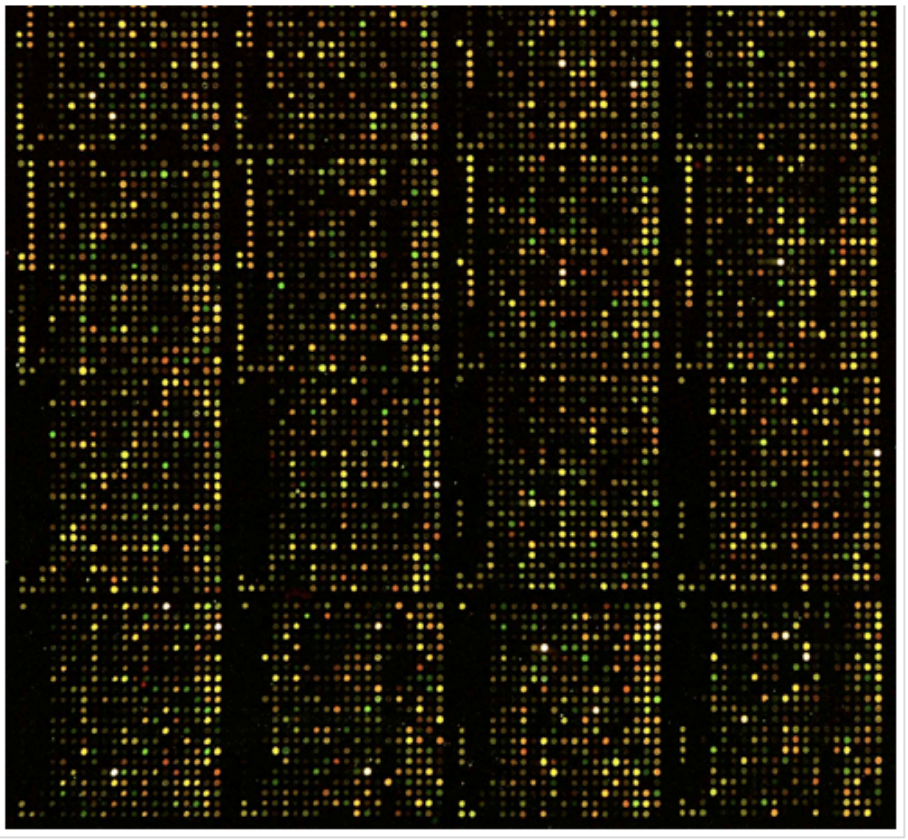

FIG. 2.2 – Exemple 4 : observation d'une biopuce en laboratoire : chaque carré lumineux mesure l'intensité d'expression d'un gène (en fait d'une séquence d'ARNm codante suffisamment longue pour être mise en correspondance avec un gène via la production de peptides pour lesquels code la séquence d'ADN correspondante). La représentation « en carrés » est donnée pour économiser la représentation : il n'y a pas *a priori* de structure bi-dimensionnelle associée à cette « image ».

où le minimum est pris sur l'ensemble des fonctions boréliennes. C'est une caractérisation de l'espérance conditionnelle pour des variables de carré intégrable (voir, par exemple, Jacod et Protter [3]).

On traduit « la taille d'un individu est fonction de son âge » par « la variance du bruit $\sigma^2 = \mathbb{E}\left[\xi^2\right]$ est petite » par exemple. On collecte les âges et tailles (X_i, Y_i) d'une population de n individus. Les observations sont les (X_i, Y_i), avec

$$Y_i = r(X_i) + \xi_i, \quad i = 1, \ldots, n \tag{2.2}$$

et les ξ_i sont des bruits centrés de taille σ^2. On a n observations (ou $2n$ selon le point de vue). Les X_i portent le nom de covariables, ou variables explicatives.

> *Problématique statistique : comment reconstruire la fonction r – appelée fonction de régression – et estimer l'intensité σ^2 du bruit ?*

Ce contexte est proche de celui de l'exemple 1 du signal bruité, à ceci près que les points k/N sont remplacés par les données aléatoires X_i, dont les valeurs ne sont pas choisies par le statisticien. Mais si les X_i sont « bien répartis » on s'attend à ce que les deux modèles soient proches lorsque n est grand.

Les variables X et Y n'ont pas vocation à être de même dimension : on peut remplacer X par un vecteur $\boldsymbol{X} \in \mathbb{R}^k$ qui collecte un ensemble de covariables possibles. Dans ce cas, la représentation (2.2) devient $Y_i = r(\boldsymbol{X}_i) + \xi_i$ où maintenant $r : \mathbb{R}^k \to \mathbb{R}$, que l'on peut chercher à reconstruire.

Il existe aussi des situations où Y est une variable qualitative, c'est-à-dire ne prenant qu'un nombre fini de valeurs. On peut penser que le risque de maladie coronarienne chez un individu est influencé par toute une série de facteurs : pression systolique, consommation de tabac, d'alcool, taux de cholestérol, poids, âge, terrain familial, etc. On note $Y_i \in \{0, 1\}$ l'absence ou la présence de maladie coronarienne pour un individu i d'étude donné, et $\boldsymbol{X}_i$ le vecteur des covariables constitué des différentes données recueillies chez l'individu i. Dans ce cas, on a

$$r(\boldsymbol{x}) = \mathbb{P}\left[Y = 1 | \ \boldsymbol{X} = \boldsymbol{x}\right],$$

qui s'interprète comme la probabilité d'être atteint de maladie coronarienne, sachant le vecteur des covariables $\boldsymbol{X}$.

2.1.2 Définition provisoire d'une expérience statistique⋆

Construire un modèle statistique consiste à identifier trois éléments distincts :

1. *Des observations*
$$\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_n \tag{2.3}$$
où les $\mathtt{x}_i$ sont des réels, mais on peut imaginer des situation plus complexes.[6] Ces observations sont associées à la réalisation d'une expérience physique, et le point de départ du statisticien est donc le résultat de cette expérience.

2. *Un modèle stochastique* associé à l'expérience qui a engendré les observations. Les observations sont considérées comme la réalisation de variables aléatoires. La loi de ces variables aléatoire identifie le mécanisme de formation des observations. Cette loi dépend de paramètres inconnus.

3. *Une problématique* associée au couple (observations, modèle). Il s'agit pour le statisticien de « retrouver » – on dira estimer – les paramètres inconnus. Il faut pouvoir contrôler la qualité de cette estimation.
On peut aussi vouloir prendre une décision, par exemple sous la forme d'un test d'hypothèse sur les paramètres. Il faut pouvoir contrôler l'erreur de décision.[7]

La problématique statistique consiste à développer le point 3 dans des situations associées aux points 1–2.

Définition 2.1 (provisoire d'une expérience statistique)**.** *Une expérience statistique est la donnée d'observations et d'un modèle stochastique susceptible d'avoir engendré ces observations.*

Mathématiquement, les observations sont la réalisation d'un vecteur aléatoire Z dont la loi $\mathbb{P}^Z$ est prise dans une famille $\mathcal{P}$ de probabilités possibles donnée à l'avance et qui définit le modèle stochastique associé à l'observation.

Cette définition règle[8] provisoirement les points 1 et 2. Au moyen d'une paramétrisation appropriée, on peut toujours représenter la famille $\mathcal{P}$ sous la forme
$$\mathcal{P} = \big\{ \mathbb{P}_\vartheta, \ \ \vartheta \in \Theta \big\},$$

[6] On peut considérer des données qualitatives, que l'on pourra coder par des entiers, ou bien des données plus complexes, comme par exemple une surface où la trajectoire d'un processus La difficulté provient de l'organisation des $\mathtt{x}_i$ qui peut être complexe (vecteurs, tableaux) et ne transparaît pas dans l'écriture (2.3).

[7] C'est-à-dire la probabilité d'accepter une hypothèse sur les paramètres alors qu'elle est fausse, ou de la rejeter alors qu'elle est vraie.

[8] Avec les notations de la définition 2.1, les observations s'écrivent sous la forme
$$(\mathtt{x}_1, \ldots, \mathtt{x}_n)^T = Z(\omega)$$

où Θ est un ensemble de paramètres possibles. Le point 3 se traduit ainsi :

Définition 2.2. *La problématique statistique (ou l'inférence statistique) consiste, à partir d'une réalisation d'un vecteur aléatoire Z, dont la loi $\mathbb{P}^Z$ est prise dans une famille $\{\mathbb{P}_\vartheta,\ \mathbb{P}_\vartheta \in \Theta\}$ donnée, à retrouver le paramètre ϑ tel que $\mathbb{P}^Z = \mathbb{P}_\vartheta$.*

Le paramètre ϑ résume toute l'information que peut apporter l'observation $Z(\omega)$. Identifier ϑ est équivalent à identifier $\mathbb{P}_\vartheta$, c'est-à-dire la loi de la variable aléatoire Z dont on a observé une réalisation $Z(\omega)$.

2.2 Formulation mathématique

2.2.1 Expérience engendrée par une observation

Situation

Une expérience statistique est la donnée d'un vecteur aléatoire Z à valeurs dans un espace mesurable $(\mathfrak{Z}, \mathcal{Z})$, le plus souvent $(\mathbb{R}^n, \mathcal{B}^n)$ et définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. La problématique statistique consiste à supposer que $\mathbb{P}^Z$ appartient à une famille de probabilités sur $(\mathfrak{Z}, \mathcal{Z})$, et le but est de « retrouver » les propriétés de $\mathbb{P}^Z$ à partir de l'observation d'une réalisation de Z seulement.

On représente cette famille sous la forme $\left\{\mathbb{P}_\vartheta, \vartheta \in \Theta\right\}$, où ϑ est un paramètre et Θ un ensemble de paramètres. Dans un problème statistique, seul « l'espace d'état » $(\mathfrak{Z}, \mathcal{Z})$ et la famille de probabilités $\left\{\mathbb{P}_\vartheta, \vartheta \in \Theta\right\}$ comptent. Une fois ces éléments spécifiés, la donnée de Z et de l'espace $(\Omega, \mathcal{F}, \mathbb{P})$ deviennent superflus.

Définition 2.3 (Expérience statistique)**.** *Une expérience (un modèle) statistique $\mathcal{E}$ est la donnée d'un triplet*

$$\mathcal{E} = \left(\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\vartheta, \vartheta \in \Theta\}\right)$$

où $(\mathfrak{Z}, \mathcal{Z})$ est un espace mesurable et $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ une famille de probabilités définie sur $(\mathfrak{Z}, \mathcal{Z})$. On appelle Θ l'ensemble des paramètres.

On parle indifféremment d'expérience statistique ou de modèle statistique. On parlera parfois simplement du modèle $\left\{\mathbb{P}_\vartheta, \vartheta \in \Theta\right\}$ lorsque le contexte ne prête pas à confusion[9].

Définition 2.4 (Expérience engendrée par une observation)**.** *Si l'expérience statistique $\mathcal{E}$ est construite à partir d'une observation Z par le procédé ci-dessus, on dit que $\mathcal{E}$ est engendrée par l'observation Z.*

---

et sont donc appréhendées comme la réalisation d'un vecteur aléatoire Z défini implicitement sur un espace mesurable $(\Omega, \mathcal{A})$. La famille $\mathcal{P}$ est un ensemble de mesures de probabilités définies sur l'image $Z(\Omega)$ de Z.

[9] sans préciser l'espace $(\mathfrak{Z}, \mathcal{Z})$ sur lequel sont définies simultanément toutes les probabilités $\mathbb{P}_\vartheta$, $\vartheta \in \Theta$.

Exemple

On observe n variables aléatoires indépendantes, gaussiennes de moyenne $\mu \in \mathbb{R}$ et de variance $\sigma^2 > 0$. L'expérience statistique associée est décrite comme l'observation de

$$X_1, \dots, X_n \text{ indépendantes, identiquement distribuées,}$$

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \ \ \mu \in \mathbb{R}, \ \ \sigma^2 > 0.$$

Il existe donc un espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ sur lequel est défini le vecteur aléatoire $Z = (X_1, \dots, X_n)^T$ et $\mathbb{P}^Z$ est la loi de n variables gaussiennes indépendantes de moyenne μ et de variance σ^2. La probabilité $\mathbb{P}^Z$, définie sur $\big(\mathbb{R}^n, \mathcal{B}^n\big)$, dépend de μ et σ^2 même si cela ne transparaît pas dans les notations. On a

$$\mathbb{P}^Z[A] = (2\pi)^{-n/2} \int_A \exp\big(-\tfrac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\big) dx_1 \cdots dx_n, \ A \in \mathcal{B}^n.$$

Dans ce cas, on construit l'expérience $\mathcal{E}$ associée de la façon suivante : on pose

$$\big(\mathfrak{Z}, \mathcal{Z}\big) = \big(\mathbb{R}^n, \mathcal{B}^n\big), \ \ \vartheta = (\mu, \sigma^2), \ \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}, \ \ \mathbb{P}_\vartheta = \mathbb{P}^Z.$$

Remarque 2.1. Si on est rigoureux, on ne peut pas dire que l'on observe Z, mais plutôt que l'on observe une réalisation $Z(\omega)$ de Z, qui correspond aux « données physiques » $\mathtt{x}_1, \mathtt{x}_2, \dots, \mathtt{x}_n$ que l'on traite effectivement en pratique. Mathématiquement, cela n'a aucune importance, et on s'autorisera cet abus de langage. Le paragraphe suivant permet de lever cette ambiguité[10] sur laquelle nous ne reviendrons plus.

2.2.2 Observation canonique*

Lorsque l'on spécifie directement une expérience statistique $\mathcal{E}$ via la Définition 2.3, il n'y a pas d'observation Z. Une façon immédiate de « considérer » $\mathcal{E}$ comme engendrée par une observation Z consiste à poser

$$(\Omega, \mathcal{F}) = \big(\mathfrak{Z}, \mathcal{Z}\big) \ \text{ et } \ Z(\omega) = \omega, \ \ \omega \in \Omega,$$

et $\mathbb{P}^Z = \mathbb{P}_\vartheta$ est la loi de Z qui dépend ici explicitement de ϑ dans les notations.

Définition 2.5 (Observation canonique)**.** *Si l'observation Z est construite à partir d'une expérience statistique $\mathcal{E}$ par le procédé ci-dessus, on dit que Z est l'observation canonique associée à $\mathcal{E}$.*

Ces deux points de vue peuvent parfois être source de confusion, principalement dans les notations. Dans la pratique (mathématique) on n'aura pas besoin de se soucier du point de vue sous lequel on se place, les Définitions 2.4 et 2.5 étant équivalentes.

[10] En statistique, on parle de Z pour désigner $Z(\omega)$, à l'inverse de la pratique qui consiste à écrire parfois $f(x)$ pour désigner la fonction f.

2.2.3 Domination

Appréhender une famille de mesure $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ sans plus d'hypothèse est très ambitieux, comme on le verra au Chapitre 3. Sous une hypothèse de régularité, dite de domination, on ramène le problème de l'étude des $\mathbb{P}_\vartheta$ à une famille de fonctions sur $(\mathfrak{Z}, \mathcal{Z})$.

Définition 2.6. *Etant données deux mesures positives σ-finies μ et ν définies sur $(\mathfrak{Z}, \mathcal{Z})$, on dit que μ domine ν et on écrit $\nu \ll \mu$ si*

$$\mu[A] = 0 \Rightarrow \nu[A] = 0.$$

Le théorème de Radon-Nikodym (voir par exemple Jacod et Protter [3], Chapitre 28) entraîne l'existence d'une fonction mesurable positive $x \rightsquigarrow p(x)$, notée $x \rightsquigarrow \frac{d\nu}{d\mu}(x)$, appelée densité de ν par rapport à μ, définie à un ensemble μ-négligeable près, de sorte que

$$\nu(dx) = p(x)\mu(dx),$$

au sens où

$$\mu[A] = \int_A p(x)\mu(dx) = \int_A \frac{d\nu}{d\mu}(x)\mu(dx),\ A \in \mathcal{Z}.$$

Définition 2.7. *Une expérience statistique $\mathcal{E} = (\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\vartheta, \vartheta \in \Theta\})$ est dominée par la mesure σ-finie μ définie sur $(\mathfrak{Z}, \mathcal{Z})$ si pour tout $\vartheta \in \Theta$, la mesure μ domine $\mathbb{P}_\vartheta$.*

Dans ce cas, il existe, pour tout $\vartheta \in \Theta$ une densité $x \rightsquigarrow p(\vartheta, x) = \frac{d\mathbb{P}_\vartheta}{d\mu}(x)$ de sorte que

$$\mathbb{P}_\vartheta(dx) = p(\vartheta, x)\mu(dx),\ \ x \in \mathfrak{Z}.$$

L'hypothèse de domination permet de « réduire » l'étude de la complexité de la famille de mesure $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ à celle de l'application

$$p : \Theta \times \mathfrak{Z} \to \mathbb{R}_+$$

et de la mesure dominante μ. Nous verrons dans les chapitres suivants comment l'étude systématique des propriétés de $p(\bullet, \bullet)$ rend compte des propriétés de $\mathcal{E}$.

2.2.4 Modèles paramétriques, non-paramétriques*

On distingue deux types d'expériences statistiques : les expériences paramétriques, où Θ peut s'écrire comme un sous-ensemble de $\mathbb{R}^d$, le paramètre ϑ pouvant être décrit par une nombre fini de composantes, et les expériences non-paramétriques, où ϑ est un élément d'un espace fonctionnel.

Par exemple, dans les exemples 2 –signal bruité et 6 –influence d'un variables sur une autre de la Section 2.1.1, le paramètre inconnu est le signal f ou la fonction de régression r. Si l'on postule que f (ou r) se représente sous la forme

$$f(\vartheta, x) = \sum_{i=1}^{d} \vartheta_i \varphi_i(x), \quad x \in \mathbb{R}$$

où les fonctions φ_i sont données, l'expérience statistique est paramétrique, et

$$\vartheta = (\vartheta_1, \ldots, \vartheta_d)^T \in \Theta \subset \mathbb{R}^d .$$

Le choix $d = 2$ et $r(\vartheta, x) = \vartheta_0 + \vartheta_1 x$ correspond à « la droite de régression », que l'on étudiera en détails dans la Section 5.2.

Si f est un élément quelconque d'un espace fonctionnel (décrit le plus souvent par des propriétés de régularité fonctionnelles : par exemple, f est de carré intégrable et dérivable un certain nombre de fois dans L^2), alors l'expérience associée est non-paramétrique et le paramètre ϑ est la fonction f elle-même. Si les fonctions φ_i sont les d-premiers éléments d'une base orthogonale de L^2 , alors la transition d'une situation paramétrique vers une situation non-paramétrique consiste formellement à passer à la limite dans le nombre de dimensions d qui décrivent le paramètre inconnu.

La distinction paramétrique ou non-paramétrique est un choix de modélisation . Pour l'exemple 2 de la transmission d'un signal bruité ou de la reconstruction d'une image de la Section 2.1.1, un modèle non-paramétrique semble plus approprié que pour l'exemple du sondage. Pour l'exemple 3 de l'estimation de la volatilité, on a choisi de prendre $\sigma > 0$ constant. Si on veut tenir compte des fluctuations de la volatilité dans le temps, une représentation fonctionnelle $(\sigma(t), t \geq 0)$ est plus appropriée. Le modèle sera plus proche de la réalité, mais le problème statistique plus difficile .

Dans ce cours, hormis le Chapitre 3, nous nous restreindrons à l'étude d'expériences paramétriques.

2.3 Exemples

2.3.1 Modèle d'échantillonnage ou du n–échantillon

De par la simplicité de sa structure, c'est une des expérience statistiques les plus étudiées, et qui occupe trois chapitres de ce cours.

Pour $n \geq 1$, on considère (la suite) d'expérience(s) engendrée par l'observation de n-variables aléatoires réelles

$$X_1, \ldots, X_n \quad \text{indépendantes, identiquement distribuées,}$$

de loi inconnue F sur $\mathbb{R}$, où $F \in \mathfrak{F}$ appartient à une famille de loi $\mathfrak{F}$ donnée. L'expérience statistique $\mathcal{E}^n$ correspondante est engendrée par le vecteur $Z = (X_1, \ldots, X_n)^T$ et on peut écrire

$$\mathcal{E}^n = \big(\mathbb{R}^n, \mathcal{B}^n, \{\mathbb{P}_F^n, F \in \mathfrak{F}\}\big)$$

où $\mathbb{P}_F^n$ est la loi sur $\mathbb{R}^n$ de n-variables aléatoires indépendantes de loi F. Cela signifie en particulier, que, pour tous $x_1, \ldots, x_n \in \mathbb{R}$, on a

$$\mathbb{P}_F^n \left[X_1 \leq x_1, \ldots, X_n \leq x_n\right] = \prod_{i=1}^n F(x_i).$$

En particulier, si $\mathfrak{F}$ est constituées de distributions F absolument continues, de densité f, alors le vecteur $(X_1, \ldots, X_n)$ admet une densité par rapport à la mesure de Lebesgue donnée par

$$(x_1, \ldots, x_n) \rightsquigarrow p(x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i).$$

Dans ce cas, on a

$$\mathbb{P}_F(dx_1 \ldots dx_n) = p(x_1, \ldots, x_n) dx_1 \ldots dx_n \tag{2.4}$$

et l'expérience $\mathcal{E}^n$ est dominée par la mesure de Lebesgue sur $\mathbb{R}^n$.

Si $\mathcal{E}$ désigne l'expérience engendrée par une seule observation $X \sim F$, c'est-à-dire

$$\mathcal{E} = \big(\mathbb{R}, \mathcal{B}, \{F \in \mathfrak{F}\}\big)$$

alors $\mathcal{E}^n$ est le « produit » de n copies indépendantes de $\mathcal{E}$ et on écrit parfois

$$\mathcal{E}^n = \mathcal{E} \times \ldots \times \mathcal{E} \quad (n\text{-fois}).$$

Si la famille $\mathfrak{F}$ est dominée par une mesure μ sur $\mathbb{R}$, alors l'expérience $\mathcal{E}^n$ est dominée par la mesure produit $\mu^n = \mu \otimes \ldots \otimes \mu$ sur $\mathbb{R}^n$. En particulier, si μ est la mesure de Lebesgue sur $\mathbb{R}$, on retrouve (2.4).

Les exemples 1 –sondage , 3 –risque d'un actif financier et 5 –contrôle de qualité de la Section 2.1.1 sont des modèles d'échantillonnage :

1. Pour l'exemple 1 –sondage ou lancer de dé, on peut associer à chaque votant une variable X_i prenant la valeur 0 ou 1 selon que l'on vote pour A (pile) ou B (face). La loi de X_i est une loi de Bernoulli de paramètre inconnu $\vartheta \in \Theta = [0, 1]$. Si $\vartheta < 1/2$, A gagne. Si $\vartheta \neq \frac{1}{2}$, la pièce est truquée.
   Si l'on récolte la suite complète $X_1, \ldots, X_n$ des votes (des lancers) supposés indépendants et de même loi de Bernoulli de paramètre ϑ, alors on est dans un modèle d'échantillonnage, et l'expérience associée s'écrit

$$\mathcal{E}^n = \Big(\{0,1\}^n, \text{tribu des parties de } \{0,1\}^n, \big\{\mathbb{P}_\vartheta^n, \vartheta \in \Theta\big\}\Big),$$

où

$$\mathbb{P}_\vartheta^n = \mathbb{P}_\vartheta \otimes \cdots \otimes \mathbb{P}_\vartheta \ \ (n \text{ fois}),$$

avec

$$\mathbb{P}_\vartheta \big[X = 1\big] = \vartheta = 1 - \mathbb{P}_\vartheta[X = 0],$$

ce que l'on peut encore écrire sous la forme

$$\mathbb{P}_\vartheta(dx) = \vartheta\delta_1(dx) + (1-\vartheta)\delta_0(dx),$$

où $\delta_a(dx)$ désigne la mesure de Dirac au point a. Cette dernière représentation permet de mettre en évidence la mesure de comptage $\mu(dx) = \delta_0(dx) + \delta_1(dx)$ sur $\{0,1\}$ comme mesure dominante pour $\mathbb{P}_\vartheta$. La mesure de comptage $\mu^n = \mu \otimes \cdots \otimes \mu$ sur le produit $\{0,1\}^n$ domine alors l'expérience $\mathcal{E}^n$.

Une autre manière de procéder est de considérer que l'on n'observe que le nombre de votants n_A pour le candidat A (ou n_P), ce qui donne aussi n_B (ou n_F), puisque $n_A + n_B = n_P + n_F = n$. Dans ce cas, on n'a qu'une seule observation X, et on modélise n_A comme la réalisation d'une variable aléatoire X binômiale de paramètres (n, ϑ), où $\vartheta \in \Theta = [0,1]$ est le paramètre inconnu. Dans ce cas, l'expérience statistique s'écrit

$$\widetilde{\mathcal{E}}^n = \Big(\{0,n\}, \text{tribu des parties de } \{0,n\}, \big\{\mathbb{Q}_\vartheta^n, \vartheta \in \Theta\big\}\Big),$$

où cette fois-ci les $\mathbb{Q}_\vartheta^n$ sont définies sur $\{0,\ldots,n\}$ et

$$\mathbb{Q}_\vartheta^n\big[X = x\big] = C_n^x \vartheta^x (1-\vartheta)^{n-x}, \ \ x = 0,\ldots,n,$$

ce qui s'écrit aussi

$$\mathbb{Q}_\vartheta^n(dx) = \sum_{k=0}^n C_n^k \vartheta^k (1-\vartheta)^{n-k} \delta_k(dx).$$

Cette dernière représentation permet de mettre en évidence la mesure de comptage $\mu_n(dx) = \sum_{k=0}^n \delta_k(dx)$ sur $\{0,\ldots,n\}$ comme mesure dominante du modèle.

Intuitivement les expériences statistiques $\mathcal{E}^n$ et $\widetilde{\mathcal{E}}^n$ contiennent la même information sur le paramètre ϑ. On verra au Chapitre 6 comment formaliser et quantifier cette idée.

2. Pour l'exemple 3, les observations s'écrivent

$$Y_i^\Delta = \mu\Delta + \sigma(B_{i\Delta} - B_{(i-1)\Delta}) \sim \mathcal{N}\big(\mu\Delta, \sigma^2\Delta\big)$$

et sont indépendantes, en utilisant les propriétés caractéristiques du mouvement brownien (que l'on pourra admettre) : $B_t - B_s \sim \mathcal{N}(0, t-s)$ et $B_t - B_s$ est indépendant du passé jusqu'à l'instant s.

La loi F de Y_i^Δ est dominée par la mesure de Lebesgue sur $\mathbb{R}$ et sa densité

$$x \rightsquigarrow f(\vartheta, x) = (2\pi\Delta\sigma^2)^{-1/2} \exp\big(-\tfrac{1}{2\sigma^2\Delta}(x - \Delta\mu)^2\big)$$

dépend du paramètre $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$.

3. Pour l'exemple 5, c'est évident. Noter qu'un modèle classique de durée de vie est fourni par la famille de lois exponentielles de paramètre $\vartheta \in \mathbb{R}_+ \setminus \{0\}$. Dans ce cas, l'expérience $\mathcal{E}$ est dominée par la mesure de Lebesgue sur $\mathbb{R}$ et la loi de Y_i s'écrit

$$\mathbb{P}_\vartheta(dx) = \vartheta e^{-\vartheta x} 1_{\{x \in \mathbb{R}_+\}} dx.$$

Si les variables Y_i sont censurées par un instant terminal T connu, on observe alors plutôt $Y_i^\star = \min\{Y_i, T\}$. Dans ce cas, la loi $\mathbb{P}^\star$ de $Y_i^\star$ n'est ni discrète, ni continue, comme dans la Section 1.1.3 du Chapitre 1.
On pourra montrer en exercice que $\mathbb{P}^\star$ est dominée par $\mu(dx) = dx + \delta_T(dx)$, où dx est la mesure de Lebesgue sur $\mathbb{R}$ et $\delta_T(dx)$ est la mesure de Dirac au point T. On a

$$\mathbb{P}^\star_\vartheta(dx) = p(\vartheta, x)\mu(dx),$$

où

$$p(\vartheta, x) = \vartheta e^{-\vartheta x} 1_{\{x<T\}} + c(\vartheta) 1_{\{x=T\}},$$

avec $c(\vartheta) = \int_T^{+\infty} \vartheta e^{-\vartheta t} dt = e^{-\vartheta T}$.

2.3.2 Modèles de régression

Régression conditionnelle ou modèle de signal bruité

On observe une fonction $r : \mathbb{R}^k \to \mathbb{R}$ échantillonnée en n points, chaque observation étant « bruitée » par une erreur systématique :

$$Y_i = r(\boldsymbol{x}_i) + \xi_i, \quad i = 1, \ldots, n.$$

Les bruits ξ_i sont des variables indépendantes, identiquement distribuées, centrées et de carré intégrable. Les $\boldsymbol{x}_i$ sont les points d'échantillonnage, appelés parfois points de « design », définis sur un domaine $\mathcal{D} \subset \mathbb{R}^k$ en général borné. Si $k = 1$, on prend le plus souvent $\mathcal{D} = [0, 1]$ et $\boldsymbol{x}_i = x_i = i/n$, $i = 1, \ldots, n$. Si $k \geq 1$ on peut imaginer que les points se « répartissent » de façon régulière sur $\mathcal{D}$, ou bien au contraire qu'ils se concentrent dans une région de $\mathcal{D}$. Dans cette acceptation du modèle de régression, le statisticien choisit les points $\boldsymbol{x}_i$.

Si $r = r(\vartheta, \bullet)$ est connue au paramètre $\vartheta \in \Theta \subset \mathbb{R}^d$ près, le modèle est paramétrique. C'est le cas qui nous intéressera. Une forme paramétrique particulièrement importante est la régression linéaire $r(\vartheta, \boldsymbol{x}) = \vartheta^T \boldsymbol{x}$, qui est bien définie dès que[11] $k \leq d \leq k + 1$.

L'expérience statistique correspondante $\mathcal{E}^n$ est engendrée par les Y_i, $i = 1, \ldots, n$. Ce sont des variables indépendantes mais pas identiquement distribuées (chaque Y_i dépend de $\boldsymbol{x}_i$). On a

$$\mathcal{E}^n = \big(\mathbb{R}^n, \mathcal{B}^n, \{\mathbb{P}^n_\vartheta, \vartheta \in \Theta\}\big),$$

[11] si $k = d - 1$ on peut toujours compléter $\boldsymbol{x}$ en ajoutant la composante 1 au vecteur $\boldsymbol{x}$, qui jouera le rôle d'ordonnée à l'origine. Par exemple, si $\vartheta = (\vartheta_0, \vartheta_1, \vartheta_2)^T$, on peut modifier $\boldsymbol{x} = (x_1, x_2)^T$ en $\tilde{\boldsymbol{x}} = (1, x_1, x_2)^T$ et on a alors $\vartheta^T \tilde{\boldsymbol{x}} = \vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2$.

où $\mathbb{P}_\vartheta^n$ est la loi conjointe des Y_i. En particulier, pour tous $y_1, \dots, y_n \in \mathbb{R}$,

$$\mathbb{P}_\vartheta^n \left[Y_1 \leq y_1, \dots, Y_n \leq y_n \right] = \prod_{i=1}^n F_{\boldsymbol{x}_i}(y_i),$$

où $y \rightsquigarrow F_{\boldsymbol{x}_i}(y)$ est la fonction de répartition de Y_i. Par exemple, si ξ_i a une densité g par rapport à la mesure de Lebesgue sur $\mathbb{R}$, on a

$$F_{\boldsymbol{x}_i}(y) = \int_\infty^y g\big(t - r(\vartheta, \boldsymbol{x}_i)\big) dt.$$

Dans ce cas, le vecteur $(Y_1, \dots, Y_n)$ a lui-même une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, donnée par

$$(y_1, \dots, y_n) \rightsquigarrow p(\vartheta, y_1, \dots, y_n) = \prod_{i=1}^n g\big(t - r(\vartheta, \boldsymbol{x}_i)\big).$$

On a alors

$$\mathbb{P}_\vartheta(dy_1 \dots dy_n) = p(\vartheta, y_1, \dots, y_n) dy_1 \dots dy_n$$

et le modèle est dominé par la mesure de Lebesgue sur $\mathbb{R}^n$.

L'exemple 2 –signal bruité de la Section 2.1.1 est un modèle de régression conditionnelle.

Le terme de régression conditionnelle pour ce modèle se justifie par opposition à la régression non-conditionnelle ou avec variables explicatives, que nous présentons maintenant.

Régression avec variables explicatives

Lorsque l'on veut étudier l'influence d'une variable aléatoire X comme dans l'exemple 6 de la Section 2.1.1, ou plus généralement d'un vecteur aléatoire $\boldsymbol{X} \in \mathbb{R}^k$ sur une variable aléatoire réelle Y, on part généralement de l'observation d'un n-échantillon

$$(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$$

de même loi que $(\boldsymbol{X}, Y)$. Formellement, on est dans le modèle du n-échantillon, mais avec une différence notoire : c'est la loi de Y qui nous intéresse, les $\boldsymbol{X}_i$ n'étant que des observations auxiliaires. Les $\boldsymbol{X}_i$ portent le nom de covariables, ou variables explicatives.

On peut postuler une représentation du type

$$Y = r(\boldsymbol{X}) + \xi, \tag{2.5}$$

où $r : \mathbb{R}^k \to \mathbb{R}$ est la fonction de régression $r(\boldsymbol{x}) = \mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x} \right]$ qui est la meilleure approximation de Y par une variable aléatoire $\boldsymbol{X}$-mesurable au sens suivant :

$$\mathbb{E}\left[\big(Y - r(\boldsymbol{X})\big)^2 \right] = \min_h \mathbb{E}\left[\big(Y - h(\boldsymbol{x})\big)^2 \right]$$

où le minimum est pris sur les fonctions boréliennes de $\mathbb{R}^k$ dans $\mathbb{R}$, comme nous l'avons déja mentionné dans l'exemple 6-influence d'une variable sur une autre.

On est alors dans une situation tout à fait analogue avec celle du paragraphe précédent, à la différence près que le statisticien ne choisit pas le « design »

$$(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).$$

Cela a des incidences pratiques bien entendu, mais d'un point de vue mathématique, on peut faire une hypothèse relativement faible qui permet d'unifier les deux points de vue :

Hypothèse 2.1 (Ancillarité du « design »)**.** *La loi de $\boldsymbol{X}$ ne dépend pas de ϑ.*

Autrement dit, toute l'information de sur la loi de Y que porte $r(\boldsymbol{X})$ est contenue dans la fonction de régression $r(\bullet)$. Dans ce cas, puisque les $\boldsymbol{X}_i$ sont observées et que leur loi ne dépend pas de ϑ, *on peut oublier ou ignorer le caractère aléatoire des $\boldsymbol{X}_i$* et raisonner dans toute la suite conditionnellement aux $\boldsymbol{X}_i = \boldsymbol{x}_i$, où les $\boldsymbol{x}_i$ sont les valeurs observées[12].

Sous l'Hypothèse 2.1, le modèle de régression avec variables explicative coïncide avec le modèle de régression conditionnelle et les formules du paragraphe précédent sont valides dans ce contexte.

Régression logistique

Si l'on veut étudier l'influence d'un vecteur $\boldsymbol{X}$ sur une variable qualitative $Y \in \{0, 1\}$ comme pour l'étude du risque de maladie coronarienne de l'exemple 6, l'écriture du modèle de régression (2.5) prend la forme

$$Y = r(\boldsymbol{X}) + \xi = \mathbb{P}\left[Y = 1 | \, \boldsymbol{X}\right] + \xi,$$

avec $\xi = Y - \mathbb{P}\left[Y = 1 | \, \boldsymbol{X}\right]$ qui vérifie bien $\mathbb{E}\left[\xi\right] = 0$.

Dans un cadre paramétrique, un choix populaire de la fonction $r(\vartheta, \bullet) : \mathbb{R}^k \to [0, 1]$ se fait de la manière suivante : on se donne un difféomorphisme $\psi : \mathbb{R} \to (0, 1)$. Dans ce cas, on peut forcer un modèle linéaire du type

$$r(\vartheta, \boldsymbol{x}) = \psi(\vartheta^T \boldsymbol{x}), \quad \vartheta \in \mathbb{R}^d, \quad x \in \mathbb{R}^k$$

avec $k \leq d \leq k + 1$. Un exemple incontournable pour les applications est celui de la fonction logistique

$$\psi(x) = \frac{e^x}{1 + e^x}, \quad x \in \mathbb{R},$$

sur lequel nous reviendrons au Chapitre 5.

[12] On reviendra ce point de vue dans le Chapitres 5.

Deuxième partie

Méthodes d'estimation

Chapitre 3

Echantillonnage et fonction de répartition empirique

3.1 Introduction

3.1.1 Situation

Nous étudions dans ce chapitre le problème très général qui consiste à « quantifier » l'information fournie par l'observation d'un n-échantillon d'une loi F sur $\mathbb{R}$, sans faire aucune (ou presque aucune) hypothèse sur cette loi. Ce chapitre est aussi un prétexte pour introduire les différentes problématiques du cours : estimation, tests et régions de confiance, point de vue asymptotique.

Le terme « quantifier » utilisé plus haut est imprécis ; nous le qualifierons à travers la construction d'estimateurs de F – ou de fonctionnelles $T(F) \in \mathbb{R}$ de F – et de leur précision d'estimation, ce qui nous amènera à parler de région (et d'intervalles) de confiance. Nous considèrerons aussi brièvement le problème de test d'hypothèse : à partir de l'observation, décider si la loi F vérifie une propriété donnée. De manière générale, nous étudierons comment la qualité des procédures statistiques augmente avec le nombre d'observations n. Nous comparerons les points de vue asymptotique (dans la limite $n \to \infty$) et non-asymptotique.

Ici, la structure probabiliste de l'expérience statistique est très simple (variables aléatoires indépendantes et identiquement distribuées) mais l'ensemble des paramètres[1] est énorme ! De ce point de vue, l'expérience statistique considéree est non-paramétrique. Nous développerons plus systématiquement des méthodes lorsque l'on fait des hypothèses supplémentaires sur l'ensemble des paramètres dans les chapitres suivants.

[1] c'est-à-dire l'ensemble de toutes les lois de probabilités F sur $\mathbb{R}$.

3.1.2 Notations et définitions préliminaires

On observe un n-échantillon

$$X_1, \dots, X_n$$

noté le plus souvent

$$\big(X_1, \dots, X_n\big)^T$$

de loi inconnue F sur $\mathbb{R}$. On ne fait pas d'hypothèse particulière sur la loi commune des X_i. L'expérience statistique sous-jacente, au sens de la Définition 2.3 du Chapitre 2, s'écrit

$$\mathcal{E}^n = \big(\mathbb{R}^n, \mathcal{B}^n, (\mathbb{P}_F^n, F \in \mathfrak{F})\big),$$

où

$$\mathfrak{F} = \big\{F,\ F \text{ fonction de répartition}\big\}$$

et $\mathbb{P}_F^n$ est la loi sur $\mathbb{R}^n$ de n variables aléatoires indépendantes de loi F. En particulier, pour tous $x_1, \dots, x_n \in \mathbb{R}$, on a

$$\mathbb{P}_F^n\big[X_1 \leq x_1, \dots, X_n \leq x_n\big] = \prod_{i=1}^n F(x_i).$$

On écrira parfois $\mathbb{P}_F$ ou $\mathbb{P}$ à la place de $\mathbb{P}_F^n$ lorsqu'il n'y aura pas de risque de confusion. On écrit aussi X pour l'une quelconque des X_i lorsque l'indice ne joue pas de rôle.

Remarque 3.1. Ici, l'ensemble des paramètres est « énorme ». En particulier, la famille de distributions $\mathfrak{F}$ n'est pas dominée (puisqu'elle contient par exemple toutes les mesures de Dirac δ_x, $x \in \mathbb{R}$).

Définition 3.1. *Une statistique, ou une procédure statistique, ou encore un estimateur, associé(e) à l'expérience $\mathcal{E}^n$, est une fonction mesurable des observations $X_1, \dots, X_n$.*

Lorsque l'on cherche à estimer une fonctionnelle $T(F) \in \mathbb{R}$ de F, un estimateur est souvent noté $\widehat{T}_n$. C'est une variable aléatoire, ne dépendant que de $X_1, \dots, X_n$, qui s'écrit donc $\widehat{T}_n = g_n(X_1, \dots, X_n)$, pour une certaine fonction borélienne $g_n : \mathbb{R}^n \to \mathbb{R}$. Se donner un estimateur, c'est se donner une telle fonction $g_n(\bullet)$.

3.2 Estimation ponctuelle

Soit $x_0 \in \mathbb{R}$. A partir de l'observation $X_1, \dots, X_n$, que pouvons-nous dire de

$$F(x_0) = \mathbb{P}\big[X \leq x_0\big] \ ?$$

3.2.1 Fonction de répartition empirique

L'idée la plus immédiate est d'estimer $F(x_0)$ par la fréquence empirique du nombre de points X_i dans l'intervalle $(-\infty, x_0]$

$$\frac{1}{n}\mathrm{Card}\Big\{X_i \in (-\infty, x_0],\ \ i = 1, \dots, n\Big\}$$

qui se rapproche de la fréquence théorique $\mathbb{P}\big[X \leq x_0\big]$ par la loi des grands nombres.

Définition 3.2. *La fonction de répartition empirique de l'échantillon $(X_1, \dots, X_n)$ est définie par*

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}, \quad x \in \mathbb{R}\,.$$

Dans la suite, nous estimerons $F(x_0)$ par $\widehat{F}_n(x_0)$.

Proposition 3.1. *Pour tout $x_0 \in \mathbb{R}$, on a*

$$\mathbb{E}\left[\widehat{F}_n(x_0)\right] = F(x_0),$$

$$\mathrm{Var}\big[\widehat{F}_n(x_0)\big] = \mathbb{E}\left[\big(\widehat{F}_n(x_0) - \mathbb{E}[\widehat{F}_n(x_0)]\big)^2\right] = \frac{F(x_0)\big(1 - F(x_0)\big)}{n}.$$

En particulier, on a $\widehat{F}_n(x_0) \xrightarrow{\mathcal{L}^2} F(x_0)$ et donc $\widehat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$.

Démonstration. Les variables aléatoires $1_{\{X_i \leq x_0\}}$ son indépendantes, de loi de Bernoulli de paramètre $\mathbb{P}[X_i \leq x_0] = F(x_0)$. Donc $n\widehat{F}_n(x_0)$ est une variables aléatoire binômiale, de paramètres $\big(n, F(x_0)\big)$. Son espérance et sa variance valent respectivement $nF(x_0)$ et $nF(x_0)\big(1 - F(x_0)\big)$. On obtient la proposition en divisant par n, et en utilisant le fait que l'espérance est linéaire et la variance quadratique. □

Remarque 3.2. La loi forte des grands nombres garantit immédiatement la convergence $\widehat{F}_n(x_0) \xrightarrow{\text{p.s.}} F(x_0)$.

3.2.2 Précision d'estimation

La Proposition 3.1 fournit un résultat de convergence en apparence très fort : si $\ell(x, y) = (x - y)^2$, avec $x, y \in \mathbb{R}$ désigne la *perte quadratique*, on a

$$\sup_{F \in \mathfrak{F}} \mathbb{E}\left[\ell\big(\widehat{F}_n(x_0), F(x_0)\big)\right] = \frac{1}{4n}, \tag{3.1}$$

Il suffit pour voir cela d'appliquer la deuxième partie de la Proposition 3.1 en utilisant le fait que

$$\sup_{F \in \mathfrak{F}} F(x_0)\big(1 - F(x_0)\big) = 1/4. \tag{3.2}$$

Cela signifie que, pour la perte quadratique, l'estimateur $\widehat{F}_n(x_0)$ approche $F(x_0)$ uniformément en F à vitesse $\sqrt{n}$. Ce résultat est-il optimal, et dans quel sens ? Comment le relier à une notion de précision d'estimation ? Si $F(x_0)$ est proche de 0 ou 1, ce qui peut nous être suggéré par la lecture de $\widehat{F}_n(x_0)$, peut-on améliorer le facteur $1/4$ dans (3.1) et améliorer la précision d'estimation ?

Une manière d'aborder la précision d'estimation consiste à construire un intervalle de confiance à partir de la borne (3.1) de la façon suivante : on a, pour tout $t > 0$

$$\mathbb{P}\left[|\widehat{F}_n(x_0) - F(x_0)| \geq t\right] \leq \frac{1}{t^2}\mathrm{Var}\left[\widehat{F}_n(x_0)\right] \leq \frac{1}{4nt^2}$$

par l'inégalité de Tchebychev (1.2). Choisissons $\alpha \in (0,1)$, et prenons $t = t(\alpha, n)$ le plus petit possible de sorte que $1/(4nt^2) \leq \alpha$. Ceci nous fournit le choix

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

On en déduit que l'intervalle[2]

$$\mathcal{I}_{n,\alpha} = \left[\widehat{F}_n(x_0) \pm \frac{1}{2\sqrt{n\alpha}}\right]$$

contient $F(x_0)$ avec probabilité plus grande que $1 - \alpha$.

Définition 3.3. *L'intervalle $\mathcal{I}_{n,\alpha}$ est appelé intervalle de confiance pour la valeur $F(x_0)$ au niveau $1 - \alpha$. La propriété*

$$\mathbb{P}\left[F(x_0) \in \mathcal{I}_{n,\alpha}\right] \geq 1 - \alpha$$

s'appelle « propriété de couverture » (coverage property).

Remarque 3.3. Un intervalle de confiance est aléatoire. Il est observable (c'est-à-dire construit à partir des observations) et ne peut dépendre de la quantité inconnue $F(x_0)$ qu'à travers la loi des observations $X_1, \ldots, X_n$.

L'interprétation de $\mathcal{I}_{n,\alpha}$ est claire : on imagine α petit[3] et on garantit avec probabilité $1 - \alpha$ que la quantité inconnue d'intérêt $F(x_0)$ appartient à $\mathcal{I}_{n,\alpha}$ que l'on observe.

Mais sans autre indication sur $\mathcal{I}_{n,\alpha}$, cette information n'a que peu d'intérêt. On s'attend à ce que la longueur $|\mathcal{I}_{n,\alpha}|$ de l'intervalle, qui joue le rôle de précision d'estimation de $F(x_0)$, soit petite lorsque n est grand[4]. On a

$$|\mathcal{I}_{n,\alpha}| = \frac{1}{2\sqrt{n\alpha}}$$

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

[3]La tradition dicte 5%, mais d'autres choix sont évidemment pertinents.

[4]Sinon, l'intervalle trivial $\mathcal{I}_{n,\alpha} = \mathbb{R}$ (ou même $\mathcal{I}_{n,\alpha} = [0,1]$ puisque $0 \leq F(x_0) \leq 1$) a la propriété de couverture au niveau de confiance 1 !

que l'on interprète comme la précision d'estimation au niveau de confiance $1-\alpha$.

L'ordre de grandeur de $\mathcal{I}_{n,\alpha}$ en n est $1/\sqrt{n}$, comme pour la perte quadratique. Mais on a aussi $|\mathcal{I}_{n,\alpha}| \to +\infty$ lorsque $\alpha \to 0$. Il s'agit d'un compromis inévitable entre précision d'estimation (vouloir $|\mathcal{I}_{n,\alpha}|$ petit) et risque (vouloir α petit) qui sont antagonistes.

Nous allons explorer plusieurs façons d'améliorer ce résultat.

3.2.3 Précision d'estimation asymptotique

Une manière de juger de la pertinence de la précision d'un estimateur est de se placer dans le régime asymptotique $n \to \infty$ et d'étudier la loi asymptotique de l'erreur renormalisée

$$\sqrt{n}\big(\widehat{F}_n(x_0) - F(x_0)\big), \quad n \to \infty,$$

la normalisation par $\sqrt{n}$ étant suggérée[5] par la Proposition 3.1.

Proposition 3.2. *On a*

$$\xi_n = \sqrt{n}\frac{\widehat{F}_n(x_0) - F(x_0)}{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

De plus, pour tout $\alpha \in (0,1)$,

$$\mathbb{P}\left[\xi_n \in \left[-\Phi^{-1}(1-\alpha/2), \Phi^{-1}(1-\alpha/2)\right]\right] \to 1-\alpha,$$

où $\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2}\frac{dt}{\sqrt{2\pi}}$ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

Démonstration. Le théorème central-limite donne la convergence

$$\sqrt{n}\frac{\widehat{F}_n(x_0) - F(x_0)}{F(x_0)^{1/2}\big(1-F(x_0)\big)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

La Proposition 3.1 assure que $\widehat{F}_n(x_0)\big(1-\widehat{F}_n(x_0)\big) \xrightarrow{\mathbb{P}} F(x_0)\big(1-F(x_0)\big)$. On en déduit la première partie en appliquant la Proposition 1.8 (Slutsky).

Puisque $\xi_n \xrightarrow{d} \mathcal{N}(0,1)$, on a

$$\begin{aligned}\mathbb{P}\left[\xi_n \in \left[-\Phi^{-1}(1-\tfrac{\alpha}{2}), \Phi^{-1}\left(1-\tfrac{\alpha}{2}\right)\right]\right] &\to \Phi\big(\Phi^{-1}(1-\tfrac{\alpha}{2})\big) - \Phi\big(-\Phi^{-1}(1-\tfrac{\alpha}{2})\big) \\ &= 1-\alpha\end{aligned}$$

en utilisant $\Phi(-x) = 1-\Phi(x)$ puisque la loi $\mathcal{N}(0,1)$ est symétrique (Définition 1.4). □

[5] D'après la Proposition 3.1, $\mathbb{E}\left[\big(\sqrt{n}\big(\widehat{F}_n(x_0) - F(x_0)\big)\big)^2\right]$ est constante, donc $\big(\sqrt{n}\big(\widehat{F}_n(x_0) - F(x_0)\big)\big)^2$, et par suite $\sqrt{n}\big(\widehat{F}_n(x_0) - F(x_0)\big)$ est « en moyenne de l'ordre de grandeur de 1 en n ».

On peut interpréter le second point de la Proposition 3.2 de la façon suivante : lorsque « n est grand »,

$$\sqrt{n}\frac{\widehat{F}_n(x_0)-F(x_0)}{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}}\in\big[-\Phi^{-1}(1-\tfrac{\alpha}{2}),\Phi^{-1}(1-\tfrac{\alpha}{2})\big]$$

avec probabilité proche de $1-\alpha$. En isolant $F(x_0)$ dans cette relation et en posant

$$\mathcal{J}_{n,\alpha}=\left[\widehat{F}_n(x_0)\pm\frac{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}}{\sqrt{n}}\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right],$$

la quantité $F(x_0)$ inconnue est dans l'intervalle $\mathcal{J}_{n,\alpha}$ avec probabilité proche de $1-\alpha$ dans la limite $n\to\infty$.

Définition 3.4. *L'intervalle $\mathcal{J}_{n,\alpha}$ est appelé intervalle de confiance asymptotique de $F(x_0)$ au niveau $1-\alpha$. La propriété*

$$\mathbb{P}\left[F(x_0)\in\mathcal{J}_{n,\alpha}\right]\to 1-\alpha,\ \ n\to\infty$$

s'appelle « propriété de couverture asymptotique ».

La précision asymptotique de $\mathcal{J}_{n,\alpha}$ est

$$|\mathcal{J}_{n,\alpha}|=2\frac{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}}{\sqrt{n}}\Phi^{-1}(1-\tfrac{\alpha}{2}).$$

L'ordre de grandeur de $\mathcal{J}_{n,\alpha}$ en n est $1/\sqrt{n}$, comme pour l'intervalle de confiance $\mathcal{I}_{n,\alpha}$ construit avec la perte quadratique. On a aussi $\Phi^{-1}(1-\alpha/2)\to\infty$ lorsque $\alpha\to 0$. Par contre,

$$\Phi^{-1}(1-\tfrac{\alpha}{2})\ll\sqrt{\alpha},\quad\alpha\to 0.$$

voir Exercice 3.1 C'est aussi un résultat plus précis en apparence que celui obtenu à l'aide de $\mathcal{I}_{n,\alpha}$ puisqu'on a remplacé le facteur $1/2$ de obtenu en prenant la racine de (3.2) dans la construction de $\mathcal{I}_{n,\alpha}$ par

$$\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}\le\frac{1}{2}$$

dans la construction de $\mathcal{J}_{n,\alpha}$. Cependant, cette amélioration n'est valide que dans le régime asymptotique $n\to\infty$.

3.2.4 Précision non-asymptotique

Nous cherchons un résultat de qualité comparable à celui de la Proposition 3.2 mais valable à n fixé.

Dans l'approche non-asymptotique à l'aide de la perte quadratique, on a perdu en utilisant l'inégalité de Markov qui s'appuie uniquement sur le contrôle de la variance de $\hat{F}_n(x)$. Le résultat suivant fournit un contrôle plus fin de la probabilité de déviation de la moyenne empirique.

Théorème 3.1 (Inégalité de Hoeffding)**.** *Soient $Y_1, \ldots, Y_n$ des variables aléatoires réelles indépendantes telles que $\mathbb{E}[Y_i] = 0$ et $a_i \leq Y_i \leq b_i$. Soit $t > 0$. Alors, pour tout $\lambda > 0$*

$$\mathbb{P}\big[\sum_{i=1}^{n} Y_i \geq t\big] \leq e^{-\lambda t} \prod_{i=1}^{n} \exp\big(\lambda^2 \frac{(b_i - a_i)^2}{8}\big).$$

Démonstration. Si Y est une variable aléatoire à valeurs dans $[a, b]$, posons

$$\psi_Y(\lambda) = \log \mathbb{E}\big[\exp\big(\lambda(Y - \mathbb{E}[Y])\big)\big], \quad \lambda > 0.$$

La fonction $\lambda \rightsquigarrow \psi_Y(\lambda)$ est deux fois dérivable et, puisque $\mathbb{E}\big[Y\big] = 0$, un calcul élémentaire conduit à

$$\psi_Y''(\lambda) = e^{-\psi_Y(\lambda)}\, \mathbb{E}\big[Y^2 \exp\big(\lambda Y\big)\big] - e^{-2\psi_Y(\lambda)} \Big(\mathbb{E}\big[Y \exp\big(\lambda Y\big)\big]\Big)^2. \tag{3.3}$$

Posons, pour $A \in \mathcal{B}$, $\mathbb{Q}\big[A\big] = e^{-\psi_Y(\lambda)}\, \mathbb{E}\big[\exp\big(\lambda Y\big) 1_A\big]$, de sorte que $\mathbb{Q}$ est une mesure de probabilité. Alors on peut interpréter (3.3) de la manière suivante :

$$\psi_Y''(\lambda) = \mathrm{Var}\big[Z\big],$$

où Z est une variable aléatoire à valeurs dans $[a, b]$ de loi $\mathbb{Q}$. Maintenant, pour toute variable Z à valeurs dans $[a, b]$, on a toujours

$$\Big|Z - \frac{b+a}{2}\Big| \leq \frac{b-a}{2},$$

et donc

$$\mathrm{Var}\big[Z\big] = \mathrm{Var}\big[Z - (b+a)/2\big] \leq \frac{(b-a)^2}{4}. \tag{3.4}$$

En intégrant (3.4), on déduit $\psi_Y''(\lambda) \leq (b-a)^2/4$, d'où

$$\psi_Y(\lambda) \leq \lambda^2 \frac{(b-a)^2}{8} \tag{3.5}$$

en utilisant $\psi_Y(0) = \psi'_Y(0) = 0$. Finalement, pour tous $t, \lambda > 0$,

$$\begin{aligned}
\mathbb{P}\big[\sum_{i=1}^n Y_i \geq t\big] &= \mathbb{P}\big[\exp\big(\lambda \sum_{i=1}^n Y_i\big) \geq \exp(\lambda t)\big] \\
&\leq e^{-\lambda t}\, \mathbb{E}\big[\exp\big(\lambda \sum_{i=1}^n Y_i\big)\big] \quad \text{(inégalité de Tchebychev)} \\
&= e^{-\lambda t} \prod_{i=1}^n \mathbb{E}\big[\exp\big(\lambda Y_i\big)\big] \quad \text{(indépendance des } Y_i\text{)} \\
&= e^{-\lambda t} \prod_{i=1}^n \exp\big(\psi_{Y_i}(\lambda)\big),
\end{aligned}$$

Puisque chaque Y_i est centrée et à valeurs dans $[a_i, b_i]$, on conclut en appliquant l'inégalité (3.5) à chaque $\psi_{Y_i}(\lambda)$. □

Corollaire 3.1. *Si $X_1, \ldots, X_n$ sont des variables aléatoire de Bernoulli de paramètre p et si $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, alors, pour tout $t > 0$*

$$\mathbb{P}\big[|\overline{X}_n - p| \geq t\big] \leq 2\exp\big(-2nt^2\big).$$

Démonstration. Appliquons l'inégalité de Hoeffding à $Y_i = X_i - p$. Les conditions du Théorème 3.1 sont vérifiées avec $b_i - a_i = 1$. Le choix $\lambda = 4t/n$ conduit à

$$\mathbb{P}\big[\sum_{i=1}^n Y_i \geq t\big] \leq \exp\big(-2t^2/n\big), \tag{3.6}$$

soit encore

$$\mathbb{P}\big[\overline{X}_n - p \geq t\big] = \mathbb{P}\big[\sum_{i=1}^n Y_i \geq nt\big] \leq \exp\big(-2nt^2\big).$$

De même

$$\mathbb{P}\big[\overline{X}_n - p \leq -t\big] = \mathbb{P}\big[\sum_{i=1}^n (-Y_i) \geq nt\big] \leq \exp\big(-2nt^2\big)$$

en appliquant (3.6) à $-Y_i$. On conclut en écrivant

$$\mathbb{P}\big[|\,\overline{X}_n - p| \geq t\big] = \mathbb{P}\big[\overline{X}_n - p \geq t\big] + \mathbb{P}\big[\overline{X}_n - p \leq -t\big].$$

□

On en déduit un intervalle de confiance non-asymptotique pour $F(x_0)$.

Proposition 3.3. *Pour tout $\alpha > 0$,*

$$\mathcal{I}^\star_{n,\alpha} = \left[\widehat{F}_n(x_0) \pm \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right]$$

est un intervalle de confiance pour $F(x_0)$ de niveau $1-\alpha$.

Démonstration. On applique le Corollaire 3.1 aux $1_{\{X_i \le x_0\}}$ qui sont des variables aléatoires de Bernoulli indépendantes, de paramètre $F(x_0)$. On a, pour tout $t > 0$

$$\mathbb{P}\left[\left|\widehat{F}_n(x_0) - F(x_0)\right| > t\right] \le 2\exp\big(-2nt^2\big).$$

On cherche $t = t(\alpha, n)$ le plus petit possible de sorte que $2\exp(-2nt^2) \le \alpha$, ce qui donne

$$t(\alpha, n) = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}.$$

□

Remarque 3.4. On a

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

où $\mathcal{I}_{n,\alpha} = \left[\widehat{F}_n(x_0) \pm \frac{1}{2\sqrt{n\alpha}}\right]$ est l'intervalle de confiance construit à l'aide de l'inégalité de Tchebychev dans la Section 3.2.2. Le gain est significatif. Par exemple, pour $\alpha = 5\%$, on a un rapport de

$$\frac{|\mathcal{I}^\star_{n,\alpha}|}{|\mathcal{I}_{n,\alpha}|} = 1,65.$$

Pour $\alpha = 1\%$, le rapport devient 0.32, soit une précision plus de 3 fois meilleure !

Remarque 3.5. Par contre, les ordres de grandeurs de $\mathcal{J}_{n,\alpha}$ et $\mathcal{I}^\star_{n,\alpha}$ sont comparables en n et en α, voir Exercice 3.1. De ce point de vue, l'intervalle $\mathcal{I}^\star_{n,\alpha}$ est satisfaisant.

3.2.5 Décision*

Notion de test et d'erreur de test

Soit F_0 une distribution donnée. On souhaite répondre à la question suivante : en vue d'un n-échantillon $X_1, \ldots, X_n$ de loi $F \in \mathfrak{F}$, est-ce que

$$F(x_0) = F_0(x_0) \text{ ou non ?}$$

On formule le problème de la manière suivante. On contruit à partir des observations une procédure (un estimateur)

$$\varphi_n = \varphi_n(X_1, \dots, X_n) \in \{0, 1\}$$

ne prenant que les valeurs 0 ou 1. La valeur $\{\varphi_n = 0\}$ correspondra à la réponse « oui » à la question, et la valeur $\{\varphi_n = 1\}$ correspondra à la réponse « non » .

On dira que l'on teste l'hypothèse nulle

$$H_0 : \quad F(x_0) = F_0(x_0),$$

contre l'alternative

$$H_1 : \quad F(x_0) \neq F_0(x_0).$$

Si φ_n est une procédure ne prenant que les valeurs 0 ou 1, on dira que φ_n est un test simple[6]. Si φ_n est un test simple, il se représente sous la forme

$$\varphi_n = \varphi_n(X_1, \dots, X_n) = 1_{\left\{(X_1 \dots, X_n) \in \mathcal{R}_n\right\}}$$

où $\mathcal{R}_n \subset \mathbb{R}^n$ est un sous-ensemble de l'espace des observations.

Définition 3.5. *L'ensemble $\mathcal{R}_n$ associé au test simple φ_n est appelé zone de rejet du test, ou encore région critique du test.*

Remarque 3.6. On définit aussi parfois la zone de rejet comme l'événement

$$\big\{(X_1, \dots, X_n) \in \mathcal{R}_n\big\}.$$

Cela n'a aucune importance : il n'y a jamais d'ambiguité[7].

Lorsque l'on procède à un test, on décide d'accepter l'hypothèse H_0 (lorsque l'événement $\big\{\varphi_n = 0\big\}$ est réalisé) ou de la rejeter (lorsque l'événement $\big\{\varphi_n = 1\big\}$ est réalisé). On peut avoir raison de deux manières : accepter l'hypothèse H_0 alors qu'elle est vraie[8] ou bien rejeter l'hypothèse H_0 alors qu'elle est fausse[9].

Mais surtout, on peut aussi se tromper de deux manières : rejeter H_0 alors qu'elle est vraie ou encore accepter H_0 alors qu'elle est fausse. Ce sont ces deux erreurs que l'on va chercher à rendre petites simultanément.

Pour cela, nous devons définir précisément les conditions

$$F(x_0) = F_0(x_0) \quad \text{et} \quad F(x_0) \neq F_0(x_0).$$

[6] On pourrait envisager des tests plus complexes, où une réponse intermédiaire entre 0 et 1 est possible.

[7] et la notion d'expérience canonique, voir Section 2.2.2 du Chapitre 2 permet d'ailleurs de concilier les deux points de vue de façon rigoureuse. Nous ne reviendrons plus sur ce point dans la suite du cours.

[8] c'est-à-dire observer $\{\varphi_n = 0\}$ et avoir $F(x_0) = F_0(x_0)$.

[9] c'est-à-dire observer $\{\varphi_n = 1\}$ et avoir $F(x_0) \neq F_0(x_0)$.

L'expérience statistique engendrée par les observations a pour ensemble de paramètres

$$\mathfrak{F} = \{F,\ F \text{ fonction de répartition}\}.$$

Posons

$$\mathfrak{F}_0 = \{F \in \mathfrak{F},\ F(x_0) = F_0(x_0)\}.$$

Alors l'hypothèse H_0 se traduit par le sous-ensemble de paramètres $\mathfrak{F}_0$, et l'alternative H_1 par le sous-ensemble de paramètre $\mathfrak{F} \setminus \mathfrak{F}_0$.

Définition 3.6. *Soit $\alpha \in [0,1]$. Le test φ_n est de niveau α (respectivement, asymptotiquement de niveau α) si*

$$\sup_{F\in\mathfrak{F}_0} \mathbb{P}_F\left[\varphi_n = 1\right] \leq \alpha \quad \textit{(respectivement} \quad \limsup_{n\to\infty} \sup_{F\in\mathfrak{F}_0} \mathbb{P}_F\left[\varphi_n = 1\right] \leq \alpha).$$

Autrement dit, si le niveau d'un test est inférieur à α, la probabilité de rejeter l'hypothèse (observer $\{\varphi_n = 1\}$) alors qu'elle est vraie ($F \in \mathfrak{F}_0$) est inférieure ou égale à α. On parle indifféremment d'erreur de première espèce du test φ_n ou de niveau du test φ_n.

Remarque 3.7. Bien que cela ne transparaîsse pas dans les notations, le test φ_n dépend de α en général.

Définition 3.7. *La puissance d'un test φ_n est l'application de $\mathfrak{F} \setminus \mathfrak{F}_0$ dans $[0,1]$ définie par*

$$F \in \mathfrak{F} \setminus \mathfrak{F}_0 \rightsquigarrow \mathbb{P}_F\left[\varphi_n = 1\right]$$

On parle indifféremment de « puissance du test » ou bien de « fonction d'erreur de seconde espèce », définie par

$$F \in \mathfrak{F} \setminus \mathfrak{F}_0 \rightsquigarrow 1 - \mathbb{P}_F\left[\varphi_n = 1\right].$$

La démarche sera la suivante : on se fixe un niveau de risque α, et on cherche un test φ_n de niveau α (d'erreur de première espèce inferieure ou égale à α) qui a la plus grande puissance possible (l'erreur de seconde espèce la plus petite possible). On étudiera systématiquement ces notions aux Chapitres 7 et 8.

Construction de tests

A partir d'estimateurs et d'intervalles de confiance de niveau $1-\alpha$, la construction d'un test φ_n est naturelle. On se restreint ici par simplicité au cadre asymptotique. On a, d'après la construction de la Section 3.2.3, pour tout $F \in \mathfrak{F}$,

$$\mathbb{P}_F\left[F(x_0) \in \mathcal{J}_{n,\alpha}\right] \to 1 - \alpha.$$

Ceci suggère la règle de décision suivante : on accepte H_0 si $F_0(x_0) \in \mathcal{J}_{n,\alpha}$ et on rejette H_0 sinon.

Proposition 3.4. *Soit $\alpha \in (0,1)$. Le test $\varphi_n = \varphi_{n,\alpha}$ de l'hypothèse nulle $H_0 : F(x_0) = F_0(x_0)$ contre l'alternative $F(x_0) \neq F_0(x_0)$ défini par la zone de rejet*

$$\mathcal{R}_{n,\alpha} = \big\{F_0(x_0) \notin \mathcal{J}_{n,\alpha}\big\}$$

est asymptotiquement de niveau α. De plus, pour tout point de l'alternative $F \in \mathfrak{F} \setminus \mathfrak{F}_0$, on a

$$\mathbb{P}_F\left[\varphi_{n,\alpha} = 0\right] = \mathbb{P}_F\left[(X_1,\dots,X_n) \notin \mathcal{R}_{n,\alpha}\right] \to 0.$$

Autrement dit, l'erreur de première espèce est asymptotiquement plus petite que α et l'erreur de seconde espèce tend vers 0 ; ou encore, la puissance du test tend vers 1 en tout point de l'alternative. On dit que le test est consistant ou convergent.

Démonstration. La première partie de la proposition découle de la propriété de couverture asymptotique de $\mathcal{J}_{n,\alpha}$ (le second point de la Proposition 3.2). Pour le contrôle de l'erreur de seconde espèce, si $F \in \mathfrak{F} \setminus \mathfrak{F}_0$, alors

$$\widehat{F}_n(x_0) \xrightarrow{\mathbb{P}_F} F(x_0) \neq F_0(x_0),$$

Ceci suggère la décomposition

$$\begin{aligned}
&\sqrt{n}\frac{\widehat{F}_n(x_0) - F_0(x_0)}{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}} \\
=&\sqrt{n}\frac{\widehat{F}_n(x_0) - F(x_0)}{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}} + \sqrt{n}\frac{F(x_0) - F_0(x_0)}{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}}.
\end{aligned}$$

Le premier terme tend en loi sous $\mathbb{P}_F$ vers un gaussienne centrée réduite d'après la Proposition 3.2. Le second terme diverge vers $\pm\infty$ lorsque $n \to \infty$. Puisque

$$\big\{\varphi_{n,\alpha} = 0\big\} = \left\{\sqrt{n}\left|\frac{\widehat{F}_n(x_0) - F_0(x_0)}{\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}}\right| \leq \Phi^{-1}\left(1-\tfrac{\alpha}{2}\right)\right\}$$

on a $\varphi_{n,\alpha} \to 1$ en $\mathbb{P}_F$-probabilité si $F \in \mathfrak{F} \setminus \mathfrak{F}_0$. Ceci implique[10] $\mathbb{P}_F\left[\varphi_{n,\alpha} = 0\right] \to 0$. □

La question de l'optimalité d'une telle construction sera discutée dans le Chapitre 8.

[10] par exemple par convergence dominée, ou plus simplement parce que la suite de variable aléatoires discrètes $\varphi_{n,\alpha}$ tend en probabilité vers 1, donc en loi vers la loi dégénérée $\delta_1(dx)$, ce qui entraîne la convergence voulue.

3.3 Estimation uniforme

Les trois problèmes développés précédemment, estimation, intervalle de confiance et test, que ce soit d'un point de vue asymptotique ou non, ne font intervenir la distribution F qu'en un point x_0 donné. Ceci est peut satisfaisant si l'on envisage F globalement.

Nous reprenons la problématique de la Section 3.2 simultanément pour toutes les valeurs possibles de $(F(x),\, x \in \mathbb{R})$. A partir de l'observation de $(X_1, \ldots, X_n)$, que peut-on dire de

$$\big(F(x), x \in \mathbb{R}\big) \ ?$$

3.3.1 Estimation uniforme

Théorème 3.2 (Glivenko-Cantelli)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires réelle indépendantes, de même loi F, et $\widehat{F}_n$ leur fonction de répartition empirique. Alors*

$$\sup_{x\in\mathbb{R}} \big|\widehat{F}_n(x) - F(x)\big| \overset{\text{p.s.}}{\to} 0, \quad n \to \infty.$$

Démonstration. Soit $k \geq 1$ un entier, et pour tout $0 \leq k \leq m$,

$$x_k^m = \inf\{x \in \mathbb{R},\ F(x) \geq \tfrac{k}{m}\}.$$

(Les points x_k^m ne sont pas nécessairement distincts si F n'est pas continue.) Par construction, pour $0 \leq k \leq m-1$,

$$F\big(x_k^m\big) \geq \frac{k}{m} \geq F\big(x_k^m - \big)$$

car F est continue à droite, et donc

$$F\big(x_k^m\big) + \frac{1}{m} \geq F\big(x_{k+1}^m - \big).$$

Soit $x \in [x_k^m, x_k^{m+1})$. Puisque F et $\widehat{F}_n$ sont croissantes, on a, pour tout $n \geq 1$,

$$\widehat{F}_n\big(x_k^m\big) - F\big(x_{k+1}^m - \big) \leq \widehat{F}_n\big(x\big) - F\big(x\big) \leq \widehat{F}_n\big(x_{k+1}^m - \big) - F\big(x_k^m\big),$$

et aussi, d'après ce ce qui précède

$$\widehat{F}_n\big(x_k^m\big) - F\big(x_{k+1}^m - \big) - \frac{1}{m} \leq \widehat{F}_n\big(x\big) - F\big(x\big) \leq \widehat{F}_n\big(x_{k+1}^m - \big) - F\big(x_{k+1}^m - \big) + \frac{1}{m}.$$

Il vient

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} \big|\widehat{F}_n(t) - F(t)\big| \\
&\leq \max\left\{ \max_{0\leq k\leq m} \big|\widehat{F}_n\big(x_k^m\big) - F\big(x_k^m\big)\big|, \max_{0\leq k\leq m} \big|\widehat{F}_n\big(x_k^m - \big) - F\big(x_k^m - \big)\big| \right\} + \frac{1}{m}.
\end{aligned}$$

On a $\widehat{F}_n(x) \overset{\text{p.s.}}{\to} F(x)$ par la loi forte des grands nombres. Il existe donc un ensemble négligeable $\mathcal{N}'(m)$ en dehors duquel

$$\max_{0\le k\le m} \left|\widehat{F}_n\big(x_k^m\big) - F\big(x_k^m\big)\right| \to 0.$$

De même, en appliquant la loi des grand nombres aux variables $1_{X_i<x}$, il existe une ensemble négligeable $\mathcal{N}''(m)$ en dehors duquel

$$\max_{0\le k\le m} \left|\widehat{F}_n\big(x_k^m - \big) - F\big(x_k^m - \big)\right| \to 0.$$

On en déduit qu'en dehors d'un ensemble négligeable $\mathcal{N}(m) = \mathcal{N}'(m) \cup \mathcal{N}''(m)$, on a

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \le \frac{1}{m}.$$

Puis on fait tendre m vers l'infini :

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| = 0$$

en dehors de $\bigcup_{m\ge1} \mathcal{N}(m)$ qui est de probabilité 0. □

3.3.2 Vitesse d'estimation uniforme

Théorème 3.3 (Kolmogorov-Smirnov)**.** *Si la fonction de répartition F est continue, alors*

$$\sqrt{n} \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{(d)} \mathbb{B}$$

où $\mathbb{B}$ est une variable aléatoire dont la loi ne dépend pas de F, de fonction de répartition

$$\mathbb{P}\left[\mathbb{B} \le x\right] = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2x}, \quad x \ge 0.$$

Remarque 3.8. La variable aléatoire se représente comme $\mathbb{B} = \sup_{t\in[0,1]} B_t$, où $(B_t, t \in [0,1])$ est un processus aléatoire appelé pont brownien. Ce résultat découle de la théorie des processus empiriques et sa preuve dépasse le cadre de ce cours[11].

Nous admettons la convergence en loi de $\sqrt{n}\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right|$. Nous allons cependant démontrer que cette loi ne dépend pas de F, ce qui est très important en vue des applications statistiques.

[11] on pourra consulter, par exemple, le livre de van der Vaart [7] pour les liens entre statistique et processus empiriques.

Lemme 3.3.1. *Soit $U_1, \dots, U_n$ une suite de variables aléatoires indépendantes, uniformes sur $[0,1]$. On note G_n leur fonction de répartition empirique. Si F est continue, on a l'égalité en loi*

$$\sup_{x\in\mathbb{R}} \big|\widehat{F}_n(x) - F(x)\big| \stackrel{d}{=} \sup_{x\in\mathbb{R}} \big|G_n(x) - x\big|.$$

En particulier, la loi de $\mathbb{B}$ ne dépend pas de F.

Démonstration. Posons, $U_i = F(X_i)$. Alors les U_i sont des variables aléatoires uniformes sur $[0,1]$, et il existe un ensemble négligeable $\mathcal{N}_i$ tel que, pour tout $x \in \mathbb{R}$ et pour tout $\omega \notin \mathcal{N}_i$ on a

$$F\big(X_i(\omega)\big) \le F(x) \quad \text{si et seulement et} \quad X_i(\omega) \le x,$$

voir par exemple Méléard [4], paragraphe 4.2.4 p. 78. Donc, on peut écrire, pour tout $x \in \mathbb{R}$

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i\le x} = \frac{1}{n}\sum_{i=1}^{n} 1_{F(X_i)\le F(x)} = G_n\big(F(x)\big)$$

en dehors de $\mathcal{N} = \bigcup_i \mathcal{N}_i$ qui est encore négligeable. Il vient

$$\sup_{x\in\mathbb{R}} \big|\widehat{F}_n(x) - F(x)\big| = \sup_{x\in\mathbb{R}} \big|G_n\big(F(x)\big) - F(x)\big| = \sup_{x\in\mathbb{R}} \big|G_n(x) - x\big|.$$

□

On en déduit un intervalle de confiance, uniforme en $x \in \mathbb{R}$ (une régions de confiance) asymptotique. Pour tout $\alpha \in (0,1)$, désignons pas $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de $\mathbb{B}$, de sorte que

$$\mathbb{P}\big[\mathbb{B} \le q_{1-\alpha}\big] = 1 - \alpha.$$

Proposition 3.5. *La région*

$$\big\{\mathcal{J}_{n,\alpha}(x), x \in \mathbb{R}\,\big\} = \left\{\Big[\widehat{F}_n(x) \pm \frac{q_{1-\alpha}}{\sqrt{n}}\Big], x \in \mathbb{R}\right\}$$

est une région de confiance asymptotique :

$$\mathbb{P}\Big[\forall x \in \mathbb{R},\ F(x) \in \mathcal{J}_{n,\alpha}(x)\Big] \to 1 - \alpha.$$

Démonstration. On applique le Théorème 3.3 :

$$\begin{aligned}\mathbb{P}\Big[\forall x \in \mathbb{R},\ F(x) \in \mathcal{J}_{n,\alpha}(x)\Big] &= \mathbb{P}\Big[\sup_{x\in\mathbb{R}} \sqrt{n}\big|\widehat{F}_n(x) - F(x)\big| \le q_{1-\alpha}\Big] \\ &\to \mathbb{P}\big[\mathbb{B} \le q_{1-\alpha}\big] = 1 - \alpha.\end{aligned}$$

□

Remarque 3.9. Bien entendu, on a toujours $0 \leq F(x) \leq 1$, ce qui n'est pas forcément le cas de $\widehat{F}_n(x) \pm q_{1-\alpha}/\sqrt{n}$. On peut « réduire » la région $\big\{\mathcal{J}_{n,\alpha}(x), x \in \mathbb{R}\big\}$ en remplaçant $\mathcal{J}_{n,\alpha}(x)$ par

$$\overline{\mathcal{J}}_{n,\alpha}(x) := \mathcal{J}_{n,\alpha}(x) \cap [0,1]$$

sans modifier la propriété de couverture asymptotique.

3.3.3 Précision uniforme non-asymptotique*

De la même manière que l'inégalité de Hoeffding du Théorème 3.1 nous a fourni une précision ponctuelle non-asymptotique, on a le résultat suivant

Théorème 3.4 (Inégalité de Dvoretzky-Kiefer-Wolfowitz)**.** *Si la fonction de répartition F est continue, pour $n \geq 1$ et $t > 0$, on a*

$$\mathbb{P}\big[\sup_x \big|\widehat{F}_n(x) - F(x)\big| \geq t\big] \leq 2\exp(-2nt^2).$$

La preuve utilise des résultats fins sur les processus empiriques et nous l'admettons. On en déduit, pour $\alpha \in (0,1)$, une région de confiance non-asymptotique uniforme

$$\big\{\mathcal{I}_{n,\alpha}(x), x \in \mathbb{R}\big\} = \big\{\big[\widehat{F}_n(x) \pm \sqrt{\tfrac{1}{2n}\log\tfrac{1}{\alpha}}\big], x \in \mathbb{R}\big\}$$

qui vérifie, pour tout $n \geq 1$

$$\mathbb{P}\Big[\forall x \in \mathbb{R},\ F(x) \in \mathcal{I}_{n,\alpha}(x)\Big] \geq 1 - \alpha.$$

Remarque 3.10. De le même manière que dans le cadre asymptotique, on peut modifier $\mathcal{I}_{n,\alpha}(x)$ en considérant $\mathcal{I}_{n,\alpha}(x) \cap [0,1]$.

3.3.4 Test d'adéquation à une distribution donnée*

Soit F_0 une distribution donnée. On souhaite maintenant décider, en vue des observations $X_1, \ldots, X_n$ distribuées selon la loi F si $F = F_0$ contre $F \neq F_0$ « globalement » c'est-à-dire tester l'hypothèse nulle

$$H_0\ :\ \forall x \in \mathbb{R},\ F(x) = F_0(x)$$

contre l'alternative

$$H_1\ :\ \exists\, x \in \mathbb{R},\ F(x) \neq F_0(x).$$

On doit modifier par rapport à la Section 3.2.5 la traduction de l'hypothèse $\mathcal{F}_0 \subset \mathfrak{F}$. On pose

$$\mathfrak{F}_0 = \big\{F \in \mathfrak{F},\ \forall x \in \mathbb{R}, F(x) = F_0(x)\big\} = \{F_0\}$$

et on traduit l'hypothèse H_0 par la propriété $F \in \mathfrak{F}_0$.

De la même manière que dans la Section 3.2.5, on peut construire un test de l'hypothèse H_0 contre H_1 à l'aide des régions de confiance $\{\mathcal{I}_{n,\alpha}(x), x \in \mathbb{R}\}$, ou $\{\mathcal{J}_{n,\alpha}(x), x \in \mathbb{R}\}$.

Pour simplifier, nous énonçons un résultat asymptotique.

Proposition 3.6 (Test de Kolmogorov-Smirnov)**.** *Pour tout $\alpha \in (0,1)$, le test simple de l'hypothèse $H_0 : F \in \mathfrak{F}_0$ contre l'alternative $H_1 : F \in \mathfrak{F} \setminus \mathfrak{F}_0$, défini par la zone de rejet*

$$\mathcal{R}_{n,\alpha} = \Big\{ \exists\, x \in \mathbb{R},\ \ F_0(x) \notin \mathcal{J}_{n,\alpha}(x) \Big\}$$

est asymptotiquement de niveau α.

De plus, pour tout point de l'alternative $F \in \mathfrak{F} \setminus \{F_0\}$, on a

$$\mathbb{P}_F \left[(X_1, \ldots, X_n) \notin \mathcal{R}_{n,\alpha} \right] \to 0.$$

Démonstration. Sous l'hypothèse, on a $F = F_0$ et

$$\mathbb{P}_{F_0} \left[(X_1, \ldots, X_n) \notin \mathcal{R} \right] = 1 - \mathbb{P}_{F_0} \left[\forall x \in \mathbb{R},\ F_0(x) \in \mathcal{J}_{n,\alpha}(x) \right] \to \alpha$$

lorsque $n \to \infty$ par la Proposition 3.5. Donc le test de Kolmogorov-Smirnov est asymptotiquement de niveau α. Pour tout point $F \in \mathfrak{F} \setminus \{F_0\}$ de l'alternative, il existe un point $x_0 \in \mathbb{R}$ pour lequel $F(x_0) \neq F_0(x_0)$. On reprend alors point par point la fin de la démonstration de la Proposition 3.4. □

3.4 Estimation de fonctionnelles

Dans les Sections 3.2 et 3.3 nous avons rencontré deux situations opposées :

1. L'estimation « locale » de F en un point x_0. Nous nous sommes intéressé à la fonctionnelle linéaire
$$T_{x_0}(F) = F(x_0).$$
2. L'estimation « globale » de F, c'est-à-dire l'estimation simultané des fonctionelles
$$\{T_x(F) = F(x),\ x \in \mathbb{R}\}.$$

Plus généralement, on peut s'intéresser à l'estimation ou la décision relative à des fonctionnelles plus générales. Par exemple

1. Une fonctionnelle linéaire, de la forme
$$T(F) = \int_{\mathbb{R}} g(x) dF(x), \tag{3.7}$$

avec g connue (choisie par le statisticien). L'exemple prototype étant le moment d'ordre 1, pour le choix $g(x) = x$

$$m(F) = \int_{\mathbb{R}} x\, dF(x).$$

2. Une combinaison de fonctionelles linéaires : la variance

$$\sigma^2(F) = \int_{\mathbb{R}} \big(x - m(F)\big)^2 dF(x),$$

le coefficient d'asymétrie

$$\alpha(F) = \frac{\int_{\mathbb{R}} \big(x - m(F)^3\big) dF(x)}{\sigma^2(F)^{3/2}},$$

le coefficient d'applatissement de F,

$$\kappa(F) = \frac{\int_{\mathbb{R}} \big(x - m(F)\big)^4 dF(x)}{\sigma^2(F)^2}$$

parmi bien d'autres exemples.

3. Une fonctionelle non-linéaire, comme le quantile d'ordre $\alpha \in (0,1)$:

$$T(F) = q_\alpha(F) = \tfrac{1}{2}\big(\inf\{t,\ F(t) > \alpha\} + \sup\{t,\ F(t) < \alpha\}\big).$$

3.4.1 Le cas régulier : méthode de substitution

Un estimateur naturel de $T(F)$ est l'estimateur par substitution, où l'on remplace formellement F par sa répartition empirique $\widehat{F}_n(\bullet)$.

Définition 3.8. *L'estimateur par substitution de $T(F)$*

$$\widehat{T}_n = \widehat{T}_n(X_1, \ldots, X_n) = T(\widehat{F}_n)$$

est obtenu en remplaçant F par sa fonction de répartition empirique $\widehat{F}_n$.

Convergence dans le cas régulier

On a vu dans la Section 3.3 que les fonctions $\widehat{F}_n(\bullet)$ et $F(\bullet)$ sont proches lorsque n est grand. On imagine alors que $T(\widehat{F}_n)$ est proche de $T(F)$ dès lors que la fonction $F \rightsquigarrow T(F)$ est régulière.

Proposition 3.7. *Si la fonctionnelle $T(F)$ admet la représentation*

$$T(F) = h\left(\int_{\mathbb{R}} g(x)dF(x)\right) \tag{3.8}$$

où $\int_{\mathbb{R}} |g(x)|dF(x) < +\infty$ et $h : \mathbb{R} \to \mathbb{R}$ continue, alors

$$T(\widehat{F}_n) \xrightarrow{\text{p.s.}} T(F). \tag{3.9}$$

Démonstration. Remarquons que $T(\widehat{F}_n) = h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$. On a

$$\frac{1}{n}\sum_{i=1}^n g(X_i) \xrightarrow{\text{p.s.}} \mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}} g(x)dF(x)$$

par la loi forte des grands nombres. La convergence reste vraie en composant par h qui est continue. □

Exemple 3.1. La variance $\sigma^2(F)$ de la distribution F s'écrit

$$\begin{aligned}\sigma^2(F) &= \int_{\mathbb{R}} \big(x - m(F)\big)^2 dF(x) \\ &= \int_{\mathbb{R}} x^2 dF(x) - \Big(\int_{\mathbb{R}} x dF(x)\Big)^2 \\ &= h_1\left(\int_{\mathbb{R}} g_1(x)dF(x)\right) + h_2\left(\int_{\mathbb{R}} g_2(x)dF(x)\right),\end{aligned}$$

avec $h_1(x) = x$, $h_2(x) = x^2$, $g_1(x) = x^2$, $g_2(x) = x$. L'estimateur pas substitution associé s'écrit

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n{}^2 = \frac{1}{n}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2.$$

La convergence $\widehat{\sigma}_n^2 \xrightarrow{\text{p.s.}} \sigma^2(F)$ découle de la Proposition 3.7 appliquée à chacun des termes $\frac{1}{n}\sum_{i=1}^n X_i^2$ et $\overline{X}_n{}^2$ respectivement. On peut faire des calculs analogues pour le coefficients d'asymétrie $\alpha(F)$ et pour le coefficient d'aplatissement $\kappa(F)$.

Remarque 3.11. Plus généralement, si l'on munit $\mathfrak{F}$ de la métrique de la convergence uniforme, le Théorème 3.2 (Glivenko-Cantelli) assure que la convergence (3.9) aura lieu si l'application $T \rightsquigarrow T(F)$ est continue.

Vitesse de convergence dans le cas régulier

Pour les fonctionnelles de type (3.8), on a une vitesse de convergence :

Proposition 3.8. *Dans la situation de la Proposition 3.7, si h est continûment dérivable et si $\mathbb{E}\big[g(X)^2\big] = \int_{\mathbb{R}} g(x)^2 dF(x) < +\infty$, alors*

$$\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big) \xrightarrow{d} \mathcal{N}\big(0, v(F)\big),$$

où

$$v(F) = h'\big(\mathbb{E}\big[g(X)\big]\big)^2 \mathrm{Var}\big[g(X)\big].$$

Démonstration. Par le théorème central limite,

$$\begin{aligned}\sqrt{n}\left(\int_{\mathbb{R}} g(x)d\widehat{F}_n(x) - \int_{\mathbb{R}} g(x)dF(x)\right) &= \sqrt{n}\Big(\tfrac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}\big[g(X)\big]\Big) \\ &\xrightarrow{d} \mathcal{N}\big(0, \mathrm{Var}\big[g(X)\big]\big).\end{aligned}$$

On applique alors la Proposition 1.10 du Chapitre 1 (méthode delta) :

$$\sqrt{n}\Big(h\big(\tfrac{1}{n}\sum_{i=1}^n g(X_i)\big) - h\big(\mathbb{E}\big[g(X)\big]\big)\Big) \xrightarrow{d} \mathcal{N}\Big(0, h'\big(\mathbb{E}\big[g(X)\big]\big)^2 \mathrm{Var}\big[g(X)\big]\Big).$$

C'est précisément le résultat recherché, puisque $h\big(\mathbb{E}\big[g(X)\big]\big) = T(F)$. □

Exemple 3.2. Etudions le comportement de l'estimateur par substitution de

$$T(F) = \frac{1}{\mathbb{E}\big[X^4\big]} = \frac{1}{\int_{\mathbb{R}} x^4 dF(x)}$$

sous l'hypothèse que $0 < \int_{\mathbb{R}} x^8 dF(x) < +\infty$. On a

$$T(\widehat{F}_n) = \frac{1}{\frac{1}{n}\sum_{i=1}^4 X_i^4}$$

(en convenant par exemple $1/0 = 0$). On applique la Proposition 3.8, avec $g(x) = x^4$ et $h(x) = x^{-1}$. (Il y a cependant une difficulté : en $x = 0$ la fonction h ne vérifie pas les hypothèses de la Proposition [12] 3.8 puisque h a une singularité en 0. En appliquant tout de même formellement de résultat de la proposition, on a

$$\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big) \xrightarrow{d} \mathcal{N}\big(0, v(F)\big),$$

avec

$$v(F) = h'\big(\mathbb{E}\big[g(X)\big]\big)^2\big(\mathbb{E}\big[g(X)^2\big] - \mathbb{E}\big[g(X)\big]^2\big) = \frac{\mu_8}{\mu_4^2} - 1$$

où $\mu_i = \mathbb{E}[X^i] = \int_{\mathbb{R}} x^i dF(x)$. On peut pousser un peu plus loin l'étude et déduire de ce résultat un intervalle de confiance asymptotique pour $T(F) = \mu_4^{-1}$ comme dans la Section 3.2.3. C'est l'objet de l'Exercice 3.3.

[12] Il s'agit en fait d'un faux problème : on a $\mathbb{E}\big[X^4\big] = \int_{\mathbb{R}} x^4 dF(x) > 0$ puisque sinon, $X = 0$ presque-sûrement et donc $F = 1_{\mathbb{R}_+}(x)$ ce qui contredirait l'hypothèse $\int_{\mathbb{R}} x^8 dF(x) > 0$. Ceci entraîne que X est « éloigné en moyenne » de la singularité 0. On pourra alors montrer en exercice que la convergence en loi voulue a bien lieu.

La Proposition 3.8 ne donne qu'un résultat en dimension 1 : elle ne permet même pas de traiter immédiatement la vitesse de convergence dans l'Exemple 3.1, et une version multidimensionnelle de la « méthode delta » s'avère nécessaire dans la cas général.

Considérons une fonctionnelle de la forme

$$T(F) = h\Big(\int_{\mathbb{R}} g_1(x)dF(x), \ldots, \int_{\mathbb{R}} g_k(x)dF(x)\Big), \tag{3.10}$$

où $h : \mathbb{R}^k \to \mathbb{R}$ est une fonction différentiable, de gradient

$$J_h(\boldsymbol{x}) = \nabla h(\boldsymbol{x}) = \big(\partial_1 h(\boldsymbol{x}), \ldots, \partial_k h(\boldsymbol{x})\big), \quad \boldsymbol{x} \in \mathbb{R}^k.$$

En appliquant la Proposition 1.11, on a le résultat suivant.

Corollaire 3.2. *Si la fonctionnelle $T(F)$ admet la représentation* (3.10) *avec une fonction h continûment différentiable, et si $\int_{\mathbb{R}} g_i(x)^2 dF(x) < +\infty$ pour tout $i = 1, \ldots, k$, alors*

$$\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big) \xrightarrow{d} \mathcal{N}\big(0, v(F)\big),$$

avec

$$v(F) = J_h(\mathbf{g})\, \Sigma_{\mathbf{g}}\, J_h(\mathbf{g})^T,$$

où

$$\mathbf{g} = \big(\mathbb{E}\left[g_1(X)\right], \ldots, \mathbb{E}\left[g_k(X)\right]\big)$$

et $\Sigma_{\mathbf{g}}$ est la matrice de variance-covariance des $g_i(X)$:

$$\big(\Sigma_{\mathbf{g}}\big)_{ij} = \mathbb{E}\left[\big(g_i(X) - \mathbb{E}[g_i(X)]\big)\big(g_j(X) - \mathbb{E}[g_j(X)]\big)\right], \;\; 1 \le i, j \le k.$$

Exemple 3.3. Reprenons le problème du calcul de la loi limite de la variance empirique de l'exemple 3.1. On a

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \big(\overline{X}_n\big)^2.$$

On applique le Corollaire 3.2 avec $h(x_1, x_2) = x_1 - x_2^2$, $g_1(x) = x^2$ et $g_2(x) = x$. On a

$$\nabla h(x_1, x_2) = (1,\ -2x_2) \quad \text{et} \quad \mathbf{g} = (\mathbb{E}[X^2],\ \mathbb{E}[X]).$$

Notons $\mu_i = \mathbb{E}[X^i]$. Un calcule simple montre que

$$\Sigma_{\mathbf{g}} = \begin{pmatrix} \mu_4 - \mu_2^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_2 - \mu_1^2 \end{pmatrix}$$

Alors

$$\sqrt{n}\big(\widehat{\sigma}_n^2 - \sigma\big) \xrightarrow{d} \mathcal{N}\big(0, v(F)\big),$$

avec

$$v(F) = (1,\ -2\mu_1) \begin{pmatrix} \mu_4 - \mu_2^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_2 - \mu_1^2 \end{pmatrix} (1,\ -2\mu_1)^T.$$

On trouve

$$v^2 = \mu_4 - \mu_2^2 - 4\mu_1(\mu_3 + \mu_1^3 - 2\mu_1\mu_2).$$

Dans le cas précis de la variance empirique, on aurait pu aussi retrouver directement ce résultat par une autre méthode, voir l'Exercice 3.2.

Avec la même technique, on peut exhiber les lois limites du coefficient d'asymétrie empirique et du coefficient d'aplatissement empirique.

3.4.2 Le cas non-régulier*

Les fonctionnelles régulières de type (3.8) sont insuffisantes pour les applications : elles ne recouvrent pas par exemple le cas très utile de l'estimation des quantiles d'une distribution inconnue.

Plus généralement, supposons que l'on dispose de l'information supplémentaire suivante sur le modèle statistique :

$$F \in \mathfrak{F}^{\mathrm{ac}} \subset \mathfrak{F},$$

où $\mathfrak{F}^{\mathrm{ac}}$ désigne l'ensemble des distributions absolument continues, c'est-à-dire pour lesquelles la fonction de répartition F est dérivable presque-partout. Alors, par exemple, en notant $f(x) = F'(x)$, la fonctionnelle

$$T(F) = \int_{\mathbb{R}} F'(x)^2 dx = \int_{\mathbb{R}} f(x)^2 dx$$

n'est pas régulière. On ne peut pas former d'estimateur par substitution en « dérivant » $\widehat{F}_n(\bullet)$ qui est constante par morceaux. Plus généralement, dans le cas où le modèle statistique a pour ensemble de paramètres $\mathfrak{F}^{\mathrm{ac}}$, on peut s'intéresser à la constuction d'un estimateur $\widehat{f}_n(\bullet)$ qui soit une bonne approximation de la densité $f(\bullet)$ de F.

Dans le reste de cette section, nous étudions deux cas particuliers : l'estimation des quantiles, et le lissage de la distribution empirique.

Estimation des quantiles

On considère la statistique d'ordre associée à l'échantillon $(X_1, \ldots, X_n)$, c'est-à-dire le vecteur $(X_{(1)}, \ldots, X_{(n)})$ obtenu par la permutation (aléatoire) qui fournit le réarrangement croissant des données

$$X_{(1)} \leq \cdots \leq X_{(i)} \leq \cdots \leq X_{(n)}.$$

Cette permutation n'est pas nécessairement unique (dans le cas discret, certaines valeurs des observations peuvent coïncider). Pour estimer le quantile[13] d'ordre p de la loi F, c'est-à-dire

$$T(F) = \tfrac{1}{2}\big(\inf\{q,\, F(q) > p\} + \sup\{q,\, F(q) < p\}\big)$$

[13] voir la Section 1.2.3 du Chapitre 1.

on peut choisir l'estimateur par substitution

$$\widehat{q}_{n,p} = T(\widehat{F}_n) = \tfrac{1}{2}\big(\inf\{q,\, \widehat{F}_n(q) > p\} + \sup\{q,\, \widehat{F}_n(q) < p\}\big)$$

appelé quantile empirique d'ordre p. La difficulté de cette approche réside dans le fait que $x \rightsquigarrow \widehat{F}_n(x)$ est constante par morceaux, donc, pour $p \in [0,1]$ donné, l'équation

$$\widehat{F}_n(q) = p$$

admet une infinité de solution ou n'en admet aucune. On a toujours

$$\widehat{F}_n(\widehat{q}_{n,p}) = p$$

et on a, plus précisément

$$\widehat{q}_{n,p} = \begin{cases} X_{(k)} & \text{si} \quad p \in \big((k-1)/n, k/n\big) \\ \frac{1}{2}\big(X_{(k)} + X_{(k+1)}\big) & \text{si} \quad p = k/n \end{cases}$$

pour $k = 1, \ldots, n$.

Lissage de la distribution empirique*

Etant donné l'observation $X_1, \ldots, X_n$, la fonction aléatoire

$$x \rightsquigarrow \widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i(\omega) \le x\}}$$

est constante par morceaux. On insiste ici sur l'aléa ω, pour marquer le fait que $\widehat{F}_n(\bullet)$ dépend d'une réalisation $\big(X_1(\omega), \ldots, X_n(\omega)\big)$ du vecteur aléatoire $\big(X_1, \ldots, X_n\big)$. Si on prend formellement sa dérivée (au sens des distributions), on obtient

$$\widehat{F}'_n(dx) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}(dx) \tag{3.11}$$

où $\delta_a(dx)$ est la mesure de Dirac au point a. On obtient ainsi une mesure de probabilité[14], qui assigne à chaque point $X_i(\omega)$ la masse $1/n$.

Définition 3.9. *Etant donnée une réalisation $\big(X_1(\omega), \ldots, X_n(\omega)\big)$ du vecteur aléatoire $\big(X_1, \ldots, X_n\big)$, on appelle distribution empirique la mesure de probabilité uniforme sur l'ensemble $\{X_1(\omega), \ldots, X_n(\omega)\}$ définie par* (3.11).

Remarquons qu'en posant formellement

$$d\widehat{F}_n(x) = \widehat{F}'_n(dx),$$

les notations sont cohérentes avec les calculs : pour toute fonction test φ, on a

$$\int_{\mathbb{R}} \varphi(x) d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi\big(X_i(\omega)\big) = \int_{\mathbb{R}} \varphi(x) \tfrac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}(dx).$$

[14] qui dépend de ω, il s'agit donc d'une distribution aléatoire

Estimateur à fenêtre mobile et à noyau*

La densité f est la dérivée de la fonction de répartition $x \rightsquigarrow F(x)$. Ecrivons l'approximation

$$f(x) = F'(x) \approx \frac{1}{h}\big(F(x+h/2) - F(x-h/2)\big)$$

lorsque h est petit. On approche le membre de droite par substitution. Ceci fournit l'estimateur

$$\widehat{f}_n(x) = \frac{1}{h}\big(\widehat{F}_n(x+h/2) - \widehat{F}_n(x-h/2)\big),$$

appelé estimateur par fenêtre mobile.

Posons $U_x^h = [x-h/2, x+h/2)$. Alors $\widehat{f}_n(x)$ compte le nombre d'observations X_i qui « tombent » dans la « fenêtre » U_x^h normalisé par n, puis on fait glisser la fenêtre U_x^h avec x :

$$\begin{aligned}\frac{1}{h}\big(\widehat{F}_n(x+h/2) - \widehat{F}_n(x-h/2)\big) &= \frac{1}{nh}\sum_{i=1}^{n} 1_{\{X_i \in U_x^h\}} \\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right),\end{aligned}$$

où $K(x) = 1_{-1/2<x\leq 1/2}$. La fonction aléatoire $x \rightsquigarrow \widehat{f}_n(x)$ est elle-même une densité de probabilité, constante par morceaux.

Un version plus régulière de l'estimateur à fenêtre mobile consiste à remplacer la fonction K par une fonction plus régulière $K^{(r)}$, vérifiant $\int_{\mathbb{R}} K^{(r)}(x)dx = 1$. On utilise souvent le noyau gaussien

$$K^{(r)}(x) = (2\pi)^{-1/2}\exp(-x^2/2).$$

L'estimateur à noyau

$$\widehat{f}_n^{(r)}(x) = \frac{1}{nh}\sum_{i=1}^{n} K^{(r)}\left(\frac{x-X_i}{h}\right)$$

est donc la moyenne arithmétique de n « fonctions cloches »

$$\frac{1}{h}K^{(r)}\left(\frac{\bullet - X_i}{h}\right),$$

chaque « cloche » étant une densité de probabilité centrée en X_i et d'échelle h. La fonction aléatoire $x \rightsquigarrow \widehat{f}_n^{(r)}(x)$ est une densité de probabilité : elle est positive, et

$$\int_{\mathbb{R}} \widehat{f}_n^{(r)}(x)dx = \int_{\mathbb{R}} K(x)dx = 1.$$

L'étude des estimateurs à noyaux pour l'estimation non-paramétrique de la densité est une théorie à part entière qui dépasse le cadre de ce cours.

3.5 Exercices

Exercice 3.1. Soit $\Phi(x) = (2\pi)^{-1/2}\int_{-\infty}^{x} e^{-t^2/2}dt$ la fonction de répartition de la loi gaussienne standard.

- Montrer que $1-\Phi(x) \leq \frac{1}{2}e^{-x^2/2}$ et en déduire que pour $\alpha \in (0,1)$,

$$\alpha \leq \exp(-\frac{1}{2}\Phi^{-1}(1-\alpha/2)^2).$$

- Montrer que $1-\Phi(x) = \frac{x}{\sqrt{2\pi}}e^{-x^2/2} - x^2[1-\Phi(x)]$. En déduire

$$1-\Phi(x) \geq \frac{e^{-x^2/2}}{2x\sqrt{2\pi}}.$$

  (On pourra utiliser l'inégalité t $x/(1+x^2) \geq 1/2x$ si $x \geq 1$.)
- En déduire

$$\sqrt{2\log\frac{1}{\alpha r(\alpha)}} \leq \Phi^{-1}(1-\alpha/2),$$

  où l'on a posé $r(\alpha) := 2\sqrt{\pi \log\frac{1}{\alpha}}$.

Exercice 3.2. On a étudié le comportement asymptotique de la variance empirique par la méthode « delta » dans l'exemple 3.3. On peut retrouver ce résultat de manière plus directe. On écrit

$$\sqrt{n}\big(\tfrac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)-\sigma^2\big) = \sqrt{n}\big(\tfrac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\sigma^2\big) - \sqrt{n}(\overline{X}_n-\mu)^2.$$

Montrer que le second terme converge vers 0 en probabilité. Montrer que le premier terme est asymptotiquement normal via le théorème central-limite. Conclure via la Proposition 1.8 (Slutsky).

Exercice 3.3. On cherche un intervalle de confiance asymptotique pour la fonctionnelle

$$T(F) = \frac{1}{\mathbb{E}\left[X^4\right]} = \frac{1}{\int_{\mathbb{R}} x^4 dF(x)}$$

sous l'hypothèse que $0 < \int_{\mathbb{R}} x^8 dF(x) < +\infty$. On a vu dans l'Exemple 3.2 la Section 3.4.1 la convergence

$$\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big) \xrightarrow{d} \mathcal{N}\big(0, v(F)\big),$$

avec $v(F) = \mu_8/\mu_4^2 - 1$. Montrer que $v(\widehat{F}_n) \xrightarrow{\mathbb{P}} v(F)$ et en déduire un intervalle de confiance asymptotique pour $T(F)$ à l'aide de la Proposition 1.8 (Slutsky).

Exercice 3.4. Soient $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes, de même densité f. On note $X_{(1)}, \ldots, X_{(n)}$ la statistique d'ordre associée (voir Section 3.4.2).

– Montrer que la densité de $(X_{(1)}, \ldots, X_{(n)})$ est donnée par

$$f_{(X_{(1)},\ldots,X_{(n)})} = n! \prod_{i=1}^{n} f(x_i) 1_{\{x_1<x_2<\ldots<x_n\}}.$$

– Si F désigne la fonction de répartition des X_i, montrer que $X_{(k)}$ a pour densité

$$f_{X_{(k)}}(x) = k\, C_n^k\, f(x)\big(1-F(x)\big)^{n-k} F(x)^{k-1}.$$

Exercice 3.5 (Un test asymptotique de gaussianité)**.** Soient $X_1, \ldots, X_n$ un n-échantillon de loi inconnue F_0 ayant au moins un moment d'ordre 4 et de moyenne nulle et de variance non-nulle.

– On pose, pour $k = 1, \ldots, 4$

$$T_n^{(k)} = \frac{\frac{1}{n}\sum_{i=1}^n X_i^k}{\big(\frac{1}{n}\sum_{i=1}^n X_i^2\big)^{k/2}}.$$

Montrer que

$$\frac{n}{6}\big(T_n^{(3)}\big)^2 + \frac{n}{24}\big(T_n^{(4)} - 3\big)^2 \xrightarrow{d} \chi^2(2),$$

où $\chi^2(2)$ désigne la loi du χ^2 à 2 degrés de liberté.

– En déduire un test de l'hypothèse nulle $H_0 : F = \Phi$ contre l'alternative $H_1 : F \neq \Phi$ où $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-t^2/2} dt$ est la fonction de répartition de loi loi normale standard.

– Le test est-il consistant ?

Chapitre 4

Méthodes d'estimation pour le modèle de densité

On se place dans le modèle d'échantillonnage. L'hypothèse supplémentaire par rapport au Chapitre 3 est que la famille de probabilités associée à l'expérience statistique est « paramétrique » : on peut la représenter à l'aide d'un sous-ensemble d'un espace de dimension finie.

4.1 Introduction

4.1.1 Notations et hypothèses

Situation

On observe un n-échantillon

$$X_1, \ldots, X_n$$

d'une loi inconnue sur $\mathbb{R}$, que l'on notera aussi sous forme d'un vecteur colonne

$$\big(X_1, \ldots, X_n\big)^T,$$

où les X_i sont des variables indépendantes et identiquement distribuées, et on suppose que leur loi commune appartient à une famille paramétrique de lois donnée

$$\big\{\, \mathbb{P}_\vartheta, \vartheta \in \Theta\big\}, \;\; \Theta \subset \mathbb{R}^d,$$

où ϑ est un paramètre de dimension d. L'expérience statistique sous-jacente au sens de la Definition 2.3 du Chapitre 2 s'écrit

$$\mathcal{E}^n = \big(\, \mathbb{R}^n, \mathcal{B}^n, \{\mathbb{P}^n_\vartheta, \vartheta \in \Theta\}\big)$$

où $\mathbb{P}_\vartheta^n$ est la loi de n variables aléatoires indépendantes de loi $\mathbb{P}_\vartheta$. On écrit indifféremment $\mathbb{P}_\vartheta$ ou $\mathbb{P}_\vartheta^n$ voire $\mathbb{P}$ lorsqu'il n'y a pas de confusion possible. On note aussi $\mathcal{E} = \mathcal{E}^1$, l'expérience associée à une seule observation.

Dans ce contexte, on cherche à construire des estimateurs $\widehat{\vartheta}_n$ de ϑ, ou plutôt des suites d'estimateurs, variant avec n. Un estimateur – *cf.* la Définition 3.1 – est une quantité mesurable par rapport aux observations :

$$\widehat{\vartheta}_n = \widehat{\vartheta}_n(X_, \dots, X_n)$$

à valeurs dans $\mathbb{R}^d$ (idéalement, à valeurs dans Θ). Evidemment, un estimateur raisonnable $\widehat{\vartheta}_n$ « approche » ϑ d'autant mieux que le nombre d'observations n est grand. Nous allons développer des méthodes systématiques de construction d'estimateurs « raisonnables », en faisant des hypothèses adéquates sur la famille $\big\{ \mathbb{P}_\vartheta, \vartheta \in \Theta \big\}$.

Identifiabilité

Nous supposons toujours que l'expérience est bien paramétrée, au sens où la fonction $\vartheta \in \Theta \rightsquigarrow \mathbb{P}_\vartheta$ est injective, ce qui était déjà implicite dans nos notations : deux valeurs différentes $\vartheta_1 \neq \vartheta_2$ donnent lieu à deux mesures de probabilités $\mathbb{P}_{\vartheta_1} \neq \mathbb{P}_{\vartheta_2}$ différentes.

Une expérience statistique $\mathcal{E}^n$ engendrée par l'observation d'un n-échantillon s'écrit $\mathcal{E}^n = \mathcal{E} \times \cdots \times \mathcal{E}$ (n fois), où $\mathcal{E}$ est l'expérience statistique associée à une observation ($\mathcal{E} = \mathcal{E}^1$). Alors $\mathcal{E}^n$ est identifiable si et seulement si $\mathcal{E}$ l'est.

Voici un exemple de mauvaise paramétrisation donnant lieu à un modèle non-identifiable : $\mathbb{P}_\vartheta$ est la loi sur $\mathbb{R}$ de densité par rapport à la mesure de Lebesgue

$$f(\vartheta, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\vartheta^2)}, \quad \vartheta \in \Theta = \mathbb{R}\,.$$

La donnée de $f(\vartheta, \bullet)$ ne permet par de distinguer ϑ et $-\vartheta$. Par contre, la même expérience associée à l'ensemble des paramètres $\widetilde{\vartheta} = \mathbb{R}_+$ devient identifiable.

Domination

Nous faisons une hypothèse essentielle de domination, qui permet, en un certain sens, de réduire la complexité de l'étude de $\mathcal{E}^n$ à celle d'une fonction de plusieurs variables.

Hypothèse 4.1. *L'expérience $\mathcal{E}$ est dominée : ll existe une mesure σ-finie μ sur $\mathbb{R}$ telle que, pour tout $\vartheta \in \Theta$, μ domine $\mathbb{P}_\vartheta$. On note*

$$f(\vartheta, x) = \frac{d\,\mathbb{P}_\vartheta}{d\mu}(x), \quad x \in \mathbb{R}$$

la densité de $\mathbb{P}_\vartheta$ par rapport à μ.

Remarque 4.1. Pour un n-échantillon, $\mathcal{E}^n$ est dominée si et seulement si $\mathcal{E}$ l'est. L'expérience statistique $\mathcal{E}^n$ est dominéee par la mesure produit $\mu^n = \mu \otimes \ldots \otimes \mu$ (n fois) et

$$\frac{d\,\mathbb{P}_\vartheta^n}{d\mu^n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(\vartheta, x_i), \quad x_1, \ldots, x_n \in \mathbb{R}\,.$$

Remarque 4.2. Se donner une expérience statistique satisfaisant l'Hypothèse 4.1 revient à spécifier une application $f : \Theta \times \mathbb{R} \to \mathbb{R}$. Nous verrons dans ce chapitre ainsi qu'au Chapitre 6 comment l'estimation de ϑ est intimement liée à la régularité de la fonction $(\vartheta, x) \rightsquigarrow f(\vartheta, x)$.

Dans presque toutes les situations que nous considérerons, la mesure μ est la mesure de Lebesgue sur $\mathbb{R}$ lorsque la loi des observations est absolument continue ou bien la mesure de comptage sur l'ensemble des valeurs possibles des observations lorsque la loi des observations est discrète.

Exemple 4.1.

1. Si l'expérience statistique $\mathcal{E}$ est engendrée par l'observation d'une variable exponentielle de paramètre $\vartheta, \vartheta > 0$, alors $\mathbb{P}_\vartheta(dx)$ est la loi exponentielle de paramètre ϑ et $\Theta = \mathbb{R}_+ \setminus \{0\}$. Une mesure dominante est la mesure de Lebesgue $\mu(dx) = dx$ et on a
$$\mathbb{P}_\vartheta(dx) = f(\vartheta, x)dx = \vartheta \exp(-\vartheta x) 1_{\{x \geq 0\}} dx.$$

2. Si $\mathcal{E}$ est engendrée par l'observation d'une variable de Poisson de paramètre $\vartheta > 0$, alors $\mathbb{P}_\vartheta(dx)$ est la loi de Poisson de paramètre ϑ et $\Theta = \mathbb{R}_+ \setminus \{0\}$. Dans ce cas, on peut prendre pour μ la mesure de comptage sur $\mathbb{N}$ et on a
$$\mathbb{P}_\vartheta(dx) = f(\vartheta, x)\mu(dx) = \exp(-\vartheta)\frac{\vartheta^x}{x!}\mu(dx),$$
et on a aussi
$$f(\vartheta, x) = \mathbb{P}_\vartheta\big[X = x\big].$$

3. Si $\mathcal{E}$ est engendrée par l'observation d'une variable gaussienne, de moyenne μ et de variance σ^2, alors $\vartheta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$ et $\mathbb{P}_\vartheta(dx)$ est la loi $\mathcal{N}(\mu, \sigma^2)$. Dans ce cas, on peut prendre $\mu(dx) = dx$ et on a
$$f(\vartheta, x) = (2\pi\sigma^2)^{-1/2} \exp\big(-\tfrac{1}{2\sigma^2}(x-\mu)^2\big).$$
Attention : dans certaines situations statistique, on suppose que l'on connaît l'un des valeurs μ ou σ^2. Dans ce cas, on doit changer de paramètre et d'ensemble de paramètres, même si, bien-sûr, la loi des observations reste la même. Par exemple, si l'on connaît σ^2, alors on prend $\vartheta = \mu$, $\Theta = \mathbb{R}$ et on écrit plutôt
$$f_{\sigma^2}(\vartheta, x) = (2\pi\sigma^2)^{-1/2} \exp\big(-\tfrac{1}{2\sigma^2}(x-\mu)^2\big).$$

Exemple 4.2. Un exemple où il n'existe pas de mesure dominante est la famille paramétrique $\{\mathbb{P}_\vartheta : \delta_\vartheta, \vartheta \in \mathbb{R}\}$, où δ_ϑ est la mesure de Dirac au point ϑ. Cet exemple[1] correspond à l'expérience parfaite ou une seule observation permet de connaître sans erreur le paramètre ϑ.

Calcul de lois

On note $\mathbb{P}_\vartheta^n$ (ou $\mathbb{P}_\vartheta$ lorsqu'il n'y a pas de confusion) la loi des observations, et $\mathbb{E}_\vartheta^n\left[\bullet\right]$ (ou $\mathbb{E}_\vartheta\left[\bullet\right]$) l'espérance associée. Si $\widehat{\vartheta}_n$ est un estimateur de ϑ et g une fonction test, alors

$$\begin{aligned}\mathbb{E}_\vartheta\left[g(\widehat{\vartheta}_n)\right] &= \mathbb{E}_\vartheta\left[g\big(\widehat{\vartheta}_n(X_1,\dots,X_n)\big)\right]\\ &= \int_{\mathbb{R}^n} g\big(\widehat{\vartheta}_n(x_1,\dots,x_n)\big)\,\mathbb{P}_\vartheta(dx_1)\dots\mathbb{P}_\vartheta(dx_n)\\ &= \int_{\mathbb{R}^n} g\big(\widehat{\vartheta}_n(x_1,\dots,x_n)\big)\prod_{i=1}^n f(\vartheta,x_i)\mu(dx_1)\dots\mu(dx_n).\end{aligned}$$

Si μ est la mesure de Lebesgue, cette formule devient

$$\mathbb{E}_\vartheta\left[g(\widehat{\vartheta}_n)\right] = \int_{\mathbb{R}^n} g\big(\widehat{\vartheta}_n(x_1,\dots,x_n)\big)\prod_{i=1}^n f(\vartheta,x_i)dx_1\dots dx_n.$$

Si μ est la mesure de comptage sur $\mathcal{M}\subset\mathbb{R}$ au plus dénombrable, la formule devient

$$\mathbb{E}_\vartheta\left[g(\widehat{\vartheta}_n)\right] = \sum_{x_1,\dots,x_n\in\mathcal{M}} g\big(\widehat{\vartheta}_n(x_1,\dots,x_n)\big)\prod_{i=1}^n f(\vartheta,x_i).$$

Ces formules ne sont pas toujours « praticables » : on choisit souvent des fonctions tests et des estimateurs très particuliers pour pouvoir conduire les calculs.

4.1.2 Familles paramétriques classiques

1. *Loi gaussienne réelle et vectorielle*, que nous avons déja rencontré au Chapitre 1.
2. *Dérivées des lois gaussiennes.* Il s'agit de la loi du χ^2 à n degrés de liberté, la loi de Student à n degrés de libertés, et la loi de Fisher ou Fisher-Snedecor à (n_1,n_2) degrés de liberté, que nous avons déja rencontré au Chapitre 1.

[1] Un exemple plus subtil est donné par l'expérience engendrée par l'observation de ϑX, où X suit une loi de Poisson de paramètre 1, et $\vartheta\in\Theta=\mathbb{R}_+\setminus\{0\}$ est le paramètre. Dans ce cas, l'expérience est « vraiment aléatoire », mais on pourra montrer en exercice qu'elle n'est pas dominée. (Indication : la loi de X s'écrit $\mathbb{P}_\vartheta(dx)=\sum_{k\in\mathbb{N}}\frac{1}{k!}e^{-1}\delta_{\vartheta k}(dx)$. On raisonne alors de la même manière que pour l'expérience parfaite.

3. *Loi Gamma.* Notée $\Gamma_{\lambda,\alpha}$ de paramètres $\alpha > 0$ et $\lambda > 0$, de densité $\gamma_{\lambda,\alpha}$ par rapport à la mesure de Lebesgue

$$\gamma_{\lambda,\alpha}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} 1_{\{x\geq 0\}}$$

où $\Gamma(x) = \int_0^{+\infty} u^{x-1} e^{-u/2} du$. Si $X \sim \Gamma_{\lambda,\alpha}$, alors

$$\begin{aligned}\mathbb{E}\left[X^k\right] &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} x^{\alpha+k-1} e^{-\lambda x} dx \\ &= \frac{\lambda^{-k}}{\Gamma(\alpha+k)} \int_0^{+\infty} x^{\alpha+k-1} e^{-x} dx \\ &= \frac{\lambda^{-k}\Gamma(\alpha+k)}{\Gamma(\alpha)}.\end{aligned}$$

En particulier, $\mathbb{E}\left[X\right] = \alpha/\lambda$ et $\mathrm{Var}\left[X\right] = \alpha/\lambda^2$. Le paramètre λ joue un rôle de facteur d'échelle : on montre de la même manière que si $X \sim \Gamma_{1,\alpha}$, alors $X/\lambda \sim \Gamma_{\lambda,\alpha}$. C'est donc le deuxième paramètre qui est important en modélisation. En particulier, la loi du χ^2 à n degrés de libertés est la loi $\Gamma_{1/2,n/2}$.

4. *Loi exponentielle.* C'est la loi $\Gamma_{\lambda,1}$, $\lambda > 0$, de densité $\lambda e^{-\lambda x} 1_{\{x\geq 0\}}$. En particulier, sa moyenne vaut $1/\lambda$ et sa variance $1/\lambda^2$.

5. *Loi Béta.* De paramètres $\lambda_1, \lambda_2 > -1$. C'est une loi sur $[0,1]$, de densité

$$x \rightsquigarrow \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} x^{\lambda_1-1}(1-x)^{\lambda_2-1} 1_{\{x\in(0,1)\}}.$$

Son nom vient de la fonction Béta

$$B(\lambda_1,\lambda_2) = \int_0^1 x^{\lambda_1-1}(1-x)^{\lambda_2-1} dx = \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)}{\Gamma(\lambda_1+\lambda_2)}.$$

Si X suit la loi Béta de paramètres (λ_1,λ_2), ses moments – s'ils existent – sont donnés par la formule

$$\mathbb{E}\left[X^k\right] = \int_0^1 \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} x^{\lambda_1+k-1}(1-x)^{\lambda_2-1} = \frac{\Gamma(\lambda_1+\lambda_2)\Gamma(\lambda_1+k)}{\Gamma(\lambda_1)\Gamma(\lambda_1+\lambda_2+k)}.$$

En particulier, pour $k = 1, 2$ on obtient

$$\mathbb{E}\left[X\right] = \frac{\lambda_1}{\lambda_1+\lambda_2}, \quad \mathbb{E}\left[X^2\right] = \frac{\lambda_1(\lambda_1+1)}{(\lambda_1+\lambda_2)(\lambda_1+\lambda_2+1)}.$$

6. *Loi uniforme.* Sur $[0,1]$, on peut la voir comme un cas particulier de la loi Béta[2] pour $\lambda_1 = \lambda_2 = 1$.

[2]Le lien entre loi uniforme et loi Béta intervient dans le calcul de la statistique de rang associé à des tirages uniformes, dont une application fondamentale est la loi limite d'estimation de quantiles, voir par exemple [1], p.46.

7. *Loi de Cauchy.* C'est la loi de paramètres $\alpha \in \mathbb{R}$ et $\sigma^2 > 0$ de densité

$$x \rightsquigarrow \frac{\sigma}{\pi\big(\sigma^2 + (x-\alpha)\big)^2} = \frac{1}{\pi\sigma}\frac{1}{1+\big((x-\alpha)/\sigma^2\big)^2}$$

sur $\mathbb{R}$. Ce n'est rien d'autre que la famille de translations-dilatations associée à la loi de Cauchy standard de densité

$$x \rightsquigarrow \frac{1}{\pi(1+x^2)}$$

mais à la différence de la famille des lois normales, elle n'admet pas de moment d'ordre 1 (et donc pas de variance non plus).

8. *Loi log-normale* On dit qu'un variable Y est log-normale si elle peut s'écrire $Y = \exp(X)$, avec $X \sim \mathcal{N}(\mu, \sigma^2)$. La densité de la loi log-normale est

$$x \rightsquigarrow \frac{1}{x} g\big(\log(x)\big),$$

où $g(x) = (2\pi^{-1/2})\exp(-x^2/2)$ est la densité de la loi normale standard. De plus,

$$\mathbb{E}\left[Y\right] = e^{\mu+\sigma^2/2}, \quad \mathbb{E}\left[Y^2\right] = e^{2\mu+2\sigma^2}.$$

9. *Loi de Bernoulli.* Rencontrée au Chapitre 1

10. *Loi de Poisson.* Rencontrée au Chapitre 1. Si X suit une loi de Poisson de paramètre $\lambda > 0$, alors $\mathbb{E}\left[X\right] = \mathrm{Var}\left[X\right] = \lambda$.

11. *Loi multinômiale.* Si $X_1, \ldots, X_n$ sont des variables aléatoires à valeurs dans $\{1, \ldots, d\}$, indépendantes et de même loi

$$\mathbb{P}\left[X = \ell\right] = p_\ell, \quad \ell = 1, \ldots, d,$$

alors, si l'on note $N_\ell = \sum_{i=1}^n 1_{\{X_i = \ell\}}$ le nombre de tirages ayant donné la valeur ℓ, le vecteur $(N_1, \ldots, N_\ell)$ suit la loi multinômiale de paramètres n et $(p_1, \ldots, p_d)$, donnée par

$$\mathbb{P}\left[N_1 = n_1, \ldots, N_d = n_d\right] = \frac{n!}{n_1! \cdots n_d!} p_1^{n_1} \cdots p_d^{n_d}, \quad \sum_{\ell=1}^d n_\ell = 1.$$

La loi multinômiale généralise la loi binômiale, qui est un cas particulier de loi multinômiale avec $d = 2$. Cette loi est fondamentale dans l'utilisation du test du χ^2 du Chapitre 8.

4.2 Méthode des moments

4.2.1 La cas de la dimension 1

On suppose $\vartheta \subset \mathbb{R}$. Supposons donnée une application $g : \mathbb{R} \to \mathbb{R}$ telle que

$$\vartheta \rightsquigarrow m(\vartheta) = \mathbb{E}_\vartheta \big[g(X)\big]$$

existe et soit strictement monotone et continue. Alors m réalise une bijection de Θ sur son image $m(\Theta)$ et on a la représentation

$$\vartheta = m^{-1}\big(\mathbb{E}_\vartheta \big[g(X)\big]\big), \quad \vartheta \in \Theta.$$

En remplaçant la moyenne théorique inconnue $m(\vartheta) = \mathbb{E}_\vartheta \big[g(X)\big]$ par sa version empirique $\frac{1}{n}\sum_{i=1}^n g(X_i)$, observable, un estimateur naturel de ϑ est donc

$$\widehat{\vartheta}_n = m^{-1}\left(\frac{1}{n}\sum_{i=1}^n g(X_i)\right). \tag{4.1}$$

Une autre façon de voir cette approche est de remarquer que si F_ϑ désigne la fonction de répartition de la loi $\mathbb{P}_\vartheta$, alors

$$\vartheta = T(F_\vartheta) = m^{-1}\left(\int_{\mathbb{R}} g(x) dF_\vartheta(x)\right),$$

où T est une fonctionnelle de type (3.7) étudiée au chapitre précédent. On a donc aussi

$$\widehat{\vartheta}_n = T(\widehat{F}_n) = m^{-1}\left(\frac{1}{n}\sum_{i=1}^n g(X_i)\right).$$

Remarque 4.3. Si $\frac{1}{n}\sum_{i=1}^n g(X_i) \notin m(\Theta)$, on peut modifier $\widehat{\vartheta}_n$ de la manière suivante : on note $x \rightsquigarrow P_\Theta[x]$ la fonction qui réalise le minimum[3] de distance entre x et l'adhérence de Θ. Alors

$$\widetilde{\vartheta}_n = P_\Theta\big[m^{-1}\big(\tfrac{1}{n}\sum_{i=1}^n g(X_i)\big)\big] \tag{4.2}$$

est un autre estimateur de ϑ qui vérifie $|\widetilde{\vartheta}_n - \vartheta| \leq |\widehat{\vartheta}_n - \vartheta|$, $\mathbb{P}_\vartheta$-presque-sûrement.

Définition 4.1. *On appelle estimateur par méthode des moments tout estimateur de la forme* (4.1) *ou* (4.2).

Remarque 4.4. Dans la plupart des exemples, on choisit g de la forme $g(x) = x^k$ avec $k \geq 1$, d'où la terminologie. Le choix g est arbitraire pour le statisticien : il y a donc tout un ensemble de possibilités pour construire un estimateur par méthode des moments, mais sous la contrainte que l'application $\vartheta \rightsquigarrow m(\vartheta)$ soit régulière et inversible.

[3] qui n'est pas nécessairement unique. Mais dans la plupart des exemples, $m(\Theta)$ est un intervalle et le problème ne se pose pas.

Sous des hypothèses de régularité sur m et d'intégrabilité sur g, on a le comportement asymptotique de $\widehat{\vartheta}$ suivant.

Proposition 4.1. *Si* $\mathbb{E}_\vartheta\left[|g(X)|\right] < +\infty$ *et si* m^{-1} *est continue, on a*

$$\widehat{\vartheta}_n \overset{\text{p.s.}}{\to} \vartheta.$$

De plus, si pour tout $\vartheta \in \Theta$, $\mathbb{E}_\vartheta\left[g(X)^2\right] < +\infty$ *et si la fonction* m *est dérivable, alors*

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{m'(\vartheta)^2}\operatorname{Var}_\vartheta\left[g(X_1)\right]\right) \tag{4.3}$$

Démonstration. On applique simplement les Propositions 3.7 et 3.8 du Chapitre 3 à la fonctionnelle régulière $T(F_\vartheta)$. □

Exemple 4.3 (Loi exponentielle)**.** On considère l'expérience $\mathcal{E}^n$ engendrée par l'observation d'un n-échantillon de variables exponentielles de paramètres $\vartheta > 0$. Les fonctions les plus simples pour construire un estimateur sont par exemple $g(x) = x$ ou $\widetilde{g}(x) = x^2$. Ceci fournit deux estimateurs. On part de l'équation

$$m(\vartheta) = \mathbb{E}_\vartheta\left[g(X)\right] = \int_0^{+\infty} x\vartheta\exp(-\vartheta x)dx = \frac{1}{\vartheta}$$

ou bien

$$\widetilde{m}_2(\vartheta) = \mathbb{E}_\vartheta\left[\widetilde{g}(X)\right] = \int_0^{+\infty} x^2\vartheta\exp(-\vartheta x)dx = \frac{2}{\vartheta^2},$$

et on résout

$$m(\vartheta) = \tfrac{1}{n}\sum_{i=1}^n X_i \quad \text{ou} \quad \widetilde{m}(\vartheta) = \tfrac{1}{n}\sum_{i=1}^n X_i^2.$$

On obtient deux estimateurs par substitution :

$$\widehat{\vartheta}_{n,1} = \frac{1}{\frac{1}{n}\sum_{i=1}^n X_i}, \quad \text{et} \quad \widehat{\vartheta}_{n,2} = \left(\frac{2}{n}\sum_{i=1}^n X_i^2\right)^{1/2}.$$

La Proposition 4.1 s'applique, et, comme

$$\operatorname{Var}_\vartheta\big[g(X)\big] = \frac{1}{\vartheta^2} \quad \text{et} \quad \operatorname{Var}_\vartheta\big[\widetilde{g}(X)\big] = \frac{20}{\vartheta^4}$$

et

$$m'(\vartheta) = -\frac{1}{\vartheta^2} \quad \text{et} \quad \widetilde{m}'(\vartheta) = -\frac{4}{\vartheta^3}$$

on obtient la convergence en loi (4.3) de l'erreur renormalisée $\sqrt{n}(\widehat{\vartheta}_{n,i} - \vartheta)$ pour $i = 1, 2$ vers une gaussienne centrée de variance

$$v(\vartheta) = m'(\vartheta)^{-2}\operatorname{Var}_\vartheta\big[g(X)\big] = \vartheta^2$$

et

$$\widetilde{v}(\vartheta) = \widetilde{m}'(\vartheta)^{-2}\mathrm{Var}_\vartheta\big[\widetilde{g}(X)\big] = \frac{20}{\vartheta^4}\frac{\vartheta^6}{16} = \frac{5}{4}\vartheta^2$$

respectivement. L'erreur de l'estimateur $\widehat{\vartheta}_{n,1}$ est « moins dispersée » que celle de $\widehat{\vartheta}_{n,2}$ et de ce point de vue, $\widehat{\vartheta}_{n,1}$ semble « préférable » à $\widehat{\vartheta}_{n,2}$. Nous étudierons plus systématiquement la comparaison d'estimateurs au Chapitre 6.

Exemple 4.4 (Loi de Cauchy)**.** On considère la famille de translation (voir 4.1.2) associée à la loi de Cauchy sur $\mathbb{R}$. La loi $\mathbb{P}_\vartheta$ a une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$

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

La densité $f(\vartheta, \bullet)$ n'a pas de moment d'ordre k pour $k \geq 1$, et le choix $g(x) = x^k$ avec k entier ne s'applique pas ici. Prenons $g(x) = \mathrm{signe}(x)$, avec

$$\mathrm{signe}(x) = \begin{cases} -1 & \text{si} \quad x \leq 0 \\ 1 & \text{si} \quad x > 0. \end{cases}$$

On a

$$\mathbb{E}_\vartheta\left[g(X_1)\right] = \int_{\mathbb{R}} \mathrm{signe}(x) f(\vartheta, x) dx = 1 - 2F(-\vartheta),$$

où

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

On résout

$$\frac{2}{\pi}\mathrm{Arctg}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{signe}(X_i),$$

d'où l'estimateur

$$\widehat{\vartheta}_n = \mathrm{tg}\left(\frac{\pi}{2n}\sum_{i=1}^{n}\mathrm{signe}(X_i)\right).$$

Les propriétés asymptotiques de $\widehat{\vartheta}_n$ vers ϑ s'obtiennent en appliquant la Proposition 4.1.

4.2.2 Le cas multidimensionnel

Lorsque $\Theta \subset \mathbb{R}^d$ avec $d \geq 1$, il n'est plus possible en général d'identifier ϑ via une seule fonction g via la représentation (3.7). On étend la méthode précédente en identifiant ϑ à l'aide de d applications $g_\ell : \mathbb{R} \to \mathbb{R}$, pour $\ell = 1, \ldots, d$

$$x \rightsquigarrow \big(g_1(x), \ldots, g_d(x)\big), \quad x \in \mathbb{R},$$

de sorte que le système d'équations

$$m_\ell(\vartheta) = \mathbb{E}_\vartheta\left[g_\ell(X)\right] = \int_{\mathbb{R}} g_\ell(x) dF_\vartheta(x), \quad \ell = 1, \ldots, d \tag{4.4}$$

admette une solution unique, lorsque cela est possible. Un estimateur par méthode de moment est alors tout estimateur $\widehat{\vartheta}_n$ satisfaisant

$$m_\ell(\widehat{\vartheta}_n) = \frac{1}{n}\sum_{i=1}^{d} g_\ell(X_i), \quad \ell = 1, \ldots, d. \tag{4.5}$$

Définition 4.2. *On appelle estimateur par substitution ou par méthode de moment associé à la fonction $\boldsymbol{g}$ tout estimateur $\widehat{\vartheta}_n$ solution de* (4.5).

On note

$$\boldsymbol{m}(\vartheta) = \mathbb{E}_\vartheta\left[\boldsymbol{g}(X)\right] = \left(\mathbb{E}_\vartheta\left[g_1(X)\right], \ldots, \mathbb{E}_\vartheta\left[g_d(X)\right]\right)$$

l'application de $\mathbb{R}^d \to \mathbb{R}^d$ définie composante par composante par (4.4). On utilise donc la représentation

$$\vartheta = \boldsymbol{m}^{-1}\left(m_1(\vartheta), \ldots, m_d(\vartheta)\right)$$

pour estimer ϑ par

$$\widehat{\vartheta}_n = \boldsymbol{m}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_1(X_i), \ldots, \frac{1}{n}\sum_{i=1}^{n} g_d(X_i)\right)$$

Proposition 4.2. *Si $\boldsymbol{m}$ est continue, inversible et d'inverse continue, alors l'estimateur par méthode des moments est bien défini et on a*

$$\widehat{\vartheta}_n \overset{\text{p.s.}}{\to} \vartheta.$$

sous $\mathbb{P}_\vartheta$. De plus, si $\boldsymbol{m}^{-1}$ est différentiable et si $\mathbb{E}_\vartheta\left[g_\ell(X)^2\right] < +\infty$, on a la convergence

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \overset{d}{\to} \mathcal{N}\big(0, V(\vartheta)\big),$$

où

$$V(\vartheta) = J_{\boldsymbol{m}^{-1}} \Sigma_{\boldsymbol{m}}(\vartheta)\, J_{\boldsymbol{m}^{-1}}^T, \tag{4.6}$$

avec $\Sigma_{\boldsymbol{m}}(\vartheta)$ la matrice de variance-covariance du vecteur $\big(g_1(X), \ldots, g_d(X)\big)^T$ définie par

$$\big(\Sigma_{\boldsymbol{m}}(\vartheta)\big)_{\ell,\ell'} = \mathbb{E}_\vartheta\left[g_\ell(X) g_{\ell'}(X)\right] - \mathbb{E}_\vartheta\left[g_\ell(X)\, \mathbb{E}_\vartheta\left[g'_\ell(X)\right]\right] \tag{4.7}$$

et $J_{\boldsymbol{m}^{-1}}$ désigne la matrice de la différentielle de $\boldsymbol{m}^{-1}$.

Remarque 4.5. Ce résultat est très proche du Corollaire 3.2 du Chapitre 3 (la fonction $\boldsymbol{m}^{-1}$ jouant le rôle de $\boldsymbol{g}$ dans le Corollaire 3.2).

Démonstration. Par la loi des grands nombres, on a, composante par composante, la convergence

$$\left(\frac{1}{n}\sum_{i=1}^{n} g_1(X_i), \dots, \frac{1}{n}\sum_{i=1}^{n} g_d^{-1}(X_i)\right) \xrightarrow{\text{p.s.}} \Big(\mathbb{E}_\vartheta\left[g_1(X)\right], \dots, \mathbb{E}_\vartheta\left[g_d(X)\right]\Big) = \boldsymbol{m}(\vartheta).$$

sous $\mathbb{P}_\vartheta$. Par continuité de $\boldsymbol{m}^{-1}$, on en déduit

$$\begin{aligned}\widehat{\vartheta}_n &= \boldsymbol{m}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_1(X_i), \dots, \frac{1}{n}\sum_{i=1}^{n} g_d^{-1}(X_i)\right)\\ &\xrightarrow{\text{p.s.}} \boldsymbol{m}^{-1}\Big(\mathbb{E}_\vartheta\left[g_1(X)\right], \dots, \mathbb{E}_\vartheta\left[g_d(X)\right]\Big)\\ &= \boldsymbol{m}^{-1}\big(\boldsymbol{m}(\vartheta)\big)\\ &= \vartheta.\end{aligned}$$

La deuxième partie est la méthode « delta » multidimensionnelle. On applique d'abord Théorème 1.4 (théorème central limite vectoriel) : la suite de vecteurs

$$\left(\frac{1}{n}\sum_{i=1}^{n} g_1(X_i), \dots, \frac{1}{n}\sum_{i=1}^{n} g_d(X_i)\right)^T$$

est asymptotiquement gaussienne, et

$$\sqrt{n}\Big(\big(\tfrac{1}{n}\sum_{i=1}^{n} g_1(X_i), \dots, \tfrac{1}{n}\sum_{i=1}^{n} g_d(X_i)\big)^T - \boldsymbol{m}(\vartheta)\Big) \xrightarrow{d} \mathcal{N}\big(0, \Sigma_m(\vartheta)\big),$$

sous $\mathbb{P}_\vartheta$, de matrice de variance-covariance $\Sigma_m(\vartheta)$ donnée par (4.7). Puis, on applique la Proposition 1.11 (méthode delta) avec $\boldsymbol{g} = \boldsymbol{m}^{-1}$. □

Exemple 4.5. Si $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$ et $\mathbb{P}_\vartheta$ est la loi $\mathcal{N}(\mu, \sigma^2)$, alors $d = 2$ et les fonctions $g_1(x) = x$ et $g_2(x) = x^2$ fournissent le système d'équations

$$\mu = \overline{X}_n, \quad \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

dont la solution est

$$\widehat{\vartheta}_n = \big(\widehat{\mu}_n, \widehat{\sigma}_n^2\big)^T = \big(\,\overline{X}_n, \tfrac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n{}^2\,\big)^T. \tag{4.8}$$

On retrouve l'estimation de fonctionnelles du Chapitre 3. L'estimateur $\widehat{\vartheta}_n$ est asymptotiquement normal. On peut calculer sa variance asymptotique en appliquant la formule

(4.6) de la Proposition 4.2 ci-dessus ou bien en partant directement de la représentation (4.8) et en appliquant alors le Corollaire 3.2 du Chapitre 3. En notant $\mu_i = \mathbb{E}\left[X^i\right]$, on obtient finalement

$$V(\vartheta) = \begin{pmatrix} \mu_2 - \mu_1^2 & -3\mu_1\mu_2 + 2\mu_1^3 + \mu_3 \\ -3\mu_1\mu_2 + 2\mu_1^3 + \mu_3 & 2\mu_1(4\mu_1\mu_2 - 2\mu_1^3 - 2\mu_3) + \mu_4 - \mu_2^2 \end{pmatrix}.$$

En particulier, dans le cas d'une distribution centrée, lorsque $\mu_1 = 0$, on retrouve la forme particulièrement simple

$$V(\vartheta) = \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}.$$

4.3 Moments généralisés. Z- et M-estimation

Insuffisance de la méthode des moments

La méthode des moments repose sur l'existence d'une fonction m (réelle ou vectorielle) inversible qui n'est pas toujours facile à déterminer ou à mettre en oeuvre numériquement. On présente une extension naturelle qui fournit une classe d'estimateurs vaste que l'on va pouvoir étudier de manière systématique.

En particulier, sous des hypothèses de régularité suffisantes, on pourra construire une méthode « automatique » de sélection d'un estimateur asymptotiquement optimal, dans un sens que nous discuterons au Chapitre 6.

4.3.1 Z-estimateurs

Construction en dimension 1

Lorsque le paramètre ϑ est de dimension 1, c'est-à-dire $\Theta \subset \mathbb{R}$, la méthode des moments de la section précédente repose sur de bonnes propriétés – régularité, inversibilité – de l'application

$$m(\vartheta) = m_g(\vartheta) = \int_{\mathbb{R}} g(x)\, \mathbb{P}_\vartheta(dx) \tag{4.9}$$

pour un certain choix de fonction g. Autrement dit, on a, pour tout $\vartheta \in \Theta$

$$\int_{\mathbb{R}} \big(m_g(\vartheta) - g(x)\big)\, \mathbb{P}_\vartheta(dx) = 0, \tag{4.10}$$

où g est à choisir. Considérons de manière générale pour $\Theta \subset \mathbb{R}^d$ et $d \geq 1$ une application

$$\phi : \Theta \times \mathbb{R} \to \mathbb{R}$$

telle que pour tout $\vartheta \in \Theta$

$$\int_{\mathbb{R}} \phi(\vartheta, x)\, \mathbb{P}_\vartheta(dx) = 0 \tag{4.11}$$

dont (4.10) est un cas particulier avec $\phi(\vartheta, x) = m_g(\vartheta) - g(x)$. Pour construire un estimateur, on peut se donner une application ϕ satisfaisant l'équation (4.11) pour tout $\vartheta \in \Theta$ et résoudre sa version empirique, c'est-à-dite chercher un estimateur $\widehat{\vartheta}_n$ satisfaisant

$$\frac{1}{n}\sum_{i=1}^{n} \phi(\widehat{\vartheta}_n, X_i) = 0. \tag{4.12}$$

Définition 4.3 (Z-Estimateur ou estimateur GMM[4])**.** *Etant donnée une application $\phi : \Theta \times \mathbb{R} \to \mathbb{R}$ satisfaisant* (4.11)*, on appelle Z-estimateur associé à ϕ tout estimateur $\widehat{\vartheta}_n$ satisfaisant* (4.12)*.*

Le cas multidimensionnel

L'extension au cas multi-dimensionnel $\Theta \subset \mathbb{R}^d$, avec $d \geq 1$ est immédiate. La fonction ϕ est remplacée par une application

$$\Phi = (\phi_1, \ldots, \phi_d) : \Theta \times \mathbb{R} \to \mathbb{R}^d$$

où chaque composante $\phi_\ell : \Theta \times \mathbb{R} \to \mathbb{R}$ joue le même rôle qu'en dimension 1. Pour que la méthode ait un sens, il faut que, comme pour l'équation (4.11), le paramètre inconnu ϑ soit solution du système d'équations

$$\int_{\mathbb{R}} \phi_\ell(\vartheta, x)\, \mathbb{P}_\vartheta(dx) = 0, \quad \ell = 1, \ldots, d \tag{4.13}$$

et construire un Z-estimateur revient à résoudre une version empirique de (4.13).

Définition 4.4 (Z-estimateur, cas multidimensionnel)**.** *Etant donné une application $\Phi : \Theta \times \mathbb{R} \to \mathbb{R}^d$, on appelle Z-estimateur associé à Φ tout estimateur $\widehat{\vartheta}_n$ satisfaisant*

$$\frac{1}{n}\sum_{i=1}^{n} \Phi(\widehat{\vartheta}_n, X_i) = 0, \quad \ell = 1, \ldots, d.$$

4.3.2 M-estimateurs

Soit $\psi : \Theta \times \mathbb{R} \to \mathbb{R}_+$ une application telle que, pour tout $\vartheta \in \Theta \subset \mathbb{R}^d$, avec $d \geq 1$, la fonction

$$a \rightsquigarrow \mathbb{E}_\vartheta\left[\psi(a, X)\right] = \int_{\mathbb{R}} \psi(a, x)\, \mathbb{P}_\vartheta(dx) \tag{4.14}$$

admette un maximum en $a = \vartheta$. Une procédure naturelle pour estimer ϑ consiste à maximiser une version empirique de (4.14).

[4] Z pour zéro et GMM pour Generalized Method of Moments

Définition 4.5. *On appelle M-estimateur*[5] *associé au contraste ψ tout estimateur $\widehat{\vartheta}_n$ qui satisfait*

$$\frac{1}{n}\sum_{i=1}^{n}\psi(\widehat{\vartheta}_n, X_i) = \max_{a\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\psi(a, X_i).$$

Si le paramètre ϑ est de dimension $d = 1$ et si l'on suppose, pour tout $x \in \mathbb{R}$ que la fonction $a \rightsquigarrow \psi(a, x)$ est régulière, en posant

$$\phi(a, x) = \partial_1\psi(a, x),$$

on a

$$\sum_{i=1}^{n}\partial_1\psi(\widehat{\vartheta}_n, X_i) = \sum_{i=1}^{n}\phi(\widehat{\vartheta}_n, X_i) = 0$$

ce qui permet – dans ce cas – d'interpréter un M-estimateur comme un Z-estimateur. Cette interprétation s'étend immédiatement au cas multidimensionnel.

Exemple 4.6. On considère les lois $\{\mathbb{P}_\vartheta, \vartheta \in \Theta = \mathbb{R}\}$ qui est la famille de translations $\big\{F(\bullet - \vartheta),\, \vartheta \in \mathbb{R}\,\big\}$ associée à une distribution donnée F centrée et ayant un moment d'ordre 1. On a

$$\vartheta = \int_{\mathbb{R}} x\,\mathbb{P}_\vartheta(dx) = \int_{\mathbb{R}}(x + \vartheta)dF(x).$$

Alors $m(\vartheta) = \mathbb{E}_\vartheta\left[X\right]$ minimise la fonction

$$a \rightsquigarrow \int_{\mathbb{R}}(x - a)^2\,\mathbb{P}_\vartheta(dx) = \mathbb{E}_\vartheta\left[(X - a)^2\right]$$

d'après la Proposition 1.1. En prenant $\psi(a, x) = -(x - a)^2$, le M-estimateur associé à ψ satisfait

$$\sum_{i=1}^{n}\psi(\widehat{\vartheta}_n, X_i) = \max_{a\in\mathbb{R}}\sum_{i=1}^{n}\psi(a, X_i)$$

ou encore

$$\sum_{i=1}^{n}\phi(\widehat{\vartheta}_n, X_i) = 0$$

avec $\phi(a, x) = \partial_1\psi(a, x) = -2(x - a)$, ce qui implique $\sum_{i=1}^{n}(X_i - \widehat{\vartheta}_n) = 0$, d'où l'estimateur $\widehat{\vartheta}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Dans cet exemple simple, tous les points de vue coincident.

[5] Il y a peut-être un problème de mesurabilité à régler pour garantir que l'on obtient effectivement un estimateur. Nous ignorons ce problème éventuel.

4.3.3 Convergence des Z- et des M-estimateurs

Dans cette Section, nous donnons des critères simples sur la famille $\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\}$ et la fonction ϕ –pour les Z-estimateurs– ou ψ –pour les M-estimateurs– qui garantissent la convergence de l'estimateur correspondant. Nos conditions sont classiques et sous-optimales. La recherche de conditions minimales est un problème délicat qui dépasse le cadre de ce cours. On pourra consulter van der Vaart [7] pour une discussion accessible sur le sujet. Pour des raisons techniques, nous commençons par traiter la convergence des M-estimateurs, dont nous déduirons celle des Z-estimateurs.

Convergence des M-estimateurs

Pour une fonction de contraste $\psi : \Theta \times \mathbb{R} \to \mathbb{R}$ donnée, on définit

$$M_n(a) = \frac{1}{n}\sum_{i=1}^{n} \psi(a, X_i), \quad a \in \Theta$$

et, pour $\vartheta \in \Theta$,

$$M(a, \vartheta) = \mathbb{E}_\vartheta\left[\psi(a, X)\right].$$

Proposition 4.3 (Convergence des M-estimateurs)**.** *On suppose $\Theta \subset \mathbb{R}^d$, avec $d \geq 1$, que le M-estimateur $\widehat{\vartheta}_n$ associé à la fonction ψ est bien défini, et qu'on a*

(i) $\sup_{a\in\Theta} \left|M_n(a) - M(a,\vartheta)\right| \overset{\mathbb{P}_\vartheta}{\to} 0$,

(ii) $\forall \varepsilon > 0,\ \sup_{|a-\vartheta|\geq\varepsilon} M(a,\vartheta) < M(\vartheta,\vartheta)$, *(condition de maximum)*

(iii) $M_n(\widehat{\vartheta}_n) \geq M_n(\vartheta) - \varepsilon_n$, *où* $\varepsilon_n \xrightarrow{\mathbb{P}_\vartheta} 0$.

Alors le M-estimateur $\widehat{\vartheta}_n$ est convergent (ou consistant) :

$$\widehat{\vartheta}_n \xrightarrow{\mathbb{P}_\vartheta} \vartheta.$$

Démonstration. On écrit

$$M(\vartheta,\vartheta) - M(\widehat{\vartheta}_n,\vartheta) = T_{n,1} + T_{n,2} + T_{n,3},$$

avec

$$\begin{aligned} T_{n,1} &= M(\vartheta,\vartheta) - M_n(\vartheta),\\ T_{n,2} &= M_n(\vartheta) - M_n(\widehat{\vartheta}_n),\\ T_{n,3} &= M_n(\widehat{\vartheta}_n) - M(\widehat{\vartheta}_n,\vartheta). \end{aligned}$$

Les termes $T_{n,1}$ et $T_{n,3}$ tendent vers 0 en probabilité sous $\mathbb{P}_\vartheta$ grâce à l'hypothèse (i).

Soit $\varepsilon > 0$. D'après la condition (ii), il existe $\eta > 0$ tel que $M(a,\vartheta) \leq M(\vartheta,\vartheta) - \eta$ dès lors que $|a - \vartheta| \geq \varepsilon$. On a donc l'inclusion

$$\big\{|\,\widehat{\vartheta}_n - \vartheta| \geq \varepsilon\big\} \subset \big\{M(\widehat{\vartheta}_n,\vartheta) \leq M(\vartheta,\vartheta) - \eta\big\} \tag{4.15}$$

en prenant $a = \widehat{\vartheta}_n$. Il vient

$$\begin{aligned}
\mathbb{P}_\vartheta\left[|\,\widehat{\vartheta}_n - \vartheta| \geq \varepsilon\right] &\leq \mathbb{P}_\vartheta\left[M(\widehat{\vartheta}_n,\vartheta) < M(\vartheta,\vartheta) - \eta\right] \\
&= \mathbb{P}_\vartheta\left[M(\vartheta,\vartheta) - M(\widehat{\vartheta}_n,\vartheta) > \eta\right] \\
&\leq \mathbb{P}_\vartheta\left[T_{n,1} + \varepsilon_n + T_{n,3} \geq \eta\right] \\
&\xrightarrow{\mathbb{P}_\vartheta} 0
\end{aligned}$$

où l'on utilise successivement l'inclusion (4.15), l'hypothèse (iii) et le fait que chacun des termes $T_{n,1}$, ε_n et $T_{n,3}$ tend vers 0 en probabilité sous $\mathbb{P}_\vartheta$. □

Convergence des Z-estimateurs

On suppose d'abord $\Theta \subset \mathbb{R}$. Pour une fonction ϕ donnée, on définit

$$Z_n(a) = \frac{1}{n}\sum_{i=1}^{n} \phi(a, X_i), \quad a \in \Theta$$

et, pour $\vartheta \in \Theta$,

$$Z(a,\vartheta) = \mathbb{E}_\vartheta\left[\phi(a,X)\right] \quad a \in \Theta.$$

Proposition 4.4 (Convergence des Z-estimateurs)**.** *On suppose que le Z-estimateur $\widehat{\vartheta}_n$ associé à la fonction ϕ est bien défini, et qu'on a*

(i) $\sup_{a\in\Theta}\left|Z_n(a) - Z(a,\vartheta)\right| \xrightarrow{\mathbb{P}_\vartheta} 0$,

(ii) $\forall \varepsilon > 0, \ \inf_{|a-\vartheta|\geq\varepsilon} |Z(a,\vartheta)| > 0 = |Z(\vartheta,\vartheta)|$,

(iii) $Z_n(\widehat{\vartheta}_n) \xrightarrow{\mathbb{P}_\vartheta} 0$.

Alors le Z-estimateur $\widehat{\vartheta}_n$ est convergent (ou consistant) :

$$\widehat{\vartheta}_n \xrightarrow{\mathbb{P}_\vartheta} \vartheta.$$

Démonstration. Il suffit d'appliquer la Proposition 4.3 avec $M_n(a) = -|Z_n(a)|$. □

Le cas multidimensionnel où $\Theta \subset \mathbb{R}^d$ avec $d \geq 1$ se traite de la même manière, en remplaçant la fonction ϕ par une fonction vectorielle $\Phi = (\phi_1, \ldots, \phi_d)$ et les valeurs absolues dans les conditions (i)–(ii)–(iii) par la norme euclidienne sur $\mathbb{R}^d$.

4.3.4 Loi limite des Z- et M-estimateurs

Nous précisons les résultats de la section précédente, en cherchant une vitesse de convergence $\alpha_n \to \infty$ de sorte que l'erreur normalisée

$$\alpha_n(\widehat{\vartheta}_n - \vartheta)$$

converge vers une limite non-dégénérée. Nous donnons des hypothèses suffisantes sur les fonctions ϕ –pour les Z-estimateurs– et ψ –pour les M-estimateurs de sorte qu'on ait une convergence en loi vers une gaussienne avec la normalisation $\alpha_n = \sqrt{n}$. Ces conditions ne sont pas optimales (voir van deer Vaart [7]). A l'inverse de la section précédente, nous partons d'un résultat sur les Z-estimateurs pour en déduire un résultat sur les M-estimateurs.

Loi limite des Z-estimateurs

Nous donnons les résultats dans le cas $\Theta \subset \mathbb{R}$, lorsque le paramètre ϑ est de dimension $d = 1$, pour simplifier[6] Etant donnés, d'une part une fonction $\phi : \Theta \times \mathbb{R} \to \mathbb{R}$ définissant un Z-estimateur, et d'autre part la famille $\left\{ \mathbb{P}_\vartheta,\, \vartheta \in \Theta \right\}$, on fait le jeu d'hypothèses suivant

Hypothèse 4.2 (Hypothèse loi limite Z-estimateurs)**.** *On a*

(i) La famille $\left\{ \mathbb{P}_\vartheta,\, \vartheta \in \Theta \right\}$ est dominée par une mesure σ-finie μ sur $\mathbb{R}$.

(ii) Pour tout point $\vartheta \in \Theta$, il existe un voisinage ouvert $\mathcal{V}(\vartheta)$ tel que, pour tout $a \in \mathcal{V}(\vartheta)$

$$\left|\partial_a^2 \phi(a, x)\right| \leq g(x), \quad \text{où} \quad \mathbb{E}_\vartheta \left[g(X) \right] < +\infty.$$

(iii) Pour tout $\vartheta \in \Theta$, on a

$$\mathbb{E}_\vartheta \left[\phi(\vartheta, X) \right] = 0, \quad \mathbb{E}_\vartheta \left[\phi(\vartheta, X)^2 \right] < +\infty, \quad \mathbb{E}_\vartheta \left[\partial_\vartheta \phi(\vartheta, X) \right] \neq 0.$$

Remarque 4.6. Le jeu d'hypothèse 4.2 peut paraître un peu « repoussant » à première vue. Nous verrons que la méthode de preuve est très simple, et que ces hypothèses apparaîssent très naturellement de façon à contrôler les différents termes d'un développement asymptotique[7].

Remarque 4.7. Le jeu d'hypothèse 4.2 est local : comme le suggère l'hypothèse (ii), on doit pouvoir contrôler le comportement de la famille $\left\{ \mathbb{P}_\vartheta,\, \vartheta \in \Theta \right\}$ dans un voisinage de ϑ, pour tout ϑ. Ceci exclut les paramètres de la frontière de Θ dans le cas où Θ n'est pas un ouvert. En restreignant l'espace des paramètres (donc en considérant une expérience statistique « plus petite »), on pourra souvent se ramener au jeu d'hypothèses 4.2 à condition que Θ soit d'intérieur non vide au départ.

[6] Le passage au cas multidimensionnel ne présente essentiellement qu'une difficulté d'écriture.

[7] on peut « presque » les oublier et ne retenir que la méthode de preuve où elles réapparaîtront de façon évidente.

Sous ce jeu d'hypothèses, on a le comportement asymptotique suivant pour les Z-estimateurs

Proposition 4.5 (Loi limite des Z-estimateurs)**.** *Si la famille* $\{\mathbb{P}_\vartheta,\ \vartheta\in\Theta\}$ *et la fonction* ϕ *vérifient l'Hypothèse 4.2, alors, si* $\widehat{\vartheta}_n$ *est un* Z*-estimateur associé à* ϕ *tel que* $\widehat{\vartheta}_n \xrightarrow{\mathbb{P}_\vartheta} \vartheta$*, on a*

$$\sqrt{n}\big(\widehat{\vartheta}_n-\vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v_\phi(\vartheta)\big)$$

en loi sous $\mathbb{P}_\vartheta$*, où*

$$v_\phi(\vartheta)=\frac{\mathbb{E}_\vartheta\left[\phi(\vartheta,X)^2\right]}{\Big(\mathbb{E}_\vartheta\left[\partial_\vartheta\phi(\vartheta,X)\right]\Big)^2}.$$

Démonstration. Notons $Z_n(a)=\frac{1}{n}\sum_{i=1}^n\phi(a,X_i)$, $a\in\Theta$ comme dans la preuve de la Proposition 4.4, et introduisons les notations $Z_n'(a)=\partial_a Z_n(a)$, $Z_n''(a)=\partial_a^2 Z_n(a)$. Ecrivons un développement de Taylor de la fonction Z_n au voisinage de ϑ. On a

$$0=Z_n(\widehat{\vartheta}_n)=Z_n(\vartheta)+(\widehat{\vartheta}_n-\vartheta)Z_n'(\vartheta)+\frac{1}{2}(\widehat{\vartheta}_n-\vartheta)^2Z_n''(\widetilde{\vartheta}_n),$$

où $\widetilde{\vartheta}_n$ est un point (aléatoire) entre $\widehat{\vartheta}_n$ et ϑ, ce que l'on réécrit sous la forme

$$\sqrt{n}\big(\widehat{\vartheta}_n-\vartheta\big)=\frac{-\sqrt{n}Z_n(\vartheta)}{Z_n'(\vartheta)+\frac{1}{2}(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)}. \tag{4.16}$$

sur l'événement $\{Z_n'(\vartheta)+\frac{1}{2}(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)\neq 0\}$.

Sous $\mathbb{P}_\vartheta$, les variables $\phi(\vartheta,X_i)$ sont indépendantes, identiquement distribuées, de moyenne nulle et de variance finie $\mathbb{E}_\vartheta\left[\phi(\vartheta,X)^2\right]$ d'après l'Hypothèse 4.2 (iii). En appliquant le théorème central-limite

$$-\sqrt{n}Z_n(\vartheta)\xrightarrow{d}\mathcal{N}\big(0,\mathbb{E}_\vartheta\left[\phi(\vartheta,X)^2\right]\big)$$

en loi sous $\mathbb{P}_\vartheta$.

Considérons maintenant le dénominateur. On a $Z_n'(\vartheta)=\frac{1}{n}\sum_{i=1}^n\partial_\vartheta\phi(\vartheta,X_i)$ et les variables $\partial_\vartheta\phi(\vartheta,X_i)$ sont intégrables d'après l'Hypothèse 4.2 (iii). En appliquant la loi des grands nombres, on obtient

$$Z_n'(\vartheta)\xrightarrow{\mathbb{P}_\vartheta}\mathbb{E}_\vartheta\left[\partial_\vartheta\phi(\vartheta,X)\right]\neq 0.$$

La seule réelle difficulté de la preuve de la proposition consiste à démontrer que

$$\tfrac{1}{2}(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)\xrightarrow{\mathbb{P}_\vartheta}0. \tag{4.17}$$

En effet, dans ce cas, le dénominateur dans (4.16) tend vers $\mathbb{E}_\vartheta\left[\partial_\vartheta\phi(\vartheta,X)\right]\neq 0$ en $\mathbb{P}_\vartheta$ probabilité, et on en en déduit[8], en appliquant la Proposition 1.8 (Slutsky) que

$$\frac{-\sqrt{n}Z_n(\vartheta)}{Z_n'(\vartheta)+\frac{1}{2}(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)}\xrightarrow{d}\mathcal{N}\left(0,\frac{\mathbb{E}_\vartheta\left[\phi(\vartheta,X)^2\right]}{\left(\mathbb{E}_\vartheta\left[\partial_\vartheta\phi(\vartheta,X)\right]\right)^2}\right),$$

qui est la limite recherchée.

Il reste à montrer (4.17). D'après l'hypothèse 4.2 (ii), il existe un voisinage $\mathcal{V}(\vartheta)$ de ϑ tel que $|\partial_a^2\phi(a,x)|\leq g(x)$ si $a\in\mathcal{V}(\vartheta)$. L'hypothèse $\widehat{\vartheta}_n\xrightarrow{\mathbb{P}_\vartheta}\vartheta$ implique que

$$\mathbb{P}_\vartheta\left[\,\widehat{\vartheta}_n\in\mathcal{V}(\vartheta)\right]\to 1.$$

Posons $\mathcal{C}_n=\{\widehat{\vartheta}_n\in\mathcal{V}(\vartheta)\}$. On a

$$\begin{aligned}\mathbb{E}_\vartheta\left[\left|Z_n''(\widetilde{\vartheta}_n)\right|1_{\mathcal{C}_n}\right]&=\mathbb{E}_\vartheta\left[\frac{1}{n}\sum_{i=1}^n\partial_\vartheta^2\phi(\widetilde{\vartheta}_n,X_i)\right|1_{\mathcal{C}_n}\Big]\\&\leq\mathbb{E}_\vartheta\left[\frac{1}{n}\sum_{i=1}^n g(X_i)\right]\\&=\mathbb{E}_\vartheta\Big[g(X)\Big]<+\infty\end{aligned}$$

en appliquant les hypothèses 4.2 (i)–(ii). On en déduit

$$\sup_n\mathbb{E}_\vartheta\left[Z_n''(\widetilde{\vartheta}_n)1_{\mathcal{C}_n}\right]<+\infty.$$

Ceci entraîne $(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)1_{\mathcal{C}_n}\xrightarrow{\mathbb{P}_\vartheta}0$, puisque $\widehat{\vartheta}_n\xrightarrow{\mathbb{P}_\vartheta}\vartheta$, voir par exemple l'Exercice 1.1 du Chapitre 1. Finalement, on écrit, pour tout $\varepsilon>0$

$$\mathbb{P}_\vartheta\left[\left|\tfrac{1}{2}(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)\right|\geq\varepsilon\right]\leq\mathbb{P}_\vartheta\left[\left|\tfrac{1}{2}(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)1_{\mathcal{C}_n}\right|\geq\varepsilon\right]+\mathbb{P}_\vartheta\left[\mathcal{C}_n^c\right],$$

et chacun des deux termes du membre de droite tend vers 0 lorsque $n\to\infty$. □

Loi limite des M-estimateurs

Nous nous restreignons encore au cas où $\Theta\subset\mathbb{R}$. Nous traduisons l'Hypothèse 4.2 pour une fonction de contraste ψ en posant $\phi(a,x)=\partial_a\psi(a,x)$.

Hypothèse 4.3 (Hypothèse loi limite M-estimateurs)**.** *On a*

[8] il y a une petite difficulté provenant du fait que l'on doit se placer sur l'événement $\{Z_n'(\vartheta)+\frac{1}{2}(\widehat{\vartheta}_n-\vartheta)Z_n''(\widetilde{\vartheta}_n)\neq 0\}$, mais la $\mathbb{P}_\vartheta$ probabilité de cet événement est tend vers 1. Nous omettons les détails.

(i) La famille $\{\mathbb{P}_\vartheta,\ \vartheta\in\Theta\}$ *est dominée par une mesure* σ*-finie* μ *sur* $\mathbb{R}$*.*

(ii) Pour tout point $\vartheta\in\Theta$*, il existe un voisinage ouvert* $\mathcal{V}(\vartheta)$ *tel que, pour tout* $a\in\mathcal{V}(\vartheta)$

$$\left|\partial_a^3\psi(a,x)\right|\le g(x),\quad \text{où}\quad \int_{\mathbb{R}} g(x)\mu(dx)<+\infty.$$

(iii) Pour tout $\vartheta\in\Theta$*, on a*

$$\mathbb{E}_\vartheta\left[\partial_\vartheta\psi(\vartheta,X)\right]=0,\quad \mathbb{E}_\vartheta\left[\partial_\vartheta\psi(\vartheta,X)^2\right]<+\infty,\quad \mathbb{E}_\vartheta\left[\partial_\vartheta^2\psi(\vartheta,X)\right]\neq 0.$$

Proposition 4.6 (Loi limite des M-estimateurs)**.** *Si la famille* $\{\mathbb{P}_\vartheta,\ \vartheta\in\Theta\}$ *et la fonction* ϕ *vérifient l'Hypothèse 4.2, alors, si* $\widehat{\vartheta}_n$ *est un* M*-estimateur associé à* ψ *tel que* $\widehat{\vartheta}_n\xrightarrow{\mathbb{P}_\vartheta}\vartheta$*, on a*

$$\sqrt{n}\big(\widehat{\vartheta}_n-\vartheta\big)\xrightarrow{d}\mathcal{N}\big(0,v_\psi(\vartheta)\big)$$

en loi sous $\mathbb{P}_\vartheta$*, où*

$$v_\phi(\vartheta)=\frac{\mathbb{E}_\vartheta\left[\partial_\vartheta(\vartheta,X)^2\right]}{\Big(\mathbb{E}_\vartheta\left[\partial_\vartheta^2\phi(\vartheta,X)\right]\Big)^2}.$$

Démonstration. Comme indiqué plus haut, on applique la Proposition 4.5 à la fonction $\phi(a,x)=\partial_a\psi(a,x)$. □

4.4 Maximum de vraisemblance

4.4.1 Principe du maximum de vraisemblance

Fonction de vraisemblance

On se place sous l'Hypothèse de domination 4.1 présentée dans la Section 4.1.1 : l'expérience $\mathcal{E}$ est dominéee par une mesure μ sur $\mathbb{R}$, et on note

$$\{f(\vartheta,\bullet),\vartheta\in\Theta\}\tag{4.18}$$

la famille de densités par rapport à μ, indicée par l'ensemble des paramètres $\Theta\subset\mathbb{R}^d$, avec $d\ge 1$. Pour toute fonction test g

$$\int_{\mathbb{R}} g(x)\,\mathbb{P}_\vartheta(dx)=\int_{\mathbb{R}} g(x)\frac{d\,\mathbb{P}_\vartheta}{d\mu}(x)\mu(dx)=\int_{\mathbb{R}} g(x)f(\vartheta,x)\mu(dx).$$

Définition 4.6. *On appelle fonction de vraisemblance associée à l'expérience produit* $\mathcal{E}^n$ *l'application*

$$\vartheta\in\Theta\rightsquigarrow\mathcal{L}_n(\vartheta,X_1,\dots,X_n)=\prod_{i=1}^n f(\vartheta,X_i).$$

La[9] fonction de vraisemblance est une fonction aléatoire, observable. On la note parfois simplement $\mathcal{L}_n(\vartheta)$ lorsqu'il n'y a pas d'ambiguité.

Exemple 4.7 (cas discret)**.** Si la famille $\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\}$ est la famille des lois de Poisson de paramètre $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$, alors une mesure dominante est la mesure de comptage μ sur $\mathbb{N}$ et on a

$$\mathbb{P}_\vartheta(dx) = f(\vartheta, x) = e^{-\vartheta}\frac{\vartheta^x}{x!}\mu(dx).$$

La mesure $\mu(dx)$ est portée par $\mathbb{N}$, donc on peut prendre $f(\vartheta, x) = e^{-\lambda}\frac{\vartheta^x}{x!}$ pour $x \in \mathbb{N}$ et 0 sinon. La vraisemblance s'écrit alors , pour tout $\vartheta > 0$

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^n e^{-\vartheta}\frac{\vartheta^{X_i}}{X_i!} = \frac{1}{\prod_{i=1}^n X_i!}e^{-n\vartheta}\vartheta^{\sum_{i=1}^n X_i}.$$

Exemple 4.8 (cas continu)**.** Si la famille $\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\}$ est la famille des lois de Cauchy de paramètre $\vartheta = (\alpha, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$, – voir la Section 4.1.2 – alors une mesure dominante est la mesure de Lebesgue sur $\mathbb{R}$ et on a

$$\mathbb{P}_\vartheta(dx) = f(\vartheta, x) = \frac{\sigma}{\pi\big(\sigma^2 + (x-\alpha)^2\big)}dx.$$

La vraisemblance s'écrit alors, pour tout $\vartheta > 0$

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \frac{\sigma^n}{\pi^n}\prod_{i=1}^n \big(\sigma^2 + (X_i - \alpha)^2\big)^{-1}$$

Exemple 4.9 (cas mélange)**.** Dans les exemples emblématiques du Chapitre 2, nous avons mentionné l'expérience engendrée par l'observation de

$$X_i^\star = \min\{X_i, T\},\quad i = 1, \ldots, n$$

où les X_i sont des variables exponentielles indépendantes, de paramètre $\vartheta > 0$ que l'on n'observe pas, et $T > 0$ est un instant de censure. Les lois $\{\mathbb{P}_\vartheta^\star,\ \vartheta \in \Theta\}$ de $X^\star$ ne sont ni discrètes, ni continues. La famille est dominée par $\mu(dx) = dx + \delta_T(dx)$, où $\delta_T(dx)$ est la mesure de Dirac au point T. On a

$$\mathbb{P}_\vartheta^\star(dx) = p(\vartheta, x)\mu(dx),$$

où

$$f(\vartheta, x) = \vartheta e^{-\vartheta x}1_{x<T} + c(\vartheta)1_{x=T},$$

[9]La fonction $x \rightsquigarrow f(\vartheta, x)$ est définie à un ensemble μ-négligeable près, donc on devrait en toute rigueur parler d'une (classe d'équivalence de) fonction de vraisemblance.

avec $c(\vartheta) = \int_T^{+\infty} \vartheta e^{-\vartheta t} dt = e^{-\vartheta T}$. La vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}_n(\vartheta, X_1^\star, \ldots, X_n^\star) &= \prod_{i=1}^n f(\vartheta, X_i^\star) \\ &= \vartheta^{\operatorname{card} N_n^-} \exp\big(-\vartheta \sum_{i \in N_n^-} X_i^\star\big) c(\vartheta)^{\operatorname{card} N_n^+},\end{aligned}$$

où $N_n^- = \{i \leq n,\, X_i^\star < T\}$ et $N_n^+ = \{i \leq n,\, X_i^\star = T\}$. Elle est à comparer avec la vraisemblance du modèle sans censure, où l'on observe les X_i directement. Dans ce cas

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \vartheta^n \exp\big(-\vartheta \sum_{i=1}^n X_i\big).$$

Nous verrons au Chapitre 6 comment quantifier la perte d'information liée à la censure grâce à la vraisemblance.

Définition de l'estimateur du maximum de vraisemblance

Définition 4.7. *On appelle estimateur du maximum de vraisemblance tout estimateur* $\widehat{\vartheta}_n^{\mathtt{mv}}$ *satisfaisant*

$$\mathcal{L}_n(\widehat{\vartheta}_n^{\mathtt{mv}}, X_1, \ldots, X_n) = \max_{\vartheta \in \Theta} \mathcal{L}_n(\vartheta, X_1, \ldots, X_n),$$

autrement dit

$$\widehat{\vartheta}_n^{\mathtt{mv}} \in \arg\max_{\vartheta \in \Theta} \mathcal{L}_n(\vartheta, X_1, \ldots, X_n). \tag{4.19}$$

L'estimateur du maximum de vraisemblance peut ne pas exister. Il n'est pas non plus nécessairement unique.

Définition 4.8. *L'application*

$$\begin{aligned}\vartheta \in \Theta \rightsquigarrow \ell_n(\vartheta, X_1, \ldots, X_n) &= \frac{1}{n} \log \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) \\ &= \frac{1}{n} \sum_{i=1}^n \log f(\vartheta, X_i),\end{aligned}$$

bien définie si $f(\vartheta, \bullet) > 0$ *est appelée fonction de log-vraisemblance. En posant* $\log 0 = 0$, *on pourra parler de log-vraisemblance en toute généralité.*

On a aussi

$$\widehat{\vartheta}_n^{\mathtt{mv}} \in \arg\max_{\vartheta \in \Theta} \ell_n(\vartheta, X_1, \ldots, X_n).$$

Avant de donner des exemples de calcul effectif d'estimateurs du maximum de vraisemblance, nous allons justifier la définition (4.19)

Principe de maximum de vraisemblance à deux points

Considérons une famille de lois à deux points

$$\Theta = \{\vartheta_1, \vartheta_1\} \subset \mathbb{R},$$

où $\mathbb{P}_{\vartheta_1}$ et $\mathbb{P}_{\vartheta_2}$ sont deux lois discrètes portées par un sous-ensemble $\mathcal{M} \subset \mathbb{R}$ au plus dénombrable. On choisit pour mesure dominante μ la mesure de comptage sur $\mathcal{M}$, et la densité $f(\vartheta, \bullet)$ est donnée par

$$f(\vartheta, x) = \mathbb{P}_\vartheta \big[X = x\big], \;\; x \in \mathcal{M}, \;\; \vartheta \in \{\vartheta_1, \vartheta_2\}. \tag{4.20}$$

A priori – avant l'expérience aléatoire – si les observations $(X_1, \ldots, X_n)$ suivent la loi $\mathbb{P}_\vartheta$ (avec $\vartheta = \vartheta_1$ ou ϑ_2) la probabilité d'observer[10] $(X_1 = x_1, \ldots, X_n = x_n)$ est exactement

$$\mathbb{P}_\vartheta \left[X_1 = x_1, \ldots, X_n = x_n\right] = \prod_{i=1}^n \mathbb{P}_\vartheta \left[X_i = x_i\right] = \prod_{i=1}^n f(\vartheta, x_i).$$

A posteriori on dispose d'une réalisation de $(X_1, \ldots, X_n)$. Supposons que, pour cette réalisation, on observe

$$\Big\{ \prod_{i=1}^n f(\vartheta_1, X_i) > \prod_{i=1}^n f(\vartheta_2, X_i) \Big\},$$

c'est-à-dire

$$\Big\{ \mathcal{L}_n(\vartheta_1, X_1, \ldots, X_n) > \mathcal{L}_n(\vartheta_2, X_1, \ldots, X_n) \Big\}.$$

D'après (4.20), nous pouvons faire l'interprétation suivante :

A posteriori, la probabilité d'avoir observé $(X_1, \ldots, X_n)$ est plus grande sous $\mathbb{P}_{\vartheta_1}$ que sous $\mathbb{P}_{\vartheta_2}$. Ceci nous suggère de « suspecter » que la loi des observations est $\mathbb{P}_{\vartheta_1}$ plutôt que $\mathbb{P}_{\vartheta_2}$: la valeur ϑ_1 est « plus vraisemblable » que ϑ_2.

Si, pour la réalisation de l'observation $(X_1, \ldots, X_n)$ on a $\mathcal{L}_n(\vartheta_2) > \mathcal{L}_n(\vartheta_1)$, alors on fera la conclusion opposée : ϑ_2 est plus « vraisemblable » que ϑ_1. On a donc maximisé la fonction de vraisemblance $\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta, X_1, \ldots, X_n)$ dans le cas très simple où ϑ ne peut prendre que deux valeurs :

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} = \vartheta_1 1_{\mathcal{L}_n(\vartheta_1, X_1, \ldots, X_n) > \mathcal{L}_n(\vartheta_2, X_1, \ldots, X_n)} + \vartheta_2 1_{\mathcal{L}_n(\vartheta_1, X_1, \ldots, X_n) < \mathcal{L}_n(\vartheta_2, X_1, \ldots, X_n)}.$$

Si enfin $\mathcal{L}_n(\vartheta_2) = \mathcal{L}_n(\vartheta_1)$, alors il n'y a pas unicité de la procédure et on ne peut pas conclure.

[10] c'est-à-dire la probabilité de réalisation de l'événement $\{X_1 = x_1, \ldots, X_n = x_n\}$

Passage de deux paramètres et une famille de lois quelconque

De manière générale, si $\Theta \subset \mathbb{R}^d$ avec $d \geq 1$ est un ensemble arbitraire, la valeur, si elle est bien définie,

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} = \arg\max_{\vartheta \in \Theta} \mathcal{L}_n(\vartheta, X_1, \ldots, X_n)$$

est la plus vraisemblable.

Passage à une famille de lois continues

Le passage aux lois continues, où les $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ sont absolument continues par rapport à la mesure de Lebesgue se faite de la même manière. On peut reproduire – heuristiquement – le raisonnement du paragraphe précédent. On remplace

$$\mathbb{P}_\vartheta \left[X_1 = x_1, \ldots, X_n = x_n\right] = \prod_{i=1}^n \mathbb{P}_\vartheta \left[X_i = x_i\right] = \prod_{i=1}^n f(\vartheta, x_i),$$

par

$$\mathbb{P}_\vartheta \left[X_1 \in \mathcal{V}(x_1), \ldots, X_n \in \mathcal{V}(x_n)\right] = \prod_{i=1}^n \mathbb{P}_\vartheta \left[X_i \in \mathcal{V}(x_i)\right]$$

où $\mathcal{V}(x)$ est un « petit » voisinage de x. Alors

$$\mathbb{P}_\vartheta \left[X \in \mathcal{V}(x)\right] = \int_{\mathcal{V}(x)} f(\vartheta, u)du \approx f(\vartheta, x)\big|\mathcal{V}(x)\big|$$

dans la limite $\big|\mathcal{V}(x)\big| \to 0$, où $\big|\mathcal{V}(x)\big|$ désigne le mesure de Lebesgue de $\mathcal{V}(x)$. Donc la probabilité de l'événement

$$\Big\{X_1 \in \mathcal{V}(x_1), \ldots, X_n \in \mathcal{V}(x_n)\Big\}$$

est « essentiellement » proportionnelle à $\prod_{i=1}^n f(\vartheta, x_i)$, et ceci indépendamment de ϑ (si on accepte l'approximation précédente).

Equations de vraisemblance

Si le maximum de $\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta)$, ou encore le maximum de $\vartheta \rightsquigarrow \ell_n(\vartheta)$ n'est pas atteint sur la frontière de Θ et si l'application $\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta)$ est continûment différentiable, alors une condition nécessaire que doit satisfaire l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_n^{\,\mathtt{mv}}$ est l'annulation du gradient

$$\nabla_\vartheta \, \mathcal{L}_n(\vartheta, X_1, \ldots, X_n)\big|_{\vartheta = \widehat{\vartheta}_n^{\,\mathtt{mv}}} = 0$$

ce qui fournit un système de d équations si $\Theta \subset \mathbb{R}^d$ avec $d \geq 1$. De la même manière, une condition nécessaire sur la log-vraisemblance est

$$\nabla_\vartheta \, \ell_n(\vartheta, X_1, \ldots, X_n)|_{\vartheta=\widehat{\vartheta}_n^{\mathtt{mv}}} = 0 \tag{4.21}$$

Définition 4.9 (Equations de vraisemblance)**.** *L'équation* (4.21) *est appelée équation de vraisemblance si* $d = 1$ *et système d'équations de vraisembance si* $d > 1$*.*

En résolvant (4.21), on obtient tous les points critiques de $\vartheta \rightsquigarrow \ell_n(\vartheta)$, en particulier, tous ses maxima et minima locaux.

Définition 4.10. *On appelle racine de l'équation de vraisemblance tout (estimateur)* $\widehat{\vartheta}_n^{\mathtt{rv}}$ *solution de* (4.21)*, c'est-à-dire tel que*

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

Remarque 4.8. Supposons que pour tout $\vartheta \in \Theta$, on a $f(\vartheta, x) > 0$, $\mu(dx)$ presque-partout et $\vartheta \rightsquigarrow f(\vartheta, x)$ est différentiable, $\mu(dx)$ presque-partout. Alors, si $\vartheta \rightsquigarrow \ell_n(\vartheta)$ atteint son maximum global pour tous les ϑ tels que $\nabla_\vartheta \ell_n(\vartheta) = 0$, alors les ensembles qui définissent les solutions $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$ et $\widehat{\vartheta}_n^{\mathtt{rv}}$ coïncident.

Invariance du maximum de vraisemblance vis-à-vis de la mesure dominante

Sous l'Hypothèse 4.1, il existe une mesure positive σ-finie sur $\mathbb{R}$ qui domine la famille $\big\{ \mathbb{P}_\vartheta, \, \vartheta \in \Theta \big\}$.

C'est le choix de μ qui spécifie la famille de densités $f(\vartheta, \bullet)$ sur laquelle est construite la vraisemblance, et par suite l'estimateur du maximum de vraisemblance.

Proposition 4.7. *L'estimateur du maximum de vraisemblance ne dépend pas du choix de la mesure dominante* μ *dans le calcul de la vraisemblance.*

Démonstration. Soit ν une autre mesure dominante. Les mesures μ et ν sont elle-mêmes dominée par la mesure $\mu + \nu$, donc, pour toute fonction test φ,

$$\begin{aligned}
\int_{\mathbb{R}} g(x) \, \mathbb{P}_\vartheta(dx) &= \int_{\mathbb{R}} \varphi(x) \frac{d \, \mathbb{P}_\vartheta}{d(\mu+\nu)}(x)(\mu+\nu)(dx) \\
&= \int_{\mathbb{R}} \varphi(x) \frac{d \, \mathbb{P}_\vartheta}{d\mu}(x) \frac{d\mu}{d(\mu+\nu)}(x)(\mu+\nu)(dx) \\
&= \int_{\mathbb{R}} \varphi(x) \frac{d \, \mathbb{P}_\vartheta}{d\nu}(x) \frac{d\nu}{d(\mu+\nu)}(x)(\mu+\nu)(dx).
\end{aligned}$$

Les densités $\frac{d \, \mathbb{P}_\vartheta}{d\mu}(x)$ et $\frac{d \, \mathbb{P}_\vartheta}{d\nu}(x)$ ne différent que d'un facteur multiplicatif qui ne dépend pas de ϑ (sauf éventuellement sur un ensemble $(\mu+\nu)$-négligeable). Donc, presque-sûrement,

$$\prod_{i=1}^n \frac{d \, \mathbb{P}_\vartheta}{d\mu}(X_i) \quad \text{et} \quad \prod_{i=1}^n \frac{d \, \mathbb{P}_\vartheta}{d\nu}(X_i)$$

ne diffèrent que d'une fonction de $X_1, \ldots, X_n$ qui ne dépend pas de ϑ. On ne modifie pas $\widehat{\vartheta}_n^{\mathtt{mv}}$ selon que l'on maximise la vraisembance formée sur l'une ou l'autre des mesures dominantes. □

Equi-invariance

L'estimateur du maximum de vraisemblance reste inchangé si l'on change de (bonne) paramétrisation. Si $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ est une famille de probabilités associée à une expérience statistique, et si

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

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

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

Proposition 4.8. *Si $G : \Theta \to \widetilde{\Theta}$ est une bijection et si $\widehat{\vartheta}_n^{\mathtt{mv}}$ est l'estimateur du maximum de vraisemblance pour la famille $\big(\mathbb{P}_\vartheta, \vartheta \in \Theta\big)$, alors $G(\widehat{\vartheta}_n^{\mathtt{mv}})$ est l'estimateur du maximum de vraisemblance pour le paramètre $G(\vartheta)$, c'est-à-dire pour la famille $\{\mathbb{P}_{G^{-1}(\tau)}, \tau \in G(\Theta)\} = \{\mathbb{Q}_\tau, \tau \in G(\Theta)\}$.*

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

$$\begin{aligned}
\widetilde{\mathcal{L}}_n(\tau, X_1, \ldots, X_n) &= \mathcal{L}_n(G^{-1}(\tau), X_1, \ldots, X_n) \\
&= \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) \\
&\leq \mathcal{L}_n(\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}, X_1, \ldots, X_n) \\
&= \widetilde{\mathcal{L}}_n(\widehat{\tau}_n, X_1, \ldots, X_n).
\end{aligned}$$

□

Exemple 4.10. Si $X_1, \ldots, X_n$ est un n-échantillon de loi exponentielle de paramètre $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$, alors la loi $\mathbb{P}_\vartheta$ a une densité par rapport à la mesure de Lebesgue donnée par $f(\vartheta, x) = \vartheta e^{-\vartheta x} 1_{\{x \geq 0\}}$. la log-vraisemblance s'écrit[11]

$$\ell_n(\vartheta, X_1, \ldots, X_n) = n \log \vartheta - \vartheta \sum_{i=1}^{n} X_i,$$

donc $\partial_\vartheta \ell_n(\vartheta, X_1, \ldots, X_n) = 0$ si et seulement si $\vartheta = \frac{1}{\overline{X}_n}$. On vérifie que c'est un maximum global, donc $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} = \frac{1}{\overline{X}_n}$. Par équi-invariance, on en déduit sans calcul que l'estimateur du maximum de vraisemblance pour un n-échantillon de loi exponentielle de paramètre $\tau = 1/\vartheta$, $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ est $\widehat{\tau}_n = \overline{X}_n$.

[11] Noter que tous les X_i sont positifs $\mathbb{P}_\vartheta$ p.s., simultanément pour tous les $\vartheta \in \Theta$, donc il est inutile de faire apparaître la condition $1_{X_i \geq 0}$ dans la formule de la vraisemblance.

Exemple 4.11. Si $X_1, \dots, X_n$ est un n-échantillon de loi log-normale de moyenne $a \in \mathbb{R}$ et de variance $d^2 > 0$, alors, par la représentation $Y_i = \log X_i \sim \mathcal{N}(\mu, \sigma^2)$ avec

$$a = e^{\mu+\sigma^2/2}, \quad d^2 = a^2(e^{\sigma^2} - 1)$$

(voir Section 4.1.2), en étudiant la fonction

$$(\mu, \sigma^2) \rightsquigarrow (a, d^2) = \big(e^{\mu+\sigma^2/2}, a^2(e^{\sigma^2} - 1)\big)$$

qui établit une bijection de $\mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$, on en déduit par équi-invariance du cas gaussien que l'estimateur du maximum de vraisamblance pour (a, d^2) est

$$\big(\widehat{a}_n^{\,\mathtt{mv}}, (\widehat{d}_n^{\,2})^{\mathtt{mv}}\big) = \big(e^{\overline{Y}_n + s_n^2/2}, (\widehat{a}_n^{\,\mathtt{mv}})^2(e^{s_n^2} - 1)\big),$$

où $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \log X_i$ et $s_n^2 = \frac{1}{n}(Y_i - \overline{Y}_n)^2$.

4.4.2 Exemples de calcul

Exemple 4.12 (modèle gaussien standard)**.** L'expérience statistique est engendrée par un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, le paramètre est $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$. Une mesure dominante est la mesure de Lebesgue sur $\mathbb{R}$ et on a alors

$$f(\vartheta, x) = (2\pi\sigma^2)^{-1/2} \exp\big(-\tfrac{1}{2\sigma^2}(x - \mu)^2\big)$$

La log-vraisemblance associée s'écrit

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

L'équation de vraisemblance s'écrit

$$\begin{cases} \partial_\mu \ell_n\big((\mu, \sigma^2), X_1, \dots, X_n\big) &= \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n (X_i - \mu) \\[2ex] \partial_{\sigma^2} \ell_n\big((\mu, \sigma^2), X_1, \dots, X_n\big) &= -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n (X_i - \mu)^2, \end{cases}$$

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

$$\widehat{\vartheta}_n = \big(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2\big).$$

On vérifie ensuite que le point critique est l'unique maximum global et donc $\widehat{\vartheta}_n^{\mathtt{rv}} = \widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$.

Exemple 4.13 (modèle de Bernoulli)**.** L'expérience statistique est engendrée par un n-échantillon de loi de Bernoulli de paramètre $\vartheta \in \Theta = (0,1)$. Donc

$$\mathbb{P}_\vartheta \big[X = x\big] = \vartheta^x (1-\vartheta)^{1-x}, \quad x \in \{0,1\}.$$

On peut prendre comme mesure dominante μ la mesure de comptage sur $\{0,1\}$ et dans ce cas $f(\vartheta, x) = \vartheta^x (1-\vartheta)^{1-x}$. La vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) &= \prod_{i=1}^n \vartheta^{X_i}(1-\vartheta)^{1-X_i} \\ &= \vartheta^{\sum_{i=1}^n X_i}(1-\vartheta)^{n-\sum_{i=1}^n X_i}\end{aligned}$$

et la log-vraisemblance vaut

$$\ell_n(\vartheta, X_1, \ldots, X_n) = n\,\overline{X}_n \log\vartheta + n(1-\overline{X}_n)\log(1-\vartheta).$$

On a $\partial_\vartheta \ell_n(\vartheta, X_1, \ldots, X_n) = n\,\overline{X}_n\,\vartheta^{-1} - (n-\overline{X}_n)(1-\vartheta)^{-1} = 0$ si et seulement si $\vartheta = \overline{X}_n$. On vérifie que $\vartheta = \overline{X}_n$ est un maximum global et donc $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} = \overline{X}_n$.

Exemple 4.14 (modèle de Laplace)**.** L'expérience statistique est engendrée par un n-échantillon de loi de Laplace de paramètre $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$, dont la densité par rapport à la mesure de Lebesgue est donnée par

$$f(\vartheta, x) = \frac{1}{2\sigma} \exp\big(-\frac{|x-\vartheta|}{\sigma}\big),$$

où $\sigma > 0$ est connu. La fonction de vraisemblance s'écrit

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = (2\sigma)^{-n} \exp\big(-\frac{1}{\sigma}\sum_{i=1}^n \big|X_i - \vartheta\big|\big)$$

et la log-vraisembance vaut

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

Maximiser $\mathcal{L}_n(\vartheta, X_1, \ldots, X_n)$ revient à minimiser la fonction $\vartheta \rightsquigarrow \sum_{i=1}^n \big|X_i - \vartheta\big|$. Cette fonction est dérivable presque partout, de dérivée

$$-\sum_{i=1}^n \mathrm{sign}(X_i - \vartheta).$$

La dérivée (définie presque partout) est constante par morceaux. Si n est impair, elle s'annule en un point unique $X_{\left(\frac{n+1}{2}\right)}$, où $X_{(1)} \leq \ldots \leq X_{(n)}$ désigne la statistique d'ordre associée à l'échantillon (voir Section 3.4.2 du Chapitre 3). Si n est pair, il y a une infinité de solution : tout point de l'intervalle $\big(X_{\left(\frac{n}{2}\right)}, X_{\left(\frac{n}{2}+1\right)}\big)$ est un estimateur du maximum de vraisemblance. On retrouve la médiane empirique (voir Section 3.4.2 du Chapitre 3).

Exemple 4.15 (modèle uniforme)**.** L'expérience statistique est engendrée par un n-échantillon de loi de uniforme sur $[0, \vartheta]$, où $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ est le paramètre. Une mesure dominante est la mesure de Lebesgue et la densité de la loi uniforme est donnée par

$$f(\vartheta, x) = \frac{1}{\vartheta} 1_{[0,\vartheta]}(x).$$

La fonction de vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) &= \frac{1}{\vartheta^n} \prod_{i=1}^n 1_{0 \le X_i \le \vartheta} \\ &= \vartheta^{-n} 1_{X_{(n)} \ge \vartheta},\end{aligned}$$

où $X_{(n)} = \max_{i=1,\ldots,n} X_i$. La valeur maximale de $\mathcal{L}_n(\vartheta, X_1, \ldots, X_n)$ est obtenue pour $\vartheta = X_{(n)}$ et donc $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} = X_{(n)}$. Par contre, la fonction de log-vraisemblance n'est pas définie pour toutes les valeurs de ϑ et n'est pas dérivable.

Exemple 4.16 (modèle de Cauchy)**.** L'expérience statistique est engendrée par un n-échantillon de loi de Cauchy de paramètre $\vartheta \in \Theta = \mathbb{R}$, dont la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ est donnée par

$$f(\vartheta, x) = \frac{1}{\pi\big(1 + (x - \vartheta)^2\big)}.$$

La fonction de vraisemblance s'écrit

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

et la log-vraisemblance vaut

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

et l'équation de vraisemblance équivaut à résoudre

$$\sum_{i=1}^n \frac{X_i - \vartheta}{1 + (X_i - \vartheta)^2} = 0. \tag{4.22}$$

Cette équation n'admet pas de solution explicite et admet en général plusieurs solutions. Nous verrons plus tard comment traiter le comportement asymptotique d'une solution de (4.22) de façon indirecte.

Exemple 4.17 (absence d'estimateur du maximum de vraisemblance)**.** Considérons le modèle de translation par rapport à la densité

$$f_0(x) = \frac{e^{-\frac{|x|}{2}}}{2\sqrt{2\pi|x|}}, \quad x \in \mathbb{R},$$

c'est-à-dire le modèle dominé par rapport à la mesure de Lebesgue sur $\mathbb{R}$ de densités

$$f_0(x-\vartheta),\ \ x\in\mathbb{R}, \vartheta\in\Theta=\mathbb{R}\,.$$

La fonction de vraisemblance s'écrit

$$\mathcal{L}_n(\vartheta, X_1,\dots,X_n)=\prod_{i=1}^n f_0(X_i-\vartheta).$$

On a $\lim_{\vartheta\to X_i}\mathcal{L}_n(\vartheta,X_1,\dots,X_n)=+\infty$ pour tout $i=1,\dots,n$. Pour cette expérience statistique, il n'existe pas d'estimateur du maximum de vraisemblance.

4.4.3 Maximum de vraisemblance et M-estimation

Préliminaire : une inégalité de convexité

Lemme 4.4.1 (Inégalite d'entropie)**.** *Soit μ une mesure σ-finie sur $(\mathbb{R},\mathcal{B})$. Soient deux deux densités de probabilité $f,g:\mathbb{R}\to\mathbb{R}_+$ par rapport à μ, c'est-à-dire vérifiant*

$$\int_{\mathbb{R}} f(x)\mu(dx)=\int_{\mathbb{R}} g(x)\mu(dx)=1.$$

Alors[12]

$$\int_{\mathbb{R}} f(x)\log f(x)\mu(dx)\geq\int_{\mathbb{R}} f(x)\log g(x)\mu(dx)$$

si les deux intégrales sont finies, et l'égalité a lieu si et seulement si $f=g$ μ-presque partout.

Démonstration. On doit montrer

$$\int_{\mathbb{R}} f(x)\log\frac{g(x)}{f(x)}\mu(dx)\leq 0. \tag{4.23}$$

Pour $x>-1$, on a $\log(1+x)\leq x$ avec égalité si et seulement si $x=0$, donc

$$\log\frac{g(x)}{f(x)}=\log\left(1+(\frac{g(x)}{f(x)}-1)\right)\leq\frac{g(x)}{f(x)}-1,$$

avec égalité si et seulement si $f(x)=g(x)$. Il vient

$$\begin{aligned}\int_{\mathbb{R}} f(x)\log\frac{g(x)}{f(x)}\mu(dx)&\leq\int_{\mathbb{R}} f(x)\left(\frac{g(x)}{f(x)}-1\right)\mu(dx)\\&=\int_{\mathbb{R}} g(x)\mu(dx)-\int_{\mathbb{R}} f(x)\mu(dx)=0.\end{aligned}$$

Si on n'a pas $f=g$ μ-presque partout, alors l'inégalité est stricte. □

[12] Avec la convention $\int_{\{x,f(x)=0\}} f(x)\log g(x)\mu(dx)=0$ pour toute fonction g.

Le maximum de vraisemblance est un M-estimateur

Replaçons nous dans le contexte de la Section 4.3.2. Posons

$$\psi(a,x) = \log f(a,x), \quad a \in \Theta,\ x \in \mathbb{R}\,.$$

Alors l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_n^{\mathtt{mv}}$, s'il existe, satisfait

$$\widehat{\vartheta}_n^{\mathtt{mv}} \in \arg\max_{a\in\Theta} \sum_{i=1}^{n} \psi(a, X_i)$$

et peut s'interpréter comme le M-estimateur associé à la fonction ψ. En effet d'après le Lemme 4.4.1, la valeur $a = \vartheta$ maximise

$$a \rightsquigarrow \int_{\mathbb{R}} \psi(a,x)\, \mathbb{P}_\vartheta(dx) = \int_{\mathbb{R}} \log f(a,x) f(\vartheta,x) \mu(dx).$$

Ceci justifie *a posteriori* le principe du maximum de vraisemblance. Nous verrons au Chapitre 6 qu'il y a beaucoup plus encore : le contraste $\psi(a,x) = \log f(\vartheta,x)$ est optimal dans un certain sens. Si pour tout $\vartheta \in \Theta$ la fonction $\vartheta \rightsquigarrow \log f(\vartheta,x)$ est différentiable μ presque-partout, alors on a aussi l'interprétation du maximum de vraisemblance comme Z-estimateur associé à la fonction

$$\phi(\vartheta,x) = \partial_\vartheta \log f(\vartheta,x) = \frac{\partial_\vartheta f(\vartheta,x)}{f(\vartheta,x)}, \quad \vartheta \in \Theta, x \in \mathbb{R}$$

lorsque $\Theta \subset \mathbb{R}$, avec une généralisation immédiate en dimension plus grande que 1.

En particulier, le comportement asymptotique de l'estimateur du maximum de vraisemblance peut se déduire des Propositions 4.5 ou 4.6 si l'on dispose de conditions de régularité suffisantes. Nous reviendrons plus spécifiquement sur la convergence de l'estimateur du maximum de vraisemblance dans le Chapitre 6.

Chapitre 5

Méthodes d'estimation pour le modèle de régression

5.1 Modèles de régression

Déjà rencontrée dans les exemples 2, 4 et 6 du Chapitre 2, la régression, comme l'échantillonnage, est un modèle incontournable en statistique. Presque tous les modèles utilisés dans les applications peuvent se ramener à des généralisations plus ou moins sophistiquées du modèle de régression. Dans ce chapitre, nous présentons brièvement les résultats essentiels de l'estimation paramétrique et en particulier, la méthode des moindres carrés.

5.1.1 Modèle de régression à « design » aléatoire

On part de l'expérience statistiques engendrée par l'observation

$$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$$

où

$$Y_i = r(\vartheta, \boldsymbol{X}_i) + \xi_i, \tag{5.1}$$

pour $i = 1, \ldots, n$. Les variables aléatoires $(\boldsymbol{X}_i, Y_i)$ sont indépendantes, de même loi, à valeurs dans $\mathbb{R}^k \times \mathbb{R}$, et $\vartheta \in \Theta \subset \mathbb{R}^d$ est le paramètre inconnu.

Définition 5.1. *Le vecteur $\boldsymbol{X}_i$ est appelé vecteur de covariables (ou de variables explicatives[1]) associé à l'observation Y_i. La matrice $\begin{pmatrix} \boldsymbol{X}_1 \cdots \boldsymbol{X}_n \end{pmatrix}$ est appelée « design » ou plan d'expérience associé au modèle.*

[1] l'emploi de termes différents – et non synonymes – pour désigner les même objets provient des utilisations très différentes du modèle de régression dans les applications (économétrie, signal, biostatistique, etc.).

La fonction $x \rightsquigarrow r(\vartheta, x)$, connue au paramètre $\vartheta \in \Theta$ près, est appelée fonction de régression.

Les variables aléatoires ξ_i sont appelées « bruits » ou innovations.

On note $\mathbb{P}_\vartheta = \mathbb{P}_\vartheta(d\mathbf{x}\, dy)$ la loi jointe des $(\boldsymbol{X}_i, Y_i)$ définie sur $\mathbb{R}^k \times \mathbb{R}$ et le but est d'inférer sur le paramètre ϑ. L'expérience statistique associée à l'observation s'écrit :

$$\mathcal{E}^n_{\texttt{design-aléa}} = \Big(\mathbb{R}^{(k+1)n}, \mathcal{B}^{(k+1)n}, \big\{\, \mathbb{P}^n_\vartheta, \vartheta \in \Theta\big\}\Big)$$

où $\mathbb{P}^n_\vartheta$ désigne le produit des lois $\mathbb{P}_\vartheta$ effectué n-fois. Notons que puisque les $(\boldsymbol{X}_i, Y_i)$ sont indépendantes et équidistribuées, les ξ_i le sont aussi.

Remarque 5.1. Les variables ξ_i « polluent » l'observation de la fonction d'intérêt $r(\vartheta, \bullet)$ aux points $(\boldsymbol{X}_i, Y_i)$. En l'absence des ξ_i reconstruire $r(\vartheta, \bullet)$ et donc ϑ se ramènerait à un problème d'interpolation numérique.

Hypothèse 5.1 (Identifiabilité, « design aléatoire »)**.** *L'application $\vartheta \in \Theta \rightsquigarrow r(\vartheta, \bullet)$ est injective. De plus, la loi des ξ_i admet un moment d'ordre* 1 *et les variables ξ_i vérifient*

$$\mathbb{E}_\vartheta\left[\xi_i \,|\, \boldsymbol{X}_i\right] = 0. \tag{5.2}$$

Remarque 5.2. L'Hypothèse 5.1 garantit une bonne paramétrisation de la fonction de régression $r(\vartheta, \bullet)$. Sans (5.2), on pourrait écrire

$$Y_i = r(\vartheta, \boldsymbol{X}_i) + g(\vartheta, \boldsymbol{X}_i) + \widetilde{\xi}_i,$$

avec $g(\vartheta, \boldsymbol{X}_i) = \mathbb{E}_\vartheta\left[\xi_i \,|\, \boldsymbol{X}_i\right]$ et $\widetilde{\xi}_i = \xi_i - \mathbb{E}_\vartheta\left[\xi_i \,|\, \boldsymbol{X}_i\right]$ qui vérifie bien $\mathbb{E}\left[\widetilde{\xi}_i \,|\, \boldsymbol{X}_i\right] = 0$ et $g \neq 0$, ce qui empêche de pouvoir identifier la fonction $r(\vartheta, \bullet)$, même lorsqu'elle est réduite à une constante.

Remarque 5.3. Une manière naturelle d'obtenir la représentation (5.1) si la loi des Y_i admet un moment d'ordre 1 est de définir, pour chaque $\vartheta \in \Theta$, la fonction de régression $r(\vartheta, \bullet) : \mathbb{R}^k \to \mathbb{R}$ en posant

$$r(\vartheta, \boldsymbol{x}) = \mathbb{E}_\vartheta\left[Y_i \,|\, \boldsymbol{X}_i = \boldsymbol{x}\right].$$

Alors, on a

$$Y_i = r(\vartheta, \boldsymbol{X}_i) + \xi_i, \quad \text{avec } \xi_i = Y_i - \mathbb{E}_\vartheta\left[Y_i \,|\, \boldsymbol{X}_i\right]$$

et on vérifie immédiatement que l'on a bien l'Hypothèse 5.2.

5.1.2 Réduction au cas d'un « design » déterministe

Nous avons déjà discuté du caractère aléatoire du « design », selon que le statisticien choisit ou non le plan d'expérience. Nous allons faire une hypothèse qui va permettre de se ramener systématiquement au cas où le « design » est déterministe.

Hypothèse 5.2 (Ancillarité des covariables)**.** *La loi $\mathbb{P}^{\boldsymbol{X}}$ des covariables ne dépend par de ϑ.*

Sous l'Hypothèse 5.2, la loi des covariables $\boldsymbol{X}_i$ ne contient pas d'information sur le paramètre ϑ. On « gèle » les $\boldsymbol{X}_i$ dont le caractère aléatoire est ignoré.

Mathématiquement, cela consiste à étudier les propriétés statistiques des estimateurs conditionnellement aux $\boldsymbol{X}_i$, et donc, de remplacer formellement les $(\boldsymbol{X}_i, Y_i)$ par $(\boldsymbol{x}_i, Y_i)$ où les $\boldsymbol{x}_i$ sont donnés, sans perdre de généralité.

On remplace désormais le modèle de régression à « design aléatoire » de la Section 5.1.1 par le modèle de régression à « design déterministe » : on observe l'expérience engendrée par

$$(\boldsymbol{x}_1, Y_1), \dots, (\boldsymbol{x}_n, Y_n),$$

où

$$Y_i = r(\vartheta, \boldsymbol{x}_i) + \xi_i \tag{5.3}$$

pour $i = 1, \dots, n$. Les vecteurs $\boldsymbol{x}_i \in \mathbb{R}^k$ sont donnés, et les variables Y_i sont indépendantes – mais pas identiquement distribuées, la loi de Y_i dépend maintenant de $\boldsymbol{x}_i$ qui est fixé – et les ξ_i sont des bruits. L'expérience statistique s'écrit ici

$$\mathcal{E}^n_{\texttt{design-déter}} = \Big(\mathbb{R}^n, \mathcal{B}^n, \big\{ \mathbb{P}^n_\vartheta, \vartheta \in \Theta \big\} \Big),$$

où $\mathbb{P}^n_\vartheta$ est la loi des Y_i données par (5.3). L'hypothèse d'identifiabilité devient

Hypothèse 5.3 (Identifiabilité, « design déterministe »)**.** *L'application $\vartheta \in \Theta \rightsquigarrow r(\vartheta, \bullet)$ est injective. De plus, pour tout $i = 1, \dots, n$, les variables aléatoires ξ sont intégrables et on a*

$$\mathbb{E}^n_\vartheta \left[\xi_i \right] = 0.$$

5.1.3 Calcul de la vraisemblance

On se place une bonne fois pour toute dans le modèle de régression à « design » déterministe, c'est-à-dire nous considérons l'expérience $\mathcal{E}^n_{\texttt{design-déter}}$.

Calcul de la loi de Y_i

Nous faisons ici une hypothèse technique :

Hypothèse 5.4. *Les « bruits » ξ_i sont indépendants, identiquement distribués, et leur loi commune $\mathbb{P}^\xi$ ne dépend pas des $\boldsymbol{x}_i$ et du paramètre ϑ.*

Cette hypothèse est un peu superflue et nous nous en affranchirons dans certains exemples. Elle a l'avantage de présenter des formules de calcul très simples.

Proposition 5.1 (Loi des observations)**.** *Sous les Hypothèses 5.3 et 5.4, on a, pour toute fonction test φ, et pour $i = 1, \dots, n$*

$$\mathbb{E}_\vartheta\left[\varphi(Y_i)\right] = \int_{\mathbb{R}} \varphi\big(z + r(\vartheta, \boldsymbol{x}_i)\big)\, \mathbb{P}^\xi(dz).$$

Si, de plus, la loi $\mathbb{P}^\xi$ des « bruits » admet une densité $z \rightsquigarrow g(z)$ par rapport à la mesure de Lebesgue, on a, pour $i = 1, \dots, n$

$$\mathbb{E}_\vartheta\left[\varphi(Y_i)\right] = \int_{\mathbb{R}} \varphi(z) g\big(z - r(\vartheta, \boldsymbol{x}_i)\big) dz.$$

En particulier, Y_i admet une densité donnée par $z \rightsquigarrow g\big(z - r(\vartheta, \boldsymbol{x}_i)\big)$.

Démonstration. Les deux points de la proposition sont évidents : on a

$$\begin{aligned}\mathbb{E}_\vartheta\left[\varphi(Y_i)\right] &= \mathbb{E}_\vartheta\left[\varphi\big(r(\vartheta, \boldsymbol{x}_i) + \xi_i\big)\right] \\ &= \int_{\mathbb{R}} \varphi\big(z + r(\vartheta, \boldsymbol{x}_i)\big)\, \mathbb{P}^\xi(dz),\end{aligned}$$

en appliquant la formule de la mesure image (1.1). Si, de plus, $\mathbb{P}^\xi$ admet une densité g, cette dernière quantité s'écrit

$$\int_{\mathbb{R}} \varphi\big(z + r(\vartheta, \boldsymbol{x}_i)\big) g(z) dz = \int_{\mathbb{R}} \varphi(z) g\big(z - r(\vartheta, \boldsymbol{x}_i)\big) dz.$$

□

Remarque 5.4. L'Hypothèse 5.4 est superflue. Dans le cas général, si on note $\mathbb{P}^\xi_{\vartheta, \boldsymbol{x}_i}$ la loi de ξ, dépendante de $\boldsymbol{x}_i$ et ϑ, et si cette loi admet une densité $z \rightsquigarrow g(\vartheta, \boldsymbol{x}_i, z)$ par rapport à la mesure de Lebesgue, alors Y_i aussi et sa densité est donnée par :

$$z \rightsquigarrow g\big(\vartheta, \boldsymbol{x}_i, z - r(\vartheta, \boldsymbol{x}_i)\big)$$

Formule de vraisemblance

Les variables Y_i étant indépendantes le calcul de leur loi jointe est immédiat.

Proposition 5.2. *Sous les Hypothèses 5.3, et 5.4, si la loi $\mathbb{P}^\xi$ des « bruits » admet une densité $z \rightsquigarrow g(z)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, alors la loi de $(Y_1, \dots, Y_n)$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ donnée par*

$$(z_1, \dots, z_n) \rightsquigarrow \prod_{i=1}^n g\big(z_i - r(\vartheta, \boldsymbol{x}_i)\big).$$

Démonstration. Par construction, les variables aléatoires $Y_1, \dots, Y_n$ sont indépendantes. On applique alors la Proposition 5.2. □

On en déduit que si $\mathbb{P}^\xi$ admet une densité par rapport à la mesure de Lebesgue, alors l'expérience statistique $\mathcal{E}^n_{\texttt{design-déter}}$ est dominée par la mesure de Lebesgue $dz_1 \cdots dz_n$ sur $\mathbb{R}_n$, et on a

$$\frac{d\,\mathbb{P}^n_\vartheta}{dz_1 \cdots dz_n} = \prod_{i=1}^n g\big(z_i - r(\vartheta, \boldsymbol{x}_i)\big).$$

Corollaire 5.1 (formule de vraisemblance)**.** *Sous les Hypothèses 5.3, et 5.4, si la loi $\mathbb{P}^\xi$ des « bruits » admet une densité $z \rightsquigarrow g(z)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, alors la vraisemblance par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ est donnée par*

$$\mathcal{L}_n\big(\vartheta, Y_1, \dots, Y_n\big) = \prod_{i=1}^n g\big(Y_i - r(\vartheta, \boldsymbol{x}_i)\big).$$

5.2 Régression linéaire simple

Pour les raisons invoquées plus haut, on se place désormais dans le modèle de régression à « design » déterministe.

5.2.1 Droite de régression

Définition 5.2. *On appelle modèle linéaire simple l'expérience statistique engendrée par les variables aléatoires Y_i à valeurs dans $\mathbb{R}$ (et par le « design » $(x_1, \dots, x_n)$), où*

$$Y_i = \vartheta_0 + \vartheta_1 x_i + \xi_i, \quad i = 1, \dots, n$$

et

- *Le paramètre inconnu est $\vartheta = (\vartheta_0, \vartheta_1)^T \in \Theta = \mathbb{R}^2$.*
- *Les « bruits » ξ_i satisfont*

$$\mathbb{E}_\vartheta\left[\xi_i\right] = 0, \quad \mathrm{Var}_\vartheta\left[\xi_i^2\right] = \sigma^2 > 0.$$

Dans ce contexte, l'Hypothèse 5.3 est automatiquement vérifiée. La variance σ^2 des « bruits » peut elle-même être inconnue et être considérée comme un paramètre du modèle. On parle de modèle de régression simple à variance connue ou inconnue. Les paramètres ϑ_0 et ϑ_1 s'appellent respectivement « ordonnée à l'origine » et « coefficient directeur » de la droite d'équation

$$y = r(\vartheta, x) = \vartheta_0 + \vartheta_1 x.$$

Si $\widehat{\vartheta}_n$ est un estimateur de ϑ, on note $x \rightsquigarrow r(\widehat{\vartheta}_n, x)$ l'estimateur de la fonction de régression (ici, une droite) associée au modèle linéaire simple.

Définition 5.3. *Si $\widehat{\vartheta}_n$ est un estimateur de ϑ dans le modèle linéaire simple, on appelle $\widehat{Y}_i = r(\widehat{\vartheta}_n, x_i)$ la valeur de Y_i prédite par l'estimateur et $\widehat{\xi}_i = Y_i - \widehat{Y}_i$ son résidu. On appelle*

$$\|\widehat{\xi}\|^2 = \sum_{i=1}^{n} \widehat{\xi}_i^2 = \sum_{i=1}^{n} (\widehat{Y}_i - Y_i)^2$$

la somme résiduelle des carrés (RSS, Residual Sum of Squares)

La somme résiduelle des carrés mesure l'erreur (au sens de la norme euclidienne) entre les observations Y_i et les observations prédites par l'estimateur $r(\widehat{\vartheta}_n, x_i)$.

Définition 5.4. *L'estimateur des moindres carrés dans le modèle linéaire simple (à variance connue) est l'estimateur $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mc}}$ qui minimise la somme résiduelle des carrés :*

$$\sum_{i=1}^{n} \big(Y_i - r(\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mc}}, x_i)\big)^2 = \min_{\vartheta \in \mathbb{R}^2} \sum_{i=1}^{n} \big(Y_i - r(\vartheta, x_i)\big)^2,$$

où l'infimum est pris sur l'ensemble des estimateurs possibles de ϑ construits à partir des observations Y_i, $i = 1, \ldots, n$.

Proposition 5.3. *On a $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mc}} = \big(\widehat{\vartheta}_{n,0}^{\mathtt{mc}}, \widehat{\vartheta}_{n,1}^{\mathtt{mc}}\big)^T$, avec*

$$\widehat{\vartheta}_{n,0}^{\mathtt{mc}} = \overline{Y}_n - \widehat{\vartheta}_{n,1}^{\mathtt{mc}} \overline{x}_n,$$

et

$$\begin{aligned} \widehat{\vartheta}_{n,1}^{\mathtt{mc}} &= \frac{\sum_{i=1}^n (x_i - \overline{x}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} \\ &= \frac{\sum_{i=1}^n x_i (Y_i - \overline{Y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} = \frac{\sum_{i=1}^n (x_i - \overline{x}_n) Y_i}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}, \end{aligned}$$

où $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.

Démonstration. En anticipant, on peut appliquer la Proposition 5.6 ou bien retrouver directement le résultat : on cherche les points critiques de la fonction

$$(\vartheta_0, \vartheta_1) \rightsquigarrow L_n(\vartheta_0, \vartheta_1) = \sum_{i=1}^{n} \big(Y_i - \vartheta_0 - \vartheta_1 x_i\big)^2.$$

On a

$$\left\{ \begin{array}{rcl} \partial_{\vartheta_0} L_n(\vartheta_0, \vartheta_1) &=& -2\sum_{i=1}^n (Y_i - \vartheta_0 - \vartheta_1 x_i) \\ \partial_{\vartheta_1} L_n(\vartheta_0, \vartheta_1) &=& -2\sum_{i=1}^n x_i (Y_i - \vartheta_0 - \vartheta_1 x_i), \end{array} \right.$$

et donc $\nabla L_n(\vartheta_0, \vartheta_1) = 0$ si et seulement si

$$\left\{ \begin{array}{rcl} -\sum_{i=1}^n Y_i + n\vartheta_0 + \vartheta_1 \sum_{i=1}^n x_i &=& 0 \\ -\sum_{i=1}^n x_i Y_i + \vartheta_0 \sum_{i=1}^n x_i + \vartheta_1 \sum_{i=1}^n x_i^2 &=& 0, \end{array} \right.$$

ce qui fournit $\vartheta_0 = \overline{Y}_n - \vartheta_1 \overline{x}_n$ en isolant ϑ_0, puis $(\vartheta_0, \vartheta_1) = \big(\widehat{\vartheta}^{\mathtt{mc}}_{n,0}, \widehat{\vartheta}^{\mathtt{mc}}_{n,1}\big)$ par substitution. La fonction L_n est quadratique et tend vers $+\infty$ en l'infini, l'unique point critique est bien un minimum global.

□

Cette preuve élémentaire s'affranchit d'hypothèses probabilistes sur le modèle : le résultat de la Propostion 5.3 ne nécessite aucune propriété sur les ξ_i. L'estimation de σ^2 est en revanche plus subtile. On peut penser à prendre la moyenne empirique du carré des résidus

$$\widehat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{\xi}^2_n = \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - r(\widehat{\vartheta}^{\mathtt{mc}}_n, x_i)\big)^2,$$

mais les variables aléatoires $\widehat{\xi}^2_n$ ne sont pas indépendantes, puisque $\widehat{\vartheta}^{\mathtt{mc}}_{\mathrm{n}}$ fait intervenir toutes les variables Y_i.

Le résultat suivant donne le comportement de la moyenne et de la variance de $\widehat{\vartheta}^{\mathtt{mc}}_n$.

Proposition 5.4. *Dans le modèle de régression linéaire simple, l'estimateur des moindres carrés $\widehat{\vartheta}^{\mathtt{mc}}_n$ vérifie*

$$\mathbb{E}_\vartheta\big[\widehat{\vartheta}^{\mathtt{mc}}_n\big] = \big(\vartheta_0, \vartheta_1\big)^T,$$

et la matrice de variance-covariance de $\widehat{\vartheta}^{\mathtt{mc}}_{\mathrm{n}}$ est donnée par

$$\Sigma\big[\,\widehat{\vartheta}^{\mathtt{mc}}_{\mathrm{n}}\,\big] = \mathbb{E}_\vartheta\left[(\widehat{\vartheta}^{\mathtt{mc}}_{\mathrm{n}} - \vartheta)(\widehat{\vartheta}^{\mathtt{mc}}_{\mathrm{n}} - \vartheta)^T\right] = \frac{\sigma^2}{n s_n^2}\begin{pmatrix} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 & -\overline{x}_n \\ -\overline{x}_n & 1 \end{pmatrix},$$

où

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Démonstration. Comme pour la preuve de la Proposition 5.3 on peut appliquer en anticipant la Proposition 5.8 ou bien démontrer le résultat directement. □

Remarque 5.5. Sans hypothèse supplémentaire sur la loi des innovations, il est difficile de préciser ces résultats.

5.2.2 Moindres carrés et maximum de vraisemblance

Nous allons faire une hypothèse supplémentaire sur la distribution des « bruits » ξ_i qui nous permettra de construire un estimateur de σ^2.

Hypothèse 5.5. *Les « bruits » ξ_i sont indépendants, de même loi $\mathcal{N}(0, \sigma^2)$.*

Sous cette hypothèse forte qui renforce l'Hypothèse 5.4, l'estimateur du maximum de vraisemblance fournit un estimateur du paramètre $(\vartheta_0, \vartheta_1, \sigma^2)$ dont les deux premières composantes coïncident avec l'estimateur des moindres carrés de la Proposition 5.3.

Proposition 5.5. *Sous l'Hypothèse 5.5, l'estimateur du maximum de vraisemblance*

$$\widehat{\vartheta}_n^{\mathtt{mv}} = \big(\widehat{\vartheta}_{n,0}^{\mathtt{mv}}, \widehat{\vartheta}_{n,1}^{\mathtt{mv}}, \widehat{\sigma}_n^2\big)$$

est bien défini. On a

$$\big(\widehat{\vartheta}_{n,0}^{\mathtt{mv}}, \widehat{\vartheta}_{n,1}^{\mathtt{mv}}\big) = \big(\widehat{\vartheta}_{n,0}^{\mathtt{mc}}, \widehat{\vartheta}_{n,1}^{\mathtt{mc}}\big),$$

et

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\xi}_i^2, \quad \textit{où} \quad \widehat{\xi}_i = \big(Y_i - r(\widehat{\vartheta}_n^{\mathtt{mc}}, x_i)\big)^2.$$

Démonstration. D'après le Corollaire 5.1, si $g_\sigma(x) = (2\pi\sigma^2)^{-1/2}\exp(-x^2/2\sigma^2)$ désigne la densité de la loi $\mathcal{N}(0,\sigma^2)$, la vraisemblance de l'expérience statistique est donnée par

$$\mathcal{L}_n\big(\vartheta_0, \vartheta_1, \sigma^2, Y_1, \ldots, Y_n\big) = \prod_{i=1}^n g_\sigma\big(Y_i - r(\vartheta, x_i)\big),$$

et la log-vraisemblance vaut alors

$$\ell_n(\vartheta_0, \vartheta_1, \sigma^2, Y_1, \ldots, Y_n) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \vartheta_0 - \vartheta_1 x_i)^2.$$

On a

$$\partial_{\sigma^2}\ell_n(\vartheta_0, \vartheta_1, \sigma^2, Y_1, \ldots, Y_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (Y_i - \vartheta_0 - \vartheta_1 x_i)^2$$

et ce terme est nul si et seulement si

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \vartheta_0 - \vartheta_1 x_i)^2.$$

Par ailleurs, le calcul de $\partial_{\vartheta_0}\ell_n(\vartheta_0, \vartheta_1, \sigma^2, Y_1, \ldots, Y_n)$ et $\partial_{\vartheta_1}\ell_n(\vartheta_0, \vartheta_1, \sigma^2, Y_1, \ldots, Y_n)$ mène à une constante multiplicative près à celui des fonction $\partial_{\vartheta_i} L_n(\vartheta_0, \vartheta_1, Y_1, \ldots, Y_n)$, pour $i = 0, 1$ de la preuve de la Proposition 5.3. On en déduit le point annoncé $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$ comme l'unique point critique de la fonction de vraisemblance, et on vérifie que c'est bien un maximum global. □

5.3 Régression linéaire multiple

5.3.1 Modèle linéaire

On généralise le modèle de régression linéaire simple et considérant des points de « design » vectoriels. On considère l'expérience statistique engendrée par l'observation de

$$(\boldsymbol{x}_1, Y_1, \ldots, \boldsymbol{x}_n, Y_n)$$

avec

$$Y_i = \vartheta^T \boldsymbol{x}_i + \xi_i, \quad i = 1, \ldots, n \tag{5.4}$$

où les Y_i sont à valeurs dans $\mathbb{R}$, les variables explicatives $\boldsymbol{X}_i$ sont à valeurs dans $\mathbb{R}^k$, et le paramètre $\vartheta \in \Theta = \mathbb{R}^d$ est k-dimensionnel, c'est-à-dire $d = k$. Matriciellement, si l'on désigne par $\mathbb{M}$ la matrice dont les colonnes sont les vecteurs $\boldsymbol{x}_i$, c'est-à-dire, si l'on note $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,k})^T$,

$$\mathbb{M} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,k} \\ \ldots & \ldots & \ldots & \ldots \\ x_{i,1} & x_{i,2} & \ldots & x_{i,k} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,k} \end{pmatrix}$$

et la représentation (5.4) s'écrit de la même manière

$$\mathbf{Y} = \mathbb{M}\,\vartheta + \boldsymbol{\xi}, \tag{5.5}$$

où $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ et $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$. Comme pour le modèle de régression linéaire simple, nous faisons une hypothèse sur le « bruit » $\boldsymbol{\xi}$:

$$\mathbb{E}\left[\boldsymbol{\xi}\right] = 0, \quad \mathbb{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^T\right] = \sigma^2 \mathrm{Id}_n. \tag{5.6}$$

5.3.2 Estimateur des moindres carrés

Dans ce contexte on cherche l'estimateur des moindres carrés pour ϑ, c'est-à-dire l'estimateur $\widehat{\vartheta}_n^{\,\mathtt{mc}}$ qui minimise la somme du carré des résidus :

$$\sum_{i=1}^{n} \left(Y_i - (\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})^T \boldsymbol{x}_i\right)^2 = \min_{\vartheta \in \mathbb{R}^k} \sum_{i=1}^{n} \left(Y_i - \vartheta^T \boldsymbol{x}_i\right)^2.$$

Il existe toujours une solution à ce problème de minimisation mais elle n'est pas nécessairement unique.

Définition 5.5. *On appelle estimateur des moindres carrés tout estimateur $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ satisfaisant*

$$\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} \in \arg\min_{\vartheta \in \mathbb{R}^k} \sum_{i=1}^{n} \left(Y_i - \vartheta^T \boldsymbol{x}_i\right)^2.$$

Une condition suffisante d'unicité de l'estimateur des moindres carrés est la suivante :

Proposition 5.6. *On suppose la matrice* $\mathbb{M}^T\,\mathbb{M}$ *inversible. Alors l'estimateur des moindres carrés est unique et s'écrit*

$$\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = \big(\mathbb{M}^T\,\mathbb{M}\big)^{-1}\,\mathbb{M}^T\,\mathbf{Y}.$$

Nous donnons deux preuves et deux interprétations de ce résultat :

Méthode analytique

Démonstration. Le point $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ est nécessairement un point critique de l'application $\vartheta \rightsquigarrow h(\vartheta) = \sum_{i=1}^n \big(Y_i - \vartheta^T\,\boldsymbol{x}_i\,\big)^2$, c'est-à-dire il est solution du système de k équations

$$\partial_{\vartheta_j} h\big(\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\big) = 0, \quad j = 1, \ldots, k,$$

ce qui s'écrit

$$2\sum_{i=1}^{n} \boldsymbol{x}_i\,\big(Y_i - \big(\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\big)^T\,\boldsymbol{x}_i\,\big) = 0$$

ou encore, sous forme matricielle :

$$\mathbb{M}^T\,\mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = \mathbb{M}^T\,\mathbf{Y}. \tag{5.7}$$

L'équation (5.7) est un système de k équations qui a une solution unique dès lors que $\mathbb{M}^T\,\mathbb{M}$ est inversible, donnée par

$$\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = \big(\mathbb{M}^T\,\mathbb{M}\big)^{-1}\,\mathbb{M}^T\,\mathbf{Y}.$$

La fonction $\vartheta \rightsquigarrow h(\vartheta)$ est convexe et positive, donc la solution $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ est un minimum global. □

Définition 5.6. *L'équation* (5.7) *est appelée système d'équations normales pour la méthode des moindres carrés.*

Proposition 5.7. *La matrice* $\mathbb{M}^T\,\mathbb{M}$ *est (symétrique) positive. Elle est définie positive si et seulement si* $\mathrm{rang}(\mathbb{M}) = k$.

Démonstration. On a, pour $v \in \mathbb{R}^p$

$$v^T\big(\mathbb{M}^T\,\mathbb{M}\big)v = w^T w \geq 0$$

où l'on a posé implicitement $w = \mathbb{M}\,v$. Le cas d'égalité est vérifié si et seulement si $w = 0$, c'est-à-dire, $\mathbb{M}\,v = 0$. Si $\mathrm{rang}(\mathbb{M}) < k$, alors il existe $v \neq 0$ tel que $\mathbb{M}\,v = 0$ et dans ce cas, $\mathbb{M}^T\,\mathbb{M}$ n'est pas strictement positive. Réciproquement, si $\mathbb{M}^T\,\mathbb{M}$ n'est pas strictement positive, alors il existe $v \neq 0$ tel que $v^T\big(\mathbb{M}^T\,\mathbb{M}\big)v = 0$, et donc $\mathbb{M}\,v = 0$ d'où $\mathrm{rang}(\mathbb{M}) < k$. □

Remarque 5.6. En conséquence, si la taille de l'échantilllon est plus petite que la dimension du paramètre ϑ, c'est-à-dire si $n < k$, la matrice $\mathbb{M}^T\,\mathbb{M}$ est dégénérée.

Méthode géométrique

Deuxième démonstration de la Proposition 5.6. Soit V l'image de $\mathbb{R}^n$ par l'application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^k$ de matrice $\mathbb{M}$, c'est-à-dire

$$V = \big\{ v \in \mathbb{R}^n, \;\; v = \mathbb{M}\,\vartheta, \vartheta \in \mathbb{R}^k \big\}.$$

Alors

$$\min_{\vartheta \in \mathbb{R}^k} \|y - \mathbb{M}\,\vartheta\|^2 = \min_{v \in V} \|y - v\|^2,$$

où $\|v\|^2 = v^T v$ désigne le carré de la norme euclidienne. Notons que $\mathbb{M}$ est de rang k si et seulement si la dimension de V est k. D'après la Proposition 5.7, puisque $\mathbb{M}^T\,\mathbb{M}$ est supposée inversible, on a bien $\dim V = k$. Alors, si P_V désigne la matrice du projecteur orthogonal sur V dans $\mathbb{R}^n$, on a $\mathrm{rang}(P_V) = k$ et l'estimateur des moindres carrés vérifie

$$\mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = P_V \mathbf{Y}, \tag{5.8}$$

ce qui se traduit par

$$\langle \mathbf{Y} - P_V \mathbf{Y}, v \rangle = 0, \;\; \text{pour tout} \;\; v \in V,$$

où, pour $u, v \in \mathbb{R}^n$, on note $\langle u, v \rangle = u^T v$ le produit scalaire euclidien. En appliquant (5.8), l'équation précédente s'écrit encore pour tout $v \in V$

$$\langle \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}, v \rangle = \langle \mathbf{Y}, v \rangle,$$

c'est-à-dire, pour tout $\vartheta \in \mathbb{R}^k$

$$\langle \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}, \mathbb{M}\,\vartheta \rangle = \langle \mathbf{Y}, \mathbb{M}\,\vartheta \rangle,$$

soit, pour tout $\vartheta \in \mathbb{R}^k$

$$\langle \mathbb{M}^T\,\mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}, \vartheta \rangle = \langle \mathbb{M}^T\,\mathbf{Y}, \vartheta \rangle.$$

Puisque $\mathbb{M}^T\,\mathbb{M}$ est inversible, on en déduit $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = \big(\mathbb{M}^T\,\mathbb{M}\big)^{-1}\,\mathbb{M}^T\,\mathbf{Y}$. □

Remarque 5.7. A ce stade de l'étude, l'estimateur des moindres carrés, comme pour le cas de la régression linéaire simple, on n'a pas besoin de faire d'hypothèse probabiliste sur le modèle. La méthode des moindres carrés dépasse le cadre de l'estimation statistique et apparaît plus généralement comme une méthode de « régularisation » en analyse numérique. Cependant, des hypothèses probabilistes sur le « bruit » $\boldsymbol{\xi}$ permettent d'affiner significativement ses propriétés.

5.3.3 Propriétés de la méthode des moindres carrés

Proposition 5.8. *Supposons la matrice $\mathbb{M}^T\,\mathbb{M}$ inversible, et que le « $\boldsymbol{\xi}$ » satisfait* (5.6)*. On a*

$$\mathbb{E}_\vartheta \big[\, \widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} \,\big] = \vartheta,$$

et la matrice de variance-covariance de $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ *est donnée par*

$$\Sigma\big[\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\big] = \mathbb{E}_\vartheta\left[(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)^T\right] = \sigma^2\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}.$$

Démonstration. On a

$$\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = \big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\,\mathbb{M}^T\,\mathbf{Y} = \big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\,\mathbb{M}^T\,\big(\,\mathbb{M}\,\vartheta + \boldsymbol{\xi}\big) = \vartheta + \big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\,\mathbb{M}^T\,\boldsymbol{\xi},$$

d'où la première partie de la proposition, puisque $\mathbb{E}_\vartheta\left[\boldsymbol{\xi}\right] = 0$. Puis,

$$\begin{aligned}
&\mathbb{E}_\vartheta\left[(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)^T\right] \\
=&\,\mathbb{E}_\vartheta\left[\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\,\mathbb{M}^T\,\boldsymbol{\xi}\big(\boldsymbol{\xi}^T\,\mathbb{M}\,\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\big)\right] \\
=&\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\,\mathbb{M}^T\,\mathbb{E}_\vartheta\left[\boldsymbol{\xi}\boldsymbol{\xi}^T\right]\mathbb{M}\,\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}.
\end{aligned}$$

Puisque $\mathbb{E}_\vartheta\left[\boldsymbol{\xi}\boldsymbol{\xi}^T\right] = \sigma^2\mathrm{Id}_n$, le dernier terme devient

$$\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\,\mathbb{M}^T\,\sigma^2\,\mathbb{M}\,\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1} = \sigma^2\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}.$$

□

Proposition 5.9 (Estimation de la variance σ^2)**.** *On suppose la matrice* $\mathbb{M}^T\,\mathbb{M}$ *inversible, et que le « bruit »* $\boldsymbol{\xi}$ *satisfait* (5.6)*. Alors l'estimateur*

$$\widehat{\sigma}_n^2 = \frac{\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\|^2}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}\big(Y_i - \big(\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\big)^T\,\boldsymbol{x}_i\,\big)^2$$

vérifie

$$\mathbb{E}\left[\widehat{\sigma}_n^2\right] = \sigma^2.$$

Démonstration. On a la décomposition

$$\begin{aligned}
\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} &= \mathbb{M}(\vartheta - \widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}) + \boldsymbol{\xi} \\
&= -\,\mathbb{M}\,\big(\,\mathbb{M}^T\,\mathbb{M}\,\big)^{-1}\,\mathbb{M}^T\,\boldsymbol{\xi} + \boldsymbol{\xi} \\
&= (\mathrm{I}_n - P_V)\boldsymbol{\xi},
\end{aligned}$$

où $V \subset \mathbb{R}^n$ est l'image de $\mathbb{R}^p$ par l'application linéaire de matrice $\mathbb{M}$ comme précédemment. Par conséquent

$$\begin{aligned}
\mathbb{E}_\vartheta\left[\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\|^2\right] &= \mathbb{E}_\vartheta\left[\boldsymbol{\xi}^T\big(\mathrm{I}_n - P_V\big)^T\big(\mathrm{I}_n - P_V\big)\boldsymbol{\xi}\right] \\
&= \mathbb{E}_\vartheta\left[\boldsymbol{\xi}^T\big(\mathrm{I}_n - P_V\big)^2\boldsymbol{\xi}\right] \\
&= \mathbb{E}_\vartheta\left[\boldsymbol{\xi}^T\big(\mathrm{I}_n - P_V\big)\boldsymbol{\xi}\right],
\end{aligned}$$

où l'on utilise le fait que la matrice $\mathrm{I}_n - P_V$ est symétrique et idempotente. Il vient

$$\begin{aligned}\mathbb{E}_\vartheta\left[\boldsymbol{\xi}^T\big(\mathrm{I}_n - P_V\big)\boldsymbol{\xi}\right] &= \mathbb{E}\,\vartheta\big[\mathrm{trace}\big(\boldsymbol{\xi}^T\big(\mathrm{I}_n - P_V\big)\boldsymbol{\xi}\big)\big] \\ &= \mathbb{E}_\vartheta\left[\mathrm{trace}\big(\mathrm{I}_n - P_V\big)\boldsymbol{\xi}\boldsymbol{\xi}^T\right] \\ &= \mathrm{trace}\Big(\big(\mathrm{I}_n - P_V\big)\,\mathbb{E}_\vartheta\left[\boldsymbol{\xi}\boldsymbol{\xi}^T\right]\Big) \\ &= \sigma^2(n-k).\end{aligned}$$

□

5.3.4 Régression linéaire multiple gaussienne

Loi des estimateurs

On fait l'hypothèse supplémentaire que $\boldsymbol{\xi}$ est un vecteur gaussien, dont les composantes sont indépendantes, ce qui revient exactement à l'Hypothèse 5.5. On a alors la loi explicite de l'estimateur des moindres carrés.

Proposition 5.10. *On se place sous l'Hypothèses 5.5 et on suppose que la matrice $\mathbb{M}^T\,\mathbb{M}$ inversible. On a*

(i) l'estimateur des moindres carrés $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ est un vecteur gaussien k-dimensionnel de moyenne ϑ et de matrice de variance-covariance $\sigma^2\big(\mathbb{M}^T\,\mathbb{M}\big)^{-1}$,

(ii) les vecteurs aléatoires $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ et $\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ sont indépendants (et de même, les vecteurs aléatoires $\mathbb{M}(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)$ et $\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ sont indépendants),

(iii) la variable aléatoires $\sigma^2\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\|^2$ suit la loi $\chi^2(n-k)$ du χ^2 à $n-k$ degrés de liberté, et $\sigma^2\|\,\mathbb{M}(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)\|^2$ suit la loi $\chi^2(k)$ du χ^2 à k degrés de liberté.

Démonstration. On écrit, comme pour la preuve de la Proposition 5.8

$$\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = \vartheta + \big(\mathbb{M}^T\,\mathbb{M}\big)^{-1}\,\mathbb{M}^T\,\boldsymbol{\xi},$$

et on en déduit immédiatement le point (i) : $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ est un vecteur gaussien comme transformation affine de $\boldsymbol{\xi}$ qui est un vecteur gaussien ; la moyenne de $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ est ϑ et sa matrice de variance-covariance $\sigma^2\big(\mathbb{M}^T\,\mathbb{M}\big)^{-1}$ d'après la Proposition 5.8.

On a aussi

$$\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = (\mathrm{Id}_n - P_V)\boldsymbol{\xi}$$

avec les notations de la preuve de la Proposition 5.9. Donc $(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}, \mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})$ est un vecteur gaussien de $\mathbb{R}^{k+n}$ comme transformation affine du vecteur gaussien $\boldsymbol{\xi}$. Pour montrer

l'indépendance dans (ii), on applique la Proposition 1.6. Il vient

$$\begin{aligned}\Sigma\big[\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}, \mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\big] &= \mathbb{E}_\vartheta\left[(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)(\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})^T\right] \\ &= \mathbb{E}_\vartheta\left[(\mathbb{M}^T\,\mathbb{M})^{-1}\,\mathbb{M}^T\,\boldsymbol{\xi}\boldsymbol{\xi}^T(\mathrm{Id}_n - P_V)\right] \\ &= 0,\end{aligned}$$

car P_V s'écrit $P_V = \mathbb{M}\left(\mathbb{M}^T\,\mathbb{M}\right)^{-1}\mathbb{M}^T$. Donc $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ et $\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ sont indépendants, et par suite $\mathbb{M}(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)$ et $\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ sont indépendants.

Le point (iii) est une application de la Proposition 1.1 (Cochran) : le vecteur $\boldsymbol{\xi}' = \sigma^{-1}\boldsymbol{\xi}$ est gaussien de matrice de variance-covariance l'identité sur $\mathbb{R}^n$. De plus

$$\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} = \sigma(\mathrm{Id}_n - P_V)\boldsymbol{\xi}', \quad \mathbb{M}(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta) = \sigma P_V\boldsymbol{\xi}'$$

et les matrices P_V et $\mathrm{Id}_n - P_V$ sont idempotentes, voir la preuve de la Proposition 5.8, et on a $(\mathrm{Id}_n - P_V)P_V = 0$, avec $\mathrm{Rang}(P_V) = k$ et $\mathrm{Rang}(\mathrm{Id}_n - P_V) = n - k$. □

Remarque sur la loi des estimateurs et l'approche asymptotique

Dans le cas où $\boldsymbol{\xi}$ est un vecteur gaussien, les lois de $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ et $\widehat{\sigma}_n^2$ sont explicites, à n fixé. Il s'agit d'un résultat exact sur les lois des estimateurs dans un cadre non-asymptotique[2]. Ceci n'est plus vrai si la loi des innovations n'est pas gaussienne. Dans ce cas, on essaye de se ramener au cas gaussien pas des arguments asymptotiques.

Par exemple, dans le cas le plus simple où l'on observe

$$Y_i = \vartheta + \xi_i, \;\; i = 1, \ldots, m$$

où les innovations ξ sont indépendantes, identiquement distribuées, de moyenne 0 et de variance $\tau^2 > 0$ et $\vartheta \in \Theta = \mathbb{R}$. Alors, on observe aussi

$$\overline{Y}_m = \vartheta + \frac{\tau}{m}\,\widetilde{\xi}^{(m)},$$

où $\widetilde{\xi}^{(m)} = \frac{1}{\sqrt{m}}\sum_{i=1}^m \xi_i$ est une variable « asymptotiquement gaussienne » par le théorème central limite, dans le sens où $\widetilde{\xi}^{(m)} \xrightarrow{d} \mathcal{N}(0,1)$ dans la limite $m \to \infty$. On est donc ramené au cas de la régression gaussienne, mais dans un cadre très dégénéré : ici, on a $k = d = 1$, $\mathbb{M} = 1$ et $\sigma^2 = \frac{\sigma^2}{m}$ et $n = 1$ (une seule observation). Le cas d'une dimension plus grande et d'un « design » non-dégéneré est plus délicat à traiter : on peut chercher à « regrouper » les observations en faisant des moyennes, de sorte de se ramener au cas gaussien via le théorème central-limite. Nous ne développons pas ce point.

[2] on dit parfois « à distance finie ».

En conclusion, l'obtention de lois explicites pour l'estimateur des moindres carrés dans un cadre non-asymptotique est un fait remarquable, mais à considérer avec précaution du point de vue de la modélisation : l'hypothèse de gaussianité sur les innovations est en fait elle-même de nature asymptotique.

5.4 Régression non-linéaire

5.4.1 Moindres carrés non-linéaires et M-estimation

Situation

On se place dans le contexte général de la Section 5.1.2. On fait l'Hypothèse 5.3 et on observe

$$(\boldsymbol{x}_1, Y_1, \ldots, \boldsymbol{x}_n, Y_n),$$

où

$$Y_i = r(\vartheta, \boldsymbol{x}_i) + \xi_i, \quad i = 1, \ldots, n, \tag{5.9}$$

où les $\boldsymbol{x}_i \in \mathbb{R}^k$ sont donnés et $\vartheta \in \Theta \subset \mathbb{R}^d$ est le paramètre inconnu. Contrairement à la section précédente, on ne suppose plus $r(\vartheta, \bullet)$ linéaire, et il n'y a donc plus de raison de supposer $d = k$.

Vraisemblance et moindres carrés

Imposons pour simplifier l'hypothèse de gaussianité 5.5 sur les innovations ξ_i, qui sont donc indépendantes, de même loi $\mathcal{N}(0, \sigma^2)$. Dans ce cas, la log-vraisemblance s'écrit

$$\ell_n(\vartheta, Y_1, \ldots, Y_n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(Y_i - r(\vartheta, \boldsymbol{x}_i)\big)^2.$$

Le calcul de l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$ de ϑ consiste à minimiser la fonction

$$\vartheta \rightsquigarrow \sum_{i=1}\big(Y_i - r(\vartheta, \boldsymbol{x}_i)\big)^2.$$

Dans le cas du modèle linéaire de la Section 5.2, si l'on postule la forme $r(\vartheta, \boldsymbol{x}) = \vartheta^T \boldsymbol{x}$ avec $d = k$, on retrouve aussi l'estimateur des moindres carrés. De manière générale, sans hypothèse particulière sur les innovations ξ, on peut poser la définition

Définition 5.7 (Estimateur des moindres carrés non-linéaires)**.** *Etant donné le modèle de régression non-linéaire* (5.9)*, on appelle estimateur des moindres carrés non-linéaire, s'il existe, tout estimateur* $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mcnl}}$ *satisfaisant*

$$\sum_{i=1}^{n}\big(Y_i - r(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mcnl}}, \boldsymbol{x}_i)\big)^2 = \inf_{\vartheta\in\Theta}\sum_{i=1}^{n}\big(Y_i - r(\vartheta, \boldsymbol{x}_i)\big)^2.$$

Cette définition se généralise très naturellement à une notion de M-estimateur de la façon suivante : soit

$$\psi : \Theta \times \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}_+$$

une application jouant le même rôle que l'application $\psi(\bullet, \bullet)$ de la Section 4.3 du Chapitre 4 pour l'estimation dans le modèle de densité, à ceci près qu'on l'autorise désormais à dépendre de $\boldsymbol{x}_i$.

Définition 5.8. *On appelle M-estimateur associé à la fonction de contraste ψ tout estimateur $\widehat{\vartheta}_n$ satisfaisant*

$$\sum_{i=1}^{n} \psi(\widehat{\vartheta}_n, \boldsymbol{x}_i, Y_i) = \max_{\vartheta \in \Theta} \sum_{i=1}^{n} \psi(\vartheta, \boldsymbol{x}_i, Y_i).$$

Dans ce contexte, l'estimateur des moindres carrés non-linéaires apparaît comme le M-estimateur associé à la fonction de contraste

$$a \rightsquigarrow \psi(a, \boldsymbol{x}, y) = -\big(y - r(a, \boldsymbol{x})\big)^2, \quad a \in \Theta.$$

Une étude systématique des propriétés asymptotiques des M-estimateurs pour le modèle de la régression se fait essentiellement de la même manière que pour le modèle de densité du Chapitre 4, mais les aspects techniques sont plus développés. Nous développons – sans entrer dans les détails – quelques exemples.

5.4.2 Reconstruction d'un signal échantillonné

On considère l'expérience statistique engendrée par

$$Y_i = r(\vartheta, i/n) + \xi_i, \quad i = 1, \ldots, n$$

où les $\xi_i = \sigma\varepsilon_i$ sont indépendants et identiquement distribués, centrés et $\mathbb{E}_\vartheta\left[\varepsilon_i^2\right] = 1$. La fonction $r(\vartheta, \bullet)$ est connue au paramètre $\vartheta \in \Theta \subset \mathbb{R}^d$ près. Ici, le « design » est donc $\big(1/n, \ldots, (n-1)/n, 1\big)$.

On suppose que la fonction $(\vartheta, x) \rightsquigarrow r(\vartheta, x)$ est régulière. En particulier, $x \rightsquigarrow r(\vartheta, x)$ est au moins continue. L'estimateur des moindres carrés non-linéaires, s'il est bien défini, vérifie

$$\widehat{\vartheta}_n^{\texttt{mcnl}} = \arg\min_{\vartheta \in \Theta} \sum_{i=1}^{n} \big(Y_i - r(\vartheta, i/n)\big)^2.$$

Indiquons brièvement comment généraliser les résultats de la Section 4.3.3 sans faire d'hypothèses précises.

Consistance

Posons, pour $a \in \Theta \subset \mathbb{R}$ (traitons le cas unidimensionnel pour simplifier),

$$M_n(a) = \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - r(a, i/n)\big)^2.$$

On écrit

$$\begin{aligned} M_n(a) &= \frac{1}{n}\sum_{i=1}^{n} \big(\sigma\varepsilon_i + r(\vartheta, i/n) - r(a, i/n)\big)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \big(r(\vartheta, i/n) - r(a, i/n)\big)^2 + \frac{\sigma^2}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \frac{2\sigma}{n}\sum_{i=1}^{n}\big(r(\vartheta, i/n) - r(a, i/n)\big)\varepsilon_i, \end{aligned}$$

où la loi des ε_i sous $\mathbb{P}_\vartheta$ est centrée et réduite. Par continuité de $x \rightsquigarrow r(\vartheta, x)$, on a la convergence

$$\frac{1}{n}\sum_{i=1}^{n} \big(r(\vartheta, i/n) - r(a, i/n)\big)^2 \to \int_0^1 \big(r(\vartheta, x) - r(a, x)\big)^2 dx.$$

Par la loi des grands nombres, on a

$$\frac{\sigma^2}{n}\sum_{i=1}^{n}\varepsilon_i^2 \xrightarrow{\mathbb{P}_\vartheta} \sigma^2,$$

et, par un simple calcul de variance,

$$\frac{2\sigma}{n}\sum_{i=1}^{n}\big(r(\vartheta, i/n) - r(a, i/n)\big)\varepsilon_i \xrightarrow{\mathbb{P}_\vartheta} 0.$$

Donc

$$M_n(a) \xrightarrow{\mathbb{P}_\vartheta} M(a, \vartheta) = \int_0^1 \big(r(\vartheta, x) - r(a, x)\big)dx + \sigma^2.$$

La suite de l'étude consiste à des hypothèses d'identifiabilité adéquates sur la fonction $(\vartheta, x) \rightsquigarrow r(\vartheta, x)$, de sorte que $a \rightsquigarrow M(a, \vartheta)$ admette un minimum unique en $a = \vartheta$, et on peut généraliser la Proposition 4.3, mais cela dépasse le cadre de notre étude.

Loi limite et normalité asymptotique

Avec suffisamment de régularité, on peut faire un développement de $M_n'(a)$ au voisinage de $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mc}}$. On a

$$M_n'(\widehat{\vartheta}_n^{\mathtt{mcnl}}) = 0 \approx M_n'(\vartheta) + (\vartheta - \widehat{\vartheta}_n^{\mathtt{mcnl}})M_n''(\vartheta), \tag{5.10}$$

d'où

$$\sqrt{n}(\widehat{\vartheta}_n^{\texttt{mcnl}} - \vartheta) \approx -\frac{\sqrt{n}M_n'(\vartheta)}{M_n''(\vartheta)}.$$

On a

$$\begin{aligned}\sqrt{n}M_n'(\vartheta) &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\big(Y_i - r(\vartheta, i/n)\big)\partial_\vartheta r(\vartheta, i/n)\\ &= \frac{2\sigma}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\,\partial_\vartheta r(\vartheta, i/n),\end{aligned}$$

d'où

$$\mathbb{E}_\vartheta\left[\sqrt{n}M_n'(\vartheta)\right] = 0,$$

et

$$\begin{aligned}\mathbb{E}_\vartheta\left[nM_n'(\vartheta)^2\right] &= \frac{4\sigma^2}{n}\sum_{i=1}^{n}\partial_\vartheta r(\vartheta, i/n)^2\,\varepsilon_i^2\\ &\to 4\sigma^2\int_0^1 \partial_\vartheta r(\vartheta, x)^2 dx.\end{aligned}$$

En ré-écrivant $\sqrt{n}M_n'(\vartheta) = \big(\mathbb{E}_\vartheta\left[nM_n'(\vartheta)^2\right]\big)^{1/2}\xi^{(n)}$, on peut montrer[3] que $\xi^{(n)} \xrightarrow{d} \mathcal{N}(0,1)$ en loi sous $\mathbb{P}_\vartheta$. On a aussi

$$\begin{aligned}M_n''(\vartheta) &= \frac{2}{n}\sum_{i=1}^{n}\big(-\partial_\vartheta r(\vartheta, i/n)^2 + \sigma\varepsilon_i\,\partial_\vartheta^2 r(\vartheta, i/n)\big)\\ &\xrightarrow{\mathbb{P}_\vartheta} -2\int_0^1 \partial_\vartheta r(\vartheta, x)^2 dx.\end{aligned}$$

On en déduit, avec suffisamment de régularité et en contrôlant le reste dans l'approximation (5.10),

$$\sqrt{n}\big(\widehat{\vartheta}_n^{\texttt{mcnl}} - \vartheta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\sigma^2}{\int_0^1 \partial_\vartheta r(\vartheta, x)^2 dx}\Big).$$

5.4.3 Modèle de Poisson conditionnel

On observe

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$$

où les $\boldsymbol{x}_i \in \mathbb{R}^k$ sont donnés et les Y_i à valeurs entières. On suppose que Y_i suit la loi de Poisson de paramètre

$$\lambda_i(\vartheta) = \exp\big(\boldsymbol{x}_i^T\,\vartheta\big), \quad i = 1, \ldots, n$$

[3] Il faut disposer d'un théorème central-limite pour des variables aléatoires indépendantes non-équidistribuées.

où $\vartheta \in \Theta = \mathbb{R}^k$ est le paramètre inconnu.

Si l'on considère le modèle de régression à « design » aléatoire associé, alors on observe un n-échantillon

$$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$$

où les $(\boldsymbol{X}_i, Y_i)$ ont la même loi que $(\boldsymbol{X}, Y) \in \mathbb{R}^k \times \mathbb{R}$. La loi de $(\boldsymbol{X}, Y)$ est décrite de la façon suivante : conditionnellement[4] à $\boldsymbol{X} = \boldsymbol{x}$, la variable Y suit une loi de Poisson de paramètre $\exp\big(\boldsymbol{x}^T \vartheta\big)$. Puis, on doit spécifier[5] la loi de $\boldsymbol{X}$. En écrivant

$$Y_i = \exp\big(\boldsymbol{x}_i^T \vartheta\big) + \Big(Y_i - \exp\big(\boldsymbol{x}_i^T \vartheta\big)\Big),$$

on obtient bien la représentation $Y_i = r(\vartheta, \boldsymbol{x}_i) + \xi_i$, avec

$$r(\vartheta, \boldsymbol{x}_i) = \exp\big(\boldsymbol{x}_i^T \vartheta\big)$$

et

$$\xi_i = Y_i - \exp\big(\boldsymbol{x}_i^T \vartheta\big).$$

On a bien $\mathbb{E}_\vartheta\big[\xi_i\big] = 0$ en utilisant que l'espérance d'une variable aléatoire de Poisson de paramètre λ est égale à λ. La vraisemblance du modèle s'écrit

$$\mathcal{L}_n(\vartheta, Y_1, \ldots, Y_n) = \prod_{i=1}^n e^{-\lambda_i(\vartheta)} \frac{\lambda_i(\vartheta)^{Y_i}}{Y_i!}$$

d'où

$$\log \mathcal{L}_n(\vartheta, Y_1, \ldots, Y_n) = -\sum_{i=1}^n \exp\big(\boldsymbol{x}_i^T \vartheta\big) + \sum_{i=1}^n Y_i \, \boldsymbol{x}_i^T \vartheta - \sum_{i=1}^n \log(Y_i!),$$

et les équations de vraisemblance s'écrivent

$$-\sum_{i=1}^n x_{ij} \exp\big(\boldsymbol{x}_i^T \vartheta\big) + \sum_{i=1}^n Y_i x_{ij} = 0, \quad j = 1, \ldots, k.$$

5.4.4 Modèles à réponse binaire

Contexte général

Très utilisés en pratique, les modèles binaires correspondent à l'observation de

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$$

où $\boldsymbol{x}_i \in \mathbb{R}^k$ est un ensemble de caractéristiques de l'individu i qui est de type $Y_i \in \{0, 1\}$.

[4] d'où la terminologie de modèle de Poisson conditionnel.

[5] ce que nous ne ferons jamais ; nous supposerons simplement que la loi de $\boldsymbol{X}$ ne dépend pas de ϑ.

Par souci d'homogénéité avec la littérature, on se place – sans perdre de généralités – dans le modèle à « design » aléatoire correspondant, c'est-à-dire que l'on considère les $\boldsymbol{x}_i$ des réalisations de variables aléatoires $\boldsymbol{X}_i$. Alors, en écrivant

$$Y_i = p_{\boldsymbol{x}_i}(\vartheta) + \big(Y_i - p_{\boldsymbol{x}_i}(\vartheta)\big),$$

avec

$$p_{\boldsymbol{x}_i}(\vartheta) = \mathbb{E}_\vartheta \left[Y_i \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i \right] = \mathbb{P}_\vartheta \left[Y_i = 1 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i \right],$$

on obtient la représentation

$$Y_i = r(\vartheta, \boldsymbol{x}_i) + \xi_i,$$

avec

$$r(\vartheta, \boldsymbol{x}_i) = p_{\boldsymbol{x}_i}(\vartheta)$$

et

$$\xi_i = Y_i - p_{\boldsymbol{x}_i}(\vartheta),$$

et on a bien $\mathbb{E}_\vartheta \left[\xi_i \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i \right] = 0$.

Régression logistique

La régression logistique correspond à la modélisation

$$p_{\boldsymbol{x}_i}(\vartheta) = \frac{\exp\big(\boldsymbol{x}_i^T \vartheta\big)}{1 + \exp\big(\boldsymbol{x}_i^T \vartheta\big)} = \psi\big(\boldsymbol{x}_i^T \vartheta\big),$$

où $\psi(x) = e^x/(1+e^x)$ est la fonction logistique.

En particulier, on peut expliciter la vraisemblance du modèle

$$\mathcal{L}_n(\vartheta, Y_1, \ldots, Y_n) = \prod_{i=1}^n p_{\boldsymbol{x}_i}(\vartheta)^{Y_i} \big(1 - p_{\boldsymbol{x}_i}(\vartheta)\big)^{Y_i},$$

que l'on peut maximiser numériquement.

Une représentation équivalente est celui des modèles dits latents, où l'on observe

$$Y_i = 1_{\left\{Y_i^\star > 0\right\}}, \quad Y_i^\star = \boldsymbol{x}_i^T \vartheta + U_i, \tag{5.11}$$

où les Y_i sont des variables *latentes*, c'est-à-dire que l'on n'observe pas, et U_i est une variable ayant pour fonction de répartition

$$F(x) = \frac{1}{1+e^{-x}}.$$

En effet,

$$\begin{aligned}\mathbb{P}_\vartheta\left[Y_i^\star > 0 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i\right] &= \mathbb{P}_\vartheta\left[\boldsymbol{x}_i^T\,\vartheta + U_i > 0 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i\right] \\ &= 1 - \mathbb{P}_\vartheta\left[U_i \leq -\boldsymbol{x}_i^T\,\vartheta\right] \\ &= 1 - F\big(-\boldsymbol{x}_i^T\,\vartheta\big) \\ &= \frac{\exp\big(\boldsymbol{x}_i^T\,\vartheta\big)}{1+\exp\big(\boldsymbol{x}_i^T\,\vartheta\big)}.\end{aligned}$$

Modèles probit

Le modèle probit est proche de la régression logistique. Il s'agit simplement de remplacer dans la représentation (5.11) la variable U_i qui a pour fonction de répartition $F(x) = 1/(1+e^{-x})$ par une variable aléatoire U_i gaussienne, centrée.

Loi logistique et « odd-ratios »*

La loi logistique de fonction de répartition $F(x) = 1/(1+e^{-x})$ possède des queues de distributions plus épaisses que la loi gaussienne, et sa fonction de répartition est plus simple à manipuler numériquement.

Une spécificité du modèle logistique est l'interprétation du modèle en terme de risque. Imaginons que $Y_i = 1$ signifie la présence d'une maladie chez l'individu i (et $Y_i = 0$ signifie l'absence de la maladie). On peut interpréter $\boldsymbol{x}_i$ comme une ensemble de facteurs (qualitatifs ou marqueurs biologiques) susceptibles « d'expliquer » Y_i. Le risque (odd-ratio) de l'individu i est défini comme

$$\mathcal{R}_i = \frac{\mathbb{P}_\vartheta\left[Y_i = 1 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i\right]}{\mathbb{P}_\vartheta\left[Y_i = 0 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i\right]}$$

et $\mathcal{R}_i$ est proche de $\mathbb{P}_\vartheta\left[Y_i = 1 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i\right]$ (à l'ordre 1) lorsque la probabilité de présence de la maladie est faible. Dans le cas de la régression logistique, on a

$$\begin{aligned}\frac{\mathbb{P}_\vartheta\left[Y_i = 1 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i\right]}{\mathbb{P}_\vartheta\left[Y_i = 0 \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i\right]} &= \frac{\big(1+\exp(-\boldsymbol{x}_i^T\,\vartheta)\big)^{-1}}{\exp\big(-\boldsymbol{x}_i^T\,\vartheta\big)\big(1+\exp(-\boldsymbol{x}_i^T\,\vartheta)\big)} \\ &= \exp\big(\boldsymbol{x}_i^T\,\vartheta\big).\end{aligned}$$

Si une des variables explicatives x_{ij} est qualitative, pour un $j \in \{1,\dots,k\}$ c'est-à-dire à valeurs dans $\{0,1\}$ (par exemple, une réponses de type « oui ou non » à un questionnaire concernant le patient), on note

$$\boldsymbol{x}_i^{(-j)} = (x_{i1},\dots,x_{i,j-1},x_{i,j+1},\dots,x_{ik})^T,$$

c'est-à-dire $\boldsymbol{x}_i$ privé de sa j-ième composante. Posons

$$\mathcal{R}_i(X_j = 1) = \frac{\mathbb{P}_\vartheta \left[Y_i = 1 \,|\, \boldsymbol{X}_i^{(-j)} = \boldsymbol{x}_i^{(-j)}, X_j = 1 \right]}{\mathbb{P}_\vartheta \left[Y_i = 0 \,|\, \boldsymbol{X}_i^{(-j)} = \boldsymbol{x}_i^{(-j)}, X_j = 1 \right]}$$

et

$$\mathcal{R}_i(X_j = 0) = \frac{\mathbb{P}_\vartheta \left[Y_i = 1 \,|\, \boldsymbol{X}_i^{(-j)} = \boldsymbol{x}_i^{(-j)}, X_j = 0 \right]}{\mathbb{P}_\vartheta \left[Y_i = 0 \,|\, \boldsymbol{X}_i^{(-j)} = \boldsymbol{x}_i^{(-j)}, X_j = 0 \right]}.$$

Alors, on a

$$\exp\left(\vartheta_j x_{ij}\right) = \frac{\mathcal{R}_i(X_j = 1)}{\mathcal{R}_i(X_j = 0)}.$$

Cete identité peut s'interpréter de la manière suivante : le coefficient $\exp\left(\vartheta_j x_{ij}\right)$ est égal au rapport des risques correspondant à $X_j = 1$ et $X_j = 0$. Ce rapport est indépendant de la valeur de $\boldsymbol{x}_i^{(-j)}$.

Chapitre 6

Information et théorie asymptotique

6.1 Introduction

Situation

Nous nous plaçons dans le contexte des deux chapitres précédents. On cherche à estimer un paramètre d -dimensionnel $\vartheta \in \Theta \subset \mathbb{R}^d$ dans deux situations :

1. Pour le modèle de la densité, on observe un n-échantillon
$$(X_1, \ldots, X_n)$$
de variables aléatoires réelles. Les X_i suivent la loi $\mathbb{P}_\vartheta$ parmi une famille de probabilités $\left\{ \mathbb{P}_\vartheta, \vartheta \in \Theta \right\}$ donnée.
2. Pour le modèle de régression à « design déterministe », on observe n vecteurs de données
$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$$
admettant la représentation
$$Y_i = r(\vartheta, \boldsymbol{x}_i) + \xi_i, \quad i = 1, \ldots, n.$$
La forme de la fonction de régression $r(\vartheta, \bullet)$ est connue au paramètre ϑ près, et les ξ_i sont des innovations ou des « bruits » centrés sur lesquels on fait un jeu d'hypothèses.

En forçant un peu le trait, nous pouvons résumer les méthodes d'estimation des chapitres précédents à la construction d'estimateurs basés sur la maximisation d'un critère : pour la densité,
$$\widehat{\vartheta}_n \in \arg\max_{\vartheta \in \Theta} \sum_{i=1}^{n} \Psi(\vartheta, X_i),$$

où

$$\psi : \Theta \times \mathbb{R} \to \mathbb{R}$$

est la fonction de constraste définissant l'estimateur. Elle est choisie par le statisticien. Pour la régression à « design » déterministe,

$$\widehat{\vartheta}_n \in \arg\max_{\vartheta\in\Theta} \sum_{i=1}^{n} \psi(\vartheta, \boldsymbol{x}_i, Y_i),$$

où maintenant la fonction de contraste

$$\psi : \Theta \times \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$$

prend aussi comme argument les valeurs des points du « design » observés $\boldsymbol{x}_i$.

Loi limite d'un estimateur

Sous des hypothèses de régularité, le comportement asymptotique de $\widehat{\vartheta}_n$ prend la forme (en dimension $d = 1$)

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v_\Psi(\vartheta)\big) \tag{6.1}$$

où $v_\psi(\vartheta) > 0$ est la variance asymptotique de l'estimateur, qui dépend en général de ϑ et bien sûr du choix de la fonction de contraste ψ.

La version multidimensionnelle de (6.1) s'écrit

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, V_\Psi(\vartheta)\big) \tag{6.2}$$

avec $V_\psi(\vartheta)$ une matrice symétrique, et doit se comprendre comme la convergence du vecteur aléatoire $\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big)$ en loi vers un vecteur gaussien de $\mathbb{R}^d$, centré, de matrice de variance covariance $V_\psi(\vartheta)$ définie positive.

Un résultat de type (6.1) ou (6.2) nous apprend deux choses :

1. Le « bon » ordre de grandeur de l'erreur $\widehat{\vartheta}_n - \vartheta$ est $\frac{1}{\sqrt{n}}$. En effet, la convergence vers une loi non-dégénérée[1] avec la normalisation $\sqrt{n}$ implique que si l'on choisit une autre normalisation $\alpha_n \to \infty$, alors l'erreur normalisée

   $$\alpha_n(\widehat{\vartheta}_n - \vartheta)$$

   tend vers 0 en probabilité si $\alpha_n/\sqrt{n} \to 0$ et « explose[2] » si $\alpha_n/\sqrt{n} \to \infty$.

[1] c'est-à-dire une loi gaussienne de variance finie $v_\psi(\vartheta)$ non nulle ou de matrice de variance-covariance $V_\psi(\vartheta)$ non singulière.

[2] Dans le sens suivant : $\forall M > 0,\ \liminf_{n\to\infty} \mathbb{P}_\vartheta\left[|\alpha_n(\widehat{\vartheta}_n - \vartheta)| \geq M\right] > 0$.

2. La dispersion de l'erreur normalisée dans la bonne échelle $\sqrt{n}$ est gaussienne, de variance $v_\psi(\vartheta)$ (ou $V_\psi(\vartheta)$).

Ces deux informations, apparaîssent à deux niveaux complètement différents, mais sont de même importance et guideront les questions que nous aborderons dans ce chapitre :

- la vitesse d'estimation $\alpha_n = \sqrt{n}$ est-elle optimale ? Dans quel sens ? Quelles conditions simples sur la famille de lois $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ garantissent cette optimalité ? Sinon, quelles vitesses peut-on trouver en général ?
- au sein d'une classe d'estimateurs satisfaisant (6.1) (ou (6.2) dans le cas ou le paramètre ϑ est multidimensionnel), comment choisir un membre optimal, et dans quel sens ? Par exemple, comment choisir la « meilleure » fonction ψ ?

Un tel programme ainsi énoncé est trop ambitieux. Nous donnerons néanmoins des éléments de réponse à chacune des questions énoncées ci-dessus. Sous des hypothèses de régularité sur la famille $\big\{\mathbb{P}_\vartheta,\, \vartheta \in \Theta\big\}$, on peut définit une quantité d'information –l'information de Fisher– associée à l'expérience statistique. L'estimateur du maximum de vraisemblance est asymptotiquement normal de variance l'inverse de l'information de Fisher. Cette variance est minimale parmi la classe des Z-estimateurs (ou M-estimateurs réguliers) et ce résultat nous fournira une notion d'optimalité associée aux modèles réguliers.

Ce n'est que le premier pas vers une théorie plus générale de l'estimation optimale dans les modèles dits réguliers, qui dépasse le cadre de ce cours. Pour des développements plus complets, on pourra consulter V. Genon-Catalot et D. Picard [2] ou van der Vaart [7].

6.2 Comparaison d'estimateurs

Pour $n \geq 1$, on se donne $\mathcal{E}^n$ une suite d'expérience associée à la famille de probabilités $\big\{\mathbb{P}_\vartheta^n, \vartheta \in \Theta\big\}$.

Plaçons nous en dimension 1 pour simplifier. Etant donnés deux (suites d') estimateurs

$$\widehat{\vartheta}_{n,1} \quad \text{et} \quad \widehat{\vartheta}_{n,2}$$

lequel est préférable ? Si l'on dispose d'un résultat asymptotique de type (6.1) de la forme

$$\sqrt{n}\big(\widehat{\vartheta}_{n,j} - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v_j(\vartheta)\big) \quad j = 1, 2$$

alors on a le développement asymptotique

$$\widehat{\vartheta}_{n,j} = \vartheta + \sqrt{\frac{v_j(\vartheta)}{n}}\xi_{n,j},$$

où

$$\xi_{n,j} \xrightarrow{d} \mathcal{N}(0,1).$$

De ce point, de vue, il est préférable de choisir $\widehat{\vartheta}_{n,1}$ à $\widehat{\vartheta}_{n,2}$ si

$$v_1(\vartheta) \leq v_2(\vartheta). \tag{6.3}$$

Mais cela pose deux problèmes :

- le sens de l'inégalité (6.3) peut varier selon la valeur de ϑ qui est inconnue.
- cette représentation ne se justifie que dans la limite $n \to \infty$.

6.2.1 Risque quadratique en dimension 1

Cette approche est non-asymptotique. On suppose ici $d = 1$, c'est-à-dire $\Theta \subset \mathbb{R}$.

Définition 6.1. *Le risque quadratique d'un estimateur $\widehat{\vartheta}_n$ au point $\vartheta \in \Theta$ est*

$$\mathcal{R}(\widehat{\vartheta}_n, \vartheta) = \mathbb{E}_\vartheta \left[(\widehat{\vartheta}_n - \vartheta)^2\right].$$

Le risque quadratique mesure l'erreur moyenne quadratique lorsque l'on estime ϑ par $\widehat{\vartheta}_n$. Le choix de l'erreur quadratique est un peu arbitraire. On pourrait tout aussi bien considérer le risque

$$\mathbb{E}_\vartheta \left[|\, \widehat{\vartheta}_n - \vartheta|\right],$$

ou plus généralement un risque associée à une fonction de perte $(x, y) \rightsquigarrow \ell(x, y)$ abritraire

$$\mathbb{E}_\vartheta \left[\ell(\widehat{\vartheta}_n, \vartheta)\right]$$

satisfaisant $\ell(x, y) \geq 0$ avec égalité si et seulement si $x = y$. On a déja rencontrés les avantages de considérer comme mesure d'erreur la différence au carré au Chapitre 5, en particulier, le fait que $\vartheta \rightsquigarrow \mathcal{R}(\widehat{\vartheta}_n, \vartheta)$ est dérivable sous des hypothèses relativement faibles.

Remarquons aussi que l'inégalité de Tchebychev (1.2) entraîne, pour tout $\varepsilon > 0$

$$\mathbb{P}_\vartheta \left[|\, \widehat{\vartheta}_n - \vartheta| > \varepsilon\right] \leq \frac{1}{\varepsilon^2} \mathcal{R}(\widehat{\vartheta}_n, \vartheta)$$

et donc le risque quadratique permet de contrôler – au moins grossièrement – la probabilité que la précision de $\widehat{\vartheta}_n$ soit inférieure ou égale à un niveau $\varepsilon > 0$ donné. En particulier, si

$$\mathcal{R}(\widehat{\vartheta}_n, \vartheta) \to 0$$

alors

$$\widehat{\vartheta}_n \xrightarrow{\mathbb{P}_\vartheta^n} 0.$$

On en déduit la règle de sélection suivante :

Définition 6.2. *L'estimateur $\widehat{\vartheta}_{n,1}$ est préférable à l'estimateur $\widehat{\vartheta}_{n,2}$ au sens du risque quadratique au point ϑ si*

$$\mathcal{R}(\widehat{\vartheta}_{n,1}, \vartheta) \leq \mathcal{R}(\widehat{\vartheta}_{n,2}, \vartheta).$$

Notion d'admissibilité

Etant donné une (suite d') expérience(s) $\mathcal{E}^n$, existe-t-il un estimateur $\widehat{\vartheta}_n^\star$ optimal au sens de la Définition 6.2, c'est-à-dire vérifiant

$$\forall \vartheta \in \Theta, \quad \mathcal{R}(\widehat{\vartheta}_n^\star, \vartheta) \leq \inf_{\widehat{\vartheta}_n} \mathcal{R}(\widehat{\vartheta}_n, \vartheta) \ ? \tag{6.4}$$

La réponse est négative : prenons par exemple l'expérience engendrée par l'observation d'un n-échantillon de loi $\mathcal{N}(\vartheta, \sigma^2)$, avec $\vartheta \in \Theta = \mathbb{R}$ et σ^2 connu. L'estimateur du maximum de vraisemblance est

$$\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} = \overline{X}_n.$$

Considérons par ailleurs l'estimateur artificiel

$$\widehat{\vartheta}_n = 0$$

qui prend toujours la valeur 0 sans tenir compte des observations. Alors, pour tout $\vartheta \in \Theta$,

$$\mathcal{R}(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}, \vartheta) = \frac{\sigma^2}{n} \ \text{ et } \ \mathcal{R}(\widehat{\vartheta}_n, \vartheta) = \vartheta^2.$$

Il est clair que selon les valeurs de n et σ il existe des valeurs de ϑ où l'estimateur absurde $\widehat{\vartheta}_n = 0$ est préférable à $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$ pour le risque quadratique.

La situation générale est pire ! Même si Θ se réduit à deux points distincts, quelle que soit l'experience statistique, on ne peut pas construire d'estimateur optimal au sens de (6.4). Voir pour cela l'Exercice 6.1. La notion d'optimalité au sens naïf de (6.4) est impossible à réaliser.

On peut néanmoins aborder la notion de comparaison sous un angle plus faible : c'est la notion d'efficacité et d'admissibilité.

Définition 6.3 (Efficacité)**.** *Si $\widehat{\vartheta}_{n,1}$ est préférable à $\widehat{\vartheta}_{n,2}$ pour le risque quadratique en tout point $\vartheta \in \Theta$ et s'il existe un point $\widetilde{\vartheta} \in \Theta$ pour lequel on a*

$$\mathcal{R}(\widehat{\vartheta}_{n,1}, \widetilde{\vartheta}) < \mathcal{R}(\widehat{\vartheta}_{n,2}, \widetilde{\vartheta}),$$

on dit que $\widehat{\vartheta}_{n,1}$ est plus efficace que $\widehat{\vartheta}_{n,2}$ et que $\widehat{\vartheta}_{n,2}$ est inadmissible.

On en déduit une notion (faible) d'optimalité

Définition 6.4 (Admissibilité)**.** *L'estimateur $\widehat{\vartheta}_n$ est admissible s'il n'existe pas d'estimateur plus efficace que $\widehat{\vartheta}_n$.*

Optimalité sur une classe d'estimateurs

Une autre manière de contourner le problème de l'absence d'optimalité au sens (6.4) consiste à restreindre la classe des estimateurs, de sorte que des estimateurs absurdes soient éliminés d'office : pour cela, on part de la constatation suivante

Proposition 6.1 (Structure du risque quadratique)**.** *Pour tout estimateur $\widehat{\vartheta}_n$ et tout $\vartheta \in \Theta$, on a la décomposition*

$$\mathcal{R}(\widehat{\vartheta}_n, \vartheta) = \big(\mathbb{E}_\vartheta\left[\,\widehat{\vartheta}_n\,\right] - \vartheta\big)^2 + \mathrm{Var}_\vartheta\big[\,\widehat{\vartheta}_n\,\big] = \text{biais}^2 + \text{variance}.$$

Définition 6.5. *On dit que $\widehat{\vartheta}_n$ est sans biais, respectivement asymptotiquement sans biais, si*

$$\forall \vartheta \in \Theta,\ \ \mathbb{E}_\vartheta\left[\,\widehat{\vartheta}_n\,\right] = \vartheta,$$

respectivement $\lim_{n\to\infty} \mathbb{E}_\vartheta\left[\widehat{\vartheta}_n\right] = \vartheta$.

Une approche classique de la littérature statistique (un peu dépassée aujourd'hui) consiste à réaliser le programme suivant : parmi les estimateurs sans biais, chercher ceux de variance minimale. Un fait remarquable est que dans certaines situations, un tel programme est réalisable, voir l'Exercice 6.3. Cependant, cette approche reste limitée et nous ne la développerons pas dans ce cours car :

- les estimateurs sans biais n'apparaîssent que dans des situations assez particulières.
- même pour les expériences statistiques admettant des estimateurs sans biais, on peut presque toujours construire des estimateurs biaisés plus efficaces, comme le montre l'exemple suivant dans un cas simple.

Exemple 6.1. Dans le modèle engendré par l'observation d'un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, avec $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$, on s'intéresse au paramètre $\vartheta = \sigma^2$. On suppose $n \geq 2$. Considérons les estimateurs

$$\widehat{\vartheta}_{n,1} = \frac{1}{n}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2, \ \text{ et } \ \widehat{\vartheta}_{n,2} = \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2.$$

Alors $\mathbb{E}_\vartheta\left[\widehat{\vartheta}_{n,1}\right] = \frac{n-1}{n}\vartheta = \vartheta - n^{-1}\sigma^2$. En conséquence, le biais de $\widehat{\vartheta}_{n,1}$ vaut $-\sigma^2 n^{-1}$ et $\widehat{\vartheta}_{n,1}$ est biaisé. Par contre, $\mathbb{E}_\vartheta\left[\widehat{\vartheta}_{n,2}\right] = \sigma^2 = \vartheta$ et $\widehat{\vartheta}_{n,2}$ est sans biais. Par ailleurs,

$$\mathrm{Var}_\vartheta\big[\widehat{\vartheta}_{n,2}\big] = \left(\frac{n}{n-1}\right)^2, \quad \mathrm{Var}_\vartheta\big[\widehat{\vartheta}_{n,1}\big] = \frac{2\sigma^4}{n-1}.$$

On en déduit

$$\mathcal{R}(\widehat{\vartheta}_{n,1}, \vartheta) = \left(\frac{\sigma^2}{n}\right)^2 + 2\frac{n-1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4$$

et

$$\mathcal{R}(\widehat{\vartheta}_{n,2}, \vartheta) = \frac{2\sigma^4}{n-1} > \mathcal{R}(\widehat{\vartheta}_{n,1}, \vartheta)$$

pour tout $\vartheta \in \Theta$. Donc $\widehat{\vartheta}_{n,1}$ est plus efficace que $\widehat{\vartheta}_{n,2}$. L'estimateur sans biais est inadmissible.

Cependant, l'Exemple 6.1 n'est pas tout à fait honnête : la différence entre $\widehat{\vartheta}_{n,1}$ et $\widehat{\vartheta}_{n,2}$ s'estompe lorsque n grandit, au sens où

$$\lim_{n\to\infty} \frac{\mathcal{R}_n(\widehat{\vartheta}_{n,1}, \vartheta)}{\mathcal{R}(\widehat{\vartheta}_{n,2}, \vartheta)} = 1.$$

Cette remarque met plutôt en relief un défaut de la notion d'admissibilité et suggère une approche asymptotique. Nous verrons plus loin comment l'approche asymptotique élimine naturellement certains estimateurs artificiels. Nous concluons cette section avec la régle de comparaison suivante :

Définition 6.6. *L'estimateur $\widehat{\vartheta}_{n,1}$ est asymptotiquement préférable à l'estimateur $\widehat{\vartheta}_{n,2}$ au point $\vartheta \in \Theta$ si*

$$\limsup_{n\to\infty} \frac{\mathcal{R}_n(\widehat{\vartheta}_{n,1}, \vartheta)}{\mathcal{R}(\widehat{\vartheta}_{n,2}, \vartheta)} \leq 1.$$

On pourrait en définir une notion d'efficacité asymptotique analogue à la Définition 6.4 non-asymptotique. Nous reviendrons sur ce point plus tard.

6.2.2 Risque quadratique et normalité asymptotique

On suppose toujours $\Theta \subset \mathbb{R}$ pour simplifier. On a vu aux Chapitres 4 et 5 des résultats de type

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v(\vartheta)\big) \tag{6.5}$$

Supposons que la convergence ait aussi lieu en passant au carré et en prenant l'espérance, c'est-à-dire

$$\lim_{n\to\infty} n\mathcal{R}(\widehat{\vartheta}_n) = v(\vartheta). \tag{6.6}$$

Alors, l'estimateur $\widehat{\vartheta}_{n,1}$ sera asymptotiquement préférable à l'estimateur $\widehat{\vartheta}_{n,2}$ pour le risque quadratique au point ϑ si

$$v_1(\vartheta) \leq v_2(\vartheta) \tag{6.7}$$

dès lors que $\widehat{\vartheta}_{n,1}$ et $\widehat{\vartheta}_{n,2}$ vérifient des convergences de type (6.5) et (6.6).

Malheureusement, on n'a pas en général une inégalité de type (6.7) *simultanément pour tout* $\vartheta \in \Theta$. Une solution conservatrice consiste alors à préférer asymptotiquement $\widehat{\vartheta}_{n,1}$ à $\widehat{\vartheta}_{n,2}$ si

$$\sup_{\vartheta\in\Theta} v_1(\vartheta) \leq \sup_{\vartheta\in\Theta} v_2(\vartheta).$$

Ceci nous conduit à une définition faible mais très robuste de la notion d'optimalité asymptotique pour le risque quadratique.

Définition 6.7 (Risque minimax)**.** *Le risque d'un estimateur $\widehat{\vartheta}_n$ sur l'ensemble des paramètres Θ est*

$$\mathcal{R}(\widehat{\vartheta}_n \mid \Theta) = \sup_{\vartheta \in \Theta} \mathcal{R}(\widehat{\vartheta}_n, \vartheta).$$

Un estimateur $\widehat{\vartheta}_n^\star$ est asymptotiquement optimal au sens minimax pour le risque quadratique si

$$\limsup_{n\to\infty} \frac{\mathcal{R}(\widehat{\vartheta}_n^\star \mid \Theta)}{\inf_{\widehat{\vartheta}_n} \mathcal{R}(\widehat{\vartheta}_n \mid \Theta)} \leq 1,$$

où l'infimum est pris sur l'ensemble de tous les estimateurs.

Remarque 6.1. L'optimalité asymptotique au sens minimax se généralise immédiatement à d'autres fonctions de perte que la perte quadratique. Elle est couramment utilisée lorsque l'ensemble des paramètres est de grande dimension, et en particulier en estimation non-paramétrique.

Nous terminons cette section en présentant des conditions simples qui permettent de passer de (6.5) à (6.6). A quelle condition simple la convergence en loi (6.5) entraîne-t-elle une convergence de type (6.6) ? Plus généralement si Z_n est une suite de variables aléatoires réelles telle que

$$Z_n \xrightarrow{d} Z,$$

peut-on avoir

$$\lim_{n\to\infty} \mathbb{E}\big[g(Z_n)\big] = \mathbb{E}\big[g(Z)\big]$$

pour une fonction g continue non-bornée ? Si g est bornée, c'est la définition même de la convergence en loi. Dans le cas où g est non-bornée, il faut invoquer une propriété d'uniforme intégrabilité sur la suite Z_n.

Proposition 6.2. *Soit Z_n une suite de vecteurs aléatoires de $\mathbb{R}^d$ telle que $Z_n \xrightarrow{d} Z$. Alors, si $g : \mathbb{R}^d \to \mathbb{R}$ est une application continue et si l'une au moins des trois conditions suivantes est vérifiée :*

(i) $\lim_{t\to\infty} \limsup_{n\to\infty} \int_t^{+\infty} \mathbb{P}\big[|g(Z_n)| > x\big]dx = 0,$
(ii) $\mathbb{P}\big[|g(Z_n)| > x\big] \leq h(x)$, *avec* $\int_0^{+\infty} h(x)dx < +\infty,$
(iii) il existe $\varepsilon > 0$ *tel que* $\sup_n \mathbb{E}\big[|g(Z_n)|^{1+\varepsilon}\big] < +\infty,$

on a

$$\lim_{n\to\infty} \mathbb{E}\big[g(Z_n)\big] = \mathbb{E}\big[g(Z)\big].$$

Démonstration. Par l'inégalité de Tchebyshev, on a, pour $x > 0$,

$$\mathbb{P}\big[|g(Z_n)| > x\big] \leq \frac{\mathbb{E}\big[|g(Z_n)|^{1+\varepsilon}\big]}{x^{1+\varepsilon}},$$

donc (iii) implique (i). De même, la condition (ii) entraîne clairement la condition (i). Supposons (i). Alors, on écrit

$$\mathbb{E}\left[|g(Z_n)|\right] = \int_0^{+\infty} \mathbb{P}\left[|g(Z_n)| \geq x\right] dx.$$

Par hypothèse, la convergence en loi $Z_n \xrightarrow{d} Z$ entraîne $|g(Z_n)| \xrightarrow{d} |g(Z)|$ par continuité de $|g(\bullet)|$, et donc

$$\mathbb{P}\left[|g(Z_n)| \geq x\right] \to \mathbb{P}\left[|g(Z)| \geq x\right]$$

pour presque tout x. L'hypothèse (i) rend légitime le passage à la limite sous le signe somme :

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[|g(Z_n)|\right] &= \lim_{n\to\infty} \int_0^{+\infty} \mathbb{P}\left[|g(Z_n)| \geq x\right] dx \\ &= \int_0^{+\infty} \mathbb{P}\left[|g(Z)| \geq x\right] dx \\ &= \mathbb{E}\left[|g(Z)|\right].\end{aligned}$$

□

6.2.3 Risque quadratique : le cas multidimensionnel⋆

Si $\vartheta \in \Theta \subset \mathbb{R}^d$ avec $d \geq 1$, un estimateur $\widehat{\vartheta}_n$ de $\vartheta = (\vartheta_1, \ldots, \vartheta_d)^T$ s'écrit sous forme vectorielle

$$\widehat{\vartheta}_n = (\widehat{\vartheta}_n^{(1)}, \ldots, \widehat{\vartheta}_n^{(d)})^T,$$

où $\widehat{\vartheta}_n^{(j)}$ est la j-ème composante de $\widehat{\vartheta}_n$. Considérons dans un premier temps le risque quadratique de $\widehat{\vartheta}_n$ au point ϑ composante par composante, c'est-à-dire $\mathcal{R}(\widehat{\vartheta}_n^{(j)}, \vartheta_j)$, pour $j = 1, \ldots, d$, ou plus généralement une combinaison linéaire

$$\sum_{j=1}^{d} \alpha_j \mathcal{R}(\widehat{\vartheta}_n^{(j)}, \vartheta_j)$$

de sorte que tous les α_j soient positifs. En particulier, pour $\alpha_j = 1$ pour tout j, on a

$$\sum_{j=1}^{d} \alpha_j \mathcal{R}(\widehat{\vartheta}_n^{(j)}, \vartheta_j) = \mathbb{E}_\vartheta\left[\| \widehat{\vartheta}_n - \vartheta\|^2\right],$$

où $\|\bullet\|$ désigne la norme euclidienne sur $\mathbb{R}^d$. Plus généralement, on peut vouloir comparer les performances relatives des estimateurs pour l'estimation de combinaisons linéaires de composantes ϑ_j de ϑ, c'est-à-dire considérer

$$\mathcal{R}\Big(\sum_{j=1}^{d} \alpha_j \widehat{\vartheta}_n^{(j)}, \sum_{j=1}^{d} \alpha_j \vartheta_j\Big) = \sum_{j,k=1}^{d} \alpha_j \alpha_k \, \mathbb{E}_\vartheta\left[\big(\widehat{\vartheta}_n^{(j)} - \vartheta_j\big)\big(\widehat{\vartheta}_n^{(k)} - \vartheta_k\big)\right].$$

Pour cela, on a besoin d'une notion de dispersion dans $\mathbb{R}^d$.

Définition 6.8. *Si Z_1 et Z_2 sont deux vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ ayant des moments d'ordre deux (c'est-à-dire $\mathbb{E}\left[\|Z_i\|^2\right] < +\infty$ pour $i = 1, 2$), on dit que la dispersion de Z_1 autour de $\alpha \in \mathbb{R}^d$ est plus petite que la dispersion de Z_2 si, pour tout $v \in \mathbb{R}^d$, on a*

$$\mathbb{E}\left[\langle Z_1 - \alpha, v\rangle^2\right] \leq \mathbb{E}\left[\langle Z_2 - \alpha, v\rangle^2\right], \tag{6.8}$$

où $\langle u, v\rangle = \sum_{i=1}^d u_i v_i$ désigne le produit scalaire euclidien sur $\mathbb{R}^d$.

Si $\alpha = \mathbb{E}\left[Z_1\right] = \mathbb{E}\left[Z_2\right]$, l'inégalité (6.8) exprime le fait que la variance de Z_1 dans n'importe quelle direction v est plus grande que la variance de Z_2 dans cette même direction.

Si $\Sigma(Z_i)$ désigne la matrice de variance-covariance de Z_i pour $i = 1, 2$, la relation (6.8) se traduit pour $\alpha = 0$ par

$$\sum_{j,k=1}^{d} \Sigma(Z_1)_{jk} v_j v_k \leq \sum_{j,k=1}^{d} \Sigma(Z_2)_{jk} v_j v_k, \quad v \in \mathbb{R}^d,$$

c'est-à-dire la matrice $\Sigma(Z_2) - \Sigma(Z_1)$ est positive. Ceci nous fournit, de la même façon qu'en dimension 1 une règle de sélection non-asymptotique.

Définition 6.9. *Un estimateur $\widehat{\vartheta}_{n,1}$ du paramètre $\vartheta \in \Theta \subset \mathbb{R}^d$ est préférable à $\widehat{\vartheta}_{n,2}$ pour le risque quadratique au point ϑ si la dispersion de $\widehat{\vartheta}_{n,1}$ autour de ϑ est plus petite que celle de $\widehat{\vartheta}_{n,2}$.*

En conséquence, si $\Sigma_i(\vartheta) = \Sigma(\widehat{\vartheta}_{n,i} - \vartheta)$ est la matrice de variance-covariance du vecteur $\widehat{\vartheta}_{n,i} - \vartheta$ pour $i = 1, 2$, dire que $\widehat{\vartheta}_{n,1}$ est préférable à $\widehat{\vartheta}_{n,2}$ implique que la matrice $\Sigma_1(\vartheta) - \Sigma_2(\vartheta)$ est positive.

On peut de même donner la règle de comparaison asymptotique suivante

Définition 6.10. *Soit $v_n > 0$ une suite telle que $\lim_{n\to\infty}$. Si $\widehat{\vartheta}_{n,1}$ et $\widehat{\vartheta}_{n,1}$ sont deux suites d'estimateurs tels que*

$$v_n\big(\widehat{\vartheta}_{n,i} - \vartheta\big) \xrightarrow{d} Z_i,$$

pour $i = 1, 2$, où les variables Z_i sont centrées et de carré intégrable, on dit que $\widehat{\vartheta}_{n,1}$ est asymptotiquement préférable à $\widehat{\vartheta}_{n,2}$ au point ϑ si la dispersion de Z_1 autour de 0 est plus petite que celle de Z_2.

Remarque 6.2. En particulier, dans le cas classique où $v_n = \sqrt{n}$ et $Z_i \sim \mathcal{N}\big(0, \Sigma_i(\vartheta)\big)$, dire que $\widehat{\vartheta}_{n,1}$ est asymptotiquement préférable à $\widehat{\vartheta}_{n,2}$ au point ϑ implique que la matrice $\Sigma_2(\vartheta) - \Sigma_1(\vartheta)$ est positive.

6.3 Modèles réguliers

6.3.1 Information de Fisher

Situation

Dans toute la suite, on se placera dans le modèle de la densité : on considère une suite d'expérience $\mathcal{E}^n$ engendrée par l'observation d'un n-échantillon

$$(X_1, \dots, X_n)$$

où la loi $\mathbb{P}_\vartheta$ des variables aléatoires X_i appartient à une famille donnée de probabilités sur $\mathbb{R}$

$$\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\}$$

dominée par une mesure σ-finie[3] μ sur $\mathbb{R}$. On note

$$f(\vartheta, x) = \frac{d\,\mathbb{P}_\vartheta}{d\mu}(x), \quad \vartheta \in \Theta,\ \ x \in \mathbb{R}$$

la densité de $\mathbb{P}_\vartheta$ par rapport à μ. C'est une fonction positive, définie μ-presque partout, μ-intégrable, et si $X \sim \mathbb{P}_\vartheta$, on a la formule d'intégration (c'est la formule (1.1) de la mesure image, voir Chapitre 1)

$$\mathbb{E}_\vartheta\left[\varphi(X)\right] = \int_{\mathbb{R}} \varphi(x)\,\mathbb{P}_\vartheta(dx) = \int_{\mathbb{R}} \varphi(x) f(\vartheta, x)\mu(dx)$$

pour toute fonction test φ. On introduit aussi la notation suivante

Définition 6.11. *On pose, lorsque cela a un sens*

$$\ell(\vartheta, x) = \log f(\vartheta, x), \ \ x \in \mathbb{R}, \vartheta \in \Theta.$$

(En convenant $\log 0 = 0$ *par exemple, on pourra toujours parler de* $\ell(\vartheta, x)$*). La dérivée de la fonction* $\vartheta \rightsquigarrow \ell(\vartheta, x)$*, lorsqu'elle existe, s'appelle « fonction score » du modèle* $\left\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\right\}$.

Information de Fisher d'une famille de densités

Restreignons nous dans un premier temps au cas où $\Theta \subset \mathbb{R}$ pour simplifier.

[3]Dans presque tous les cas, μ sera la mesure de Lebesgue, dans le cas où les $\mathbb{P}_\vartheta$ sont absolument continues, ou bien la mesure de comptage sur un ensemble $\mathcal{M} \subset \mathbb{R}$ au plus dénombrable lorsque les X_i sont discrètes, à valeurs dans $\mathcal{M}$.

Définition 6.12 (Information de Fisher)**.** *Si $\vartheta \rightsquigarrow \ell(\vartheta, x)$ est dérivable $\mu(dx)$-presque partout, on appelle information de Fisher de la famille $\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\}$ au point $\vartheta \in \Theta$ la quantité*

$$\mathbb{I}(\vartheta) = \int_{\mathbb{R}} \big(\partial_\vartheta\, \ell(\vartheta, x)\big)^2 f(\vartheta, x)\mu(dx) = \mathbb{E}_\vartheta\left[\big(\partial_\vartheta\, \ell(\vartheta, X)\big)^2\right].$$

On a, pour tout $\vartheta \in \Theta$,

$$\mathbb{I}(\vartheta) = \int_{\{x,\, f(\vartheta,x)>0\}} \frac{\big(\partial_\vartheta f(\vartheta, x)\big)^2}{f(\vartheta, x)}\mu(dx),$$

et aussi

$$0 \leq \mathbb{I}(\vartheta) \leq +\infty,$$

les cas intéressants étant ceux pour lesquels on a

$$0 < \mathbb{I}(\vartheta) < +\infty.$$

Origine de l'information de Fisher

L'information de Fisher apparaît naturellement comme la variance limite de l'estimateur du maximum de vraisemblance, sous des hypothèses suffisantes de régularité sur $\{f(\vartheta, \bullet), \vartheta \in \Theta\}$. Cela signifie que l'on a

$$\sqrt{n}\big(\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} - \vartheta\big) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathbb{I}(\vartheta)}\right). \tag{6.9}$$

Donnons immédiatement l'heuristique de ce résultat, sans nous soucier des hypothèses, que nous préciserons plus loin. On va essentiellement répéter la preuve de la Proposition 4.5 du Chapitre 4 dans ce contexte particulier. Rappelons que d'après l'équation (4.21) du Chapitre 4, l'estimateur $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$ satisfait

$$\partial_\vartheta \ell_n(\vartheta)_{|\vartheta=\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}} = 0,$$

où

$$\ell_n(\vartheta) = \sum_{i=1}^n \ell(\vartheta, X_i) = \sum_{i=1}^n \log f(\vartheta, X_i)$$

est la log-vaisemblance associée à la famille $\big\{\,\mathbb{P}_\vartheta, \vartheta \in \Theta\big\}$. Au voisinage de $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$, on a, à l'ordre 1,

$$0 = \partial_\vartheta \ell_n(\vartheta)_{|\vartheta=\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}} \approx \partial_\vartheta \ell_n(\vartheta) + \big(\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} - \vartheta\big)\partial_\vartheta^2 \ell_n(\vartheta).$$

En divisant par $\partial_\vartheta^2 \ell_n(\vartheta)$ et en mutipliant par $\frac{1}{\sqrt{n}}$, on obtient l'approximation

$$\sqrt{n}\big(\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} - \vartheta\big) \approx \frac{n^{-1/2}\partial_\vartheta\, \ell_n(\vartheta)}{-n^{-1}\partial_\vartheta^2\, \ell_n(\vartheta)}.$$

C'est l'étude asymptotique du numérateur et du dénominateur respectivement qui va faire apparaître $\mathbb{I}(\vartheta)$. Notons que

$$n^{-1/2}\,\partial_\vartheta\,\ell_n(\vartheta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial_\vartheta \log f(\vartheta, X_i).$$

et

$$-n^{-1}\,\partial_\vartheta^2\,\ell_n(\vartheta) = -\frac{1}{n}\sum_{i=1}^{n}\partial_\vartheta^2 \log f(\vartheta, X_i).$$

Sous des conditions d'intégrabilité suffisantes, le dénominateur converge par la loi des grands nombres vers

$$-\,\mathbb{E}_\vartheta\left[\partial_\vartheta^2 \log f(\vartheta, X)\right]$$

en probabilité. Le comportement du numérateur $\frac{1}{\sqrt{n}}\partial_\vartheta \ell_n(\vartheta)$ est moins évident. Nous allons d'abord énoncer un lemme fondamental sur lequel nous reviendrons plus tard.

Lemme 6.3.1. *Sous des hypothèses de régularité adéquates, on a*

$$\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, X)\right] = 0.$$

Démonstration. Justifions formellement ce résultat : on a

$$\begin{aligned}
\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, X)\right] &= \int_{\mathbb{R}} \partial_\vartheta \log f(\vartheta, x)\, f(\vartheta, x)\mu(dx) \\
&= \int_{\mathbb{R}} \frac{\partial_\vartheta f(\vartheta, x)}{f(\vartheta, x)} f(\vartheta, x)\mu(dx) \\
&= \int_{\mathbb{R}} \partial_\vartheta f(\vartheta, x)\mu(dx) \\
&= \partial_\vartheta \int_{\mathbb{R}} f(\vartheta, x)\mu(dx) = \partial_\vartheta 1 = 0.
\end{aligned}$$

□

On a aussi $\int_{\mathbb{R}} \partial_\vartheta^2 f(\vartheta, x)\mu(dx) = 0$, ce qui permet de déduire la relation très utile pour les calculs

$$\mathbb{I}(\vartheta) = \mathbb{E}_\vartheta\left[\big(\partial_\vartheta \log f(\vartheta, X)\big)^2\right] = -\,\mathbb{E}_\vartheta\left[\partial_\vartheta^2 \log f(\vartheta, X\right]. \tag{6.10}$$

Revenons à l'étude du numérateur $\frac{1}{\sqrt{n}}\partial_\vartheta \ell_n(\vartheta)$. D'après le Lemme 6.3.1, les variables aléatoires $\partial_\vartheta \log f(\vartheta, X_i)$ sont indépendantes, centrées, de variance $\mathbb{I}(\vartheta)$. D'après le théorème central-limite, on a la convergence

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial_\vartheta \log f(\vartheta, X_i) \xrightarrow{d} \mathcal{N}\big(0, \mathbb{I}(\vartheta)\big).$$

On a déja vu que le dénominateur $-\frac{1}{n}\partial^2_\vartheta \ell_n(\vartheta)$ converge en probabilité vers

$$-\,\mathbb{E}_\vartheta\left[\partial^2_\vartheta(\vartheta, X)\right] = \mathbb{I}(\vartheta)$$

d'après la formule (6.10). On en déduit par la Proposition 1.8 (Slutsky) que le quotient converge en loi vers une gaussienne centrée de variance $\mathbb{I}(\vartheta)^{-1}$, c'est-à-dire

$$\sqrt{n}\big(\,\widehat{\vartheta}^{\,\mathtt{mv}}_{\mathrm{n}} - \vartheta\big) \approx \mathcal{N}\left(0, \frac{1}{\mathbb{I}(\vartheta)}\right),$$

et nous pouvons donc interpréter $\mathbb{I}(\vartheta)$ comme l'inverse de la variance asymptotique de l'estimateur du maximum de vraisemblance.

La suite de cette section va consister à rendre rigoureux ce raisonnement, à le généraliser au cas où ϑ est de dimension $d \geq 1$ et à montrer que $\mathbb{I}(\vartheta)$ est une caractéristique « géométrique » de la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$, apparentée à une notion d'information intrinsèque de l'expérience statistique associée. Ce sera un premier pas vers une notion de comparaison des expériences statistiques d'une part, et de la meilleure estimation possible d'autre part.

Information de Fisher d'une (suite d') expérience(s) statistique(s)

L'information de Fisher introduite dans la Définition 6.12 de la Section 6.3.1 porte sur une famille $\big\{f(\vartheta, \bullet),\, \vartheta \in \Theta\big\}$ de densités $(\vartheta, x) \in \Theta \times \mathbb{R} \to \mathbb{R}_+$ avec $\Theta \subset \mathbb{R}$. L'extension de cette notion pour une experience statistique dominée arbitraire – en se restreignant toujours au cas $\Theta \subset \mathbb{R}$ – est immédiate :

Définition 6.13. *Si $\mathcal{E}^n = \big(\mathfrak{Z}_n, \mathcal{Z}_n, \big\{\,\mathbb{P}^n_\vartheta,\, \vartheta \in \Theta\big\}\big)$ est une suite d'expériences statistiques dominée par une mesure $\mu_n(dz)$ σ-finie sur $(\mathfrak{Z}_n, \mathcal{Z}_n)$ et si $\Theta \subset \mathbb{R}$, alors l'information de Fisher de l'expérience au point $\vartheta \in \Theta$ est définie par*

$$\mathbb{I}(\vartheta \,|\, \mathcal{E}^n) = \int_{\mathfrak{Z}_n} \Big(\partial_\vartheta \log f_n(\vartheta, z)\Big)^2\, \mathbb{P}^n_\vartheta(dz), \tag{6.11}$$

où $f_n(\vartheta, z) = \frac{d\,\mathbb{P}^n_\vartheta}{d\mu}(z)$ pour peu que l'expression ci-dessus soit bien définie.

En particulier, si l'expérience statistique considérée est engendrée par un n-échantillon de loi $\{\mathbb{P}_\vartheta,\, \vartheta \in \Theta\}$ sur $\mathbb{R}$ dominée par un mesure μ sur $\mathbb{R}$, alors on a

$$\mathcal{E}^n = \big(\,\mathbb{R}^n, \mathcal{B}^n, \{\mathbb{P}^n_\vartheta,\, \vartheta \in \Theta\}\big),$$

avec $\mathbb{P}^n_\vartheta = \mathbb{P}_\vartheta \otimes \cdots \otimes \mathbb{P}_\vartheta$ (n-fois), $\mu_n = \mu \otimes \cdots \otimes \mu$ (n-fois), et

$$f_n(z) = f_n(x_1, \ldots, x_n) = \prod_{i=1}^n f(\vartheta, x_i), \quad z = (x_1, \ldots, x_n) \in \mathfrak{Z} = \mathbb{R}^n,$$

où $f(\vartheta, x) = \frac{d\mathbb{P}_\vartheta}{d\mu}(x)$ est la densité pour la famille de lois de probabilités sur $\mathbb{R}$. On déduit immédiatement de la formule (6.11) l'identité :

$$\mathbb{I}(\vartheta \,|\, \mathcal{E}^n) = n\,\mathbb{I}(\vartheta) = n\,\mathbb{I}(\vartheta \,|\, \mathcal{E}^1), \tag{6.12}$$

où $\mathbb{I}(\vartheta)$ est l'information de Fisher pour la famille $\big\{f(\vartheta, \bullet),\, \vartheta \in \Theta\big\}$ de la Définition 6.12.

Remarque 6.3. La formule (6.12) s'interprète de la manière suivante : pour un n-échantillon, chaque donnée X_i contribue à l'information totale du modèle au point ϑ pour une quantité $\mathbb{I}(\vartheta)$. L'information totale, après n observations, est n fois l'information qu'apporte chaque donnée. Voir la Section 6.3.3.

6.3.2 Modèle régulier en dimension 1

Nous avons vu dans la Section 4.3 du Chapitre 4 que l'estimateur du maximum de vraisemblance est un M-estimateur associé au constraste $\psi(a, x) = \log f(a, x)$ ou bien un Z-estimateur associé au « score »

$$\phi(a, x) = \partial_a \psi(a, x) = \frac{\partial_a f(a, x)}{f(a, x)},$$

pour peu que ces quantités soient bien définies et régulières[4].

Nous allons donner un jeu d'hypothèses le plus simple possible, de sorte que les calculs de la Section 6.3.1 développés précédemment soient justifiés, en particulier le Lemme 6.3.1, et que l'on puisse appliquer les Propositions 4.5 ou 4.6 qui fournissent le comportement asymptotique des Z- ou M-estimateurs.

Hypothèse 6.1 (Régularité d'un modèle (ou d'une famille))**.** *On a*

(i) L'ensemble des paramètre $\Theta \subset \mathbb{R}$ est un intervalle ouvert et pour tous $\vartheta, \vartheta' \in \Theta$, les ensembles $\{f(\vartheta, \bullet) > 0\}$ et $\{f(\vartheta', \bullet) > 0\}$ coïncident.

(ii) Les fonctions $\vartheta \rightsquigarrow f(\vartheta, \bullet)$ et $\vartheta \rightsquigarrow \ell(\vartheta, \bullet)$ sont trois fois continûment différentiables sur Θ.

(iii) Pour tout $\vartheta \in \Theta$, il existe un voisinage de $\mathcal{V}(\vartheta) \subset \Theta$ tel que pour tout $a \in \mathcal{V}(\vartheta)$:

$$|\partial_a^3 \ell(a, x)| + |\partial_a^2 \ell(a, x)| + |\partial_a \ell(a, x)| + \partial_a \ell(a, x)^2 \leq g(x),$$

où

$$\int_{\mathbb{R}} g(x)\mu(dx) < +\infty.$$

(iv) L'information de Fisher est non-dégénérée :

$$\forall \vartheta \in \Theta,\ \ \mathbb{I}(\vartheta) > 0.$$

[4]et on suppose toujours implicitement que la famille $\big\{\mathbb{P}_\vartheta,\, \vartheta \in \Theta\big\}$ est dominée par une mesure σ-finie μ sur $\mathbb{R}$, de sorte que l'on puisse parler de la famille des densités $\big\{f(\vartheta, \bullet),\, \vartheta \in \Theta\big\}$.

Les hypothèses (ii) et (iii) sont les plus restrictives. On peut significativement les améliorer. Une référence accessible est van der Waart [7].

Définition 6.14. *On dit que la famille de densités* $\{f(\vartheta,\bullet),\vartheta\in\Theta\}$ *est régulière si l'Hypothèse 6.1 est vérifiée. Par extension, l'expérience statistique* $\mathcal{E}$ *(ou* $\mathcal{E}^n$*) est régulière si elle est dominée et que la famille de densités associées est régulière.*

6.3.3 Propriétés de l'information de Fisher

Information de Fisher et maximum de vraisemblance

L'estimateur du maximum de vraisemblance est un M-estimateur, associé à la fonction

$$a\rightsquigarrow\mathcal{F}(a,\vartheta)=\mathbb{E}_\vartheta\left[\psi(a,X)\right],$$

où ψ est la fonction de contraste :

$$\psi(a,x)=\log f(a,x).$$

Proposition 6.3. *Si la famille* $\{f(\vartheta,\bullet),\vartheta\in\Theta\}$ *est régulière et si*

$$\forall\vartheta,\in\Theta,\quad\int_{\mathbb{R}}\big|\log f(\vartheta,x)\big|f(\vartheta,x)\mu(dx)<+\infty,$$

alors, pour tout $\vartheta\in\Theta$*, la fonction* $a\rightsquigarrow\mathcal{F}(\vartheta,a)$ *est deux fois continûment dérivable et on a*

$$\partial_a\mathcal{F}(a,\vartheta)_{\big|\,a=\vartheta}=0,$$

et

$$\partial_a^2\mathcal{F}(a,\vartheta)_{\big|\,a=\vartheta}=-\,\mathbb{E}_\vartheta\left[\partial_\vartheta^2\,\ell(\vartheta,X)\right].$$

Le lemme technique suivant permet de justifier la dérivation sous le signe somme

Lemme 6.3.2. *Soit* $g:\Theta\times\mathbb{R}\to\mathbb{R}$ *telle que* $a\rightsquigarrow g(a,x)$ *soit continûment différentiable* $\mu(dx)$ *presque-partout. Si, de plus, pour un ouvert* $\mathcal{U}$ *de* Θ*, et pour tout* $a\in\mathcal{U}$

$$\int_{\mathbb{R}}\big|g(a,x)\big|\mu(dx)<+\infty,\quad et\quad\int_{\mathbb{R}}\sup_{a\in\mathcal{U}}\big|\partial_a\,g(a,x)\big|\mu(dx)<+\infty$$

alors la fonction $a\rightsquigarrow G(a)=\int_{\mathbb{R}}g(a,x)\mu(dx)$ *est continûment différentiable sur* $\mathcal{U}$ *et*

$$G'(a)=\frac{d}{da}\int_{\mathbb{R}}g(a,x)\mu(dx)=\int_{\mathbb{R}}\partial_a\,g(a,x)\mu(dx).$$

Démonstration. C'est une application répétée du théorème de convergence dominée. □

Démonstration de la Proposition 6.3. L'application $a \rightsquigarrow \mathcal{F}(a,\vartheta)$ est dérivable en appliquant le Lemme 6.3.2 avec $g(a,x) = f(\vartheta,x)\log f(a,x)$. On obtient :

$$\partial_a \mathcal{F}(a,\vartheta)_{\big|a=\vartheta} = -\int_{\mathbb{R}} \partial_\vartheta \ell(\vartheta,x) f(\vartheta,x)\mu(dx).$$

On sait déja par le Lemme 4.4.1 du Chapitre 4 que le minimum de $\mathcal{F}(\bullet,\vartheta)$ est atteint en $a=\vartheta$, donc $\partial_a \mathcal{F}(a,\vartheta)_{\big|a=\vartheta} = 0$. Pour la deuxième égalité, on applique le Lemme 6.3.2 à $G(a) = \partial_a \mathcal{F}(a,\vartheta)$ en posant cette fois-ci $g(a,x) = -\partial_a^2 \ell(a,x) f(\vartheta,x)$. □

Nous allons maintenant démontrer rigoureusement l'identité 6.10 de la Section 6.3.1.

Lemme 6.3.3. *Si la famille $\{f(\vartheta,\bullet), \vartheta\in\Theta\}$ est régulière, alors, pour tout $\vartheta\in\Theta$, on a*

$$\mathbb{I}(\vartheta) = -\,\mathbb{E}_\vartheta\left[\partial_\vartheta^2\,\ell(\vartheta,X)\right] = -\int_{\mathbb{R}} \partial_\vartheta^2\,\ell(\vartheta,x) f(\vartheta,x)\mu(dx).$$

En particulier, on en déduit, sous les hypothèses de la Proposition 6.3

$$\partial_a^2 \mathcal{F}(a,\vartheta)_{\big|a=\vartheta} = \mathbb{I}(\vartheta).$$

Démonstration. On dérive deux fois sous le signe somme l'égalité

$$\int_{\mathbb{R}} f(\vartheta,x)\mu(dx) = 1.$$

On applique d'abord le Lemme 6.3.2 avec $g(\vartheta,x) = f(\vartheta,x)$. On en déduit, pour tout $\vartheta\in\Theta$

$$\int_{\mathbb{R}} \partial_\vartheta f(\vartheta,x)\mu(dx) = 0,$$

ou encore

$$\int_{\mathbb{R}} \partial_\vartheta \ell(\vartheta,x) f(\vartheta,x)\mu(dx) = 0.$$

On applique le Lemme 6.3.2 une seconde fois, avec $g(\vartheta,x) = \partial_\vartheta f(\vartheta,x) = \partial_\vartheta \ell(\vartheta,x) f(\vartheta,x)$. Alors

$$\partial_\vartheta g(\vartheta,x) = \partial_\vartheta^2 \ell(\vartheta,x) f(\vartheta,x) + \left(\partial_\vartheta\,\ell(\vartheta,x)\right)^2 f(\vartheta,x).$$

Cette identité permet de conclure

$$0 = \int_{\mathbb{R}} \partial_\vartheta g(\vartheta,x)\mu(dx) = \int_{\mathbb{R}} \partial_\vartheta^2 \ell(\vartheta,x) f(\vartheta,x)\mu(dx) + \mathbb{I}(\vartheta),$$

d'où le résultat. □

6.3.4 Interprétation géométrique de l'information de Fisher

Pour une expérience statistique régulière, la Proposition 6.3 ainsi que le Lemme 6.3.3 donnent la représentation

$$\mathbb{I}(\vartheta) = \partial_a^2 \mathcal{F}(a,\vartheta)_{\big|a=\vartheta} \geq 0,$$

et la fonction $a \rightsquigarrow \mathcal{F}(a,\vartheta)$ atteint son maximum au point $a = \vartheta$.

Si $\mathbb{I}(\vartheta)$ est petite, le rayon de courbure de la courbe représentative de $a \rightsquigarrow \mathcal{F}(a,\vartheta)$ est grand dans un voisinage de ϑ, et $\mathcal{F}(\bullet,\vartheta)$ est « plate » dans ce voisinage, et le comportement typique de $a \rightsquigarrow \ell_n(a)$ sera oscillant, rendant moins précis l'estimateur du maximum de vaisemblance. Par contre, si $\mathbb{I}(\vartheta)$ est grande, $\mathcal{F}(\bullet,\vartheta)$ est « pointue » dans un voisinage de ϑ.

Lien avec l'entropie

Si $\mathbb{P}$ et $\mathbb{Q}$ sont deux mesures de probabilités définies sur un même espace mesurable $(\Omega, \mathcal{A})$, on définit la divergence de Kullback-Leibler de $\mathbb{P}$ relativement à $\mathbb{Q}$ comme

$$K(\mathbb{P},\mathbb{Q}) = \int_\Omega \log \frac{d\,\mathbb{P}}{d\mathbb{Q}}(\omega)\, \mathbb{P}(d\omega)$$

si $\mathbb{P} \ll \mathbb{Q}$ ($\mathbb{Q}$ domine $\mathbb{P}$) et on pose $K(\mathbb{P},\mathbb{Q}) = +\infty$ sinon. On parle improprement de « distance » de Kullback-Leibler entre $\mathbb{P}$ et $\mathbb{Q}$ pour la raison suivante :

Lemme 6.3.4. *On a toujours*

$$0 \leq K(\mathbb{P},\mathbb{Q}) \leq +\infty,$$

et

$$K(\mathbb{P},\mathbb{Q}) = 0 \quad \textit{si et seulement si} \quad \mathbb{P} = \mathbb{Q}.$$

Démonstration. Introduisons la fonction définie sur $\mathbb{R}_+$ par

$$h(x) = x \log(x).$$

Si Z est une variable aléatoire positive telle que $\mathbb{E}_\mathbb{Q}\big[Z\big] < +\infty$ (espérance de Z par rapport à la mesure de probabilité $\mathbb{Q}$), on peut toujours définir la quantité

$$\mathcal{E}\big[Z\big] = \mathbb{E}_\mathbb{Q}\big[h(Z)\big] - h\big(\mathbb{E}_\mathbb{Q}\big[Z\big]\big),$$

En effet, h est minorée par $-1/e$, donc $\mathbb{E}_\mathbb{Q}\big[h(Z)\big]$ a un sens, même si $h(Z)$ n'est pas $\mathbb{Q}$-intégrable. Puisque h est convexe, l'inégalité de Jensen assure que $\mathcal{E}_\mathbb{Q}[Z] \geq 0$ (éventuellement $+\infty$). Enfin, $\mathcal{E}\big[Z\big]$ est finie si et seulement si $h(Z)$ est $\mathbb{Q}$-intégrable.

Supposons maintenant $\mathbb{P} \ll \mathbb{Q}$, et posons $Z = \frac{d\mathbb{P}}{d\mathbb{Q}}$, la densité de Radon-Nikodym[5] de $\mathbb{P}$ par rapport à $\mathbb{Q}$. Alors Z est $\mathbb{Q}$-intégrable et $\mathbb{E}_{\mathbb{Q}}\big[Z\big] = 1$. Il vient

$$\mathcal{E}\big[Z\big] = \mathbb{E}_{\mathbb{Q}}\big[h(Z)\big] = \int \frac{d\,\mathbb{P}}{d\mathbb{Q}} \log \frac{d\,\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} = K(\mathbb{P}, \mathbb{Q}),$$

d'où la première partie du lemme.

La seconde partie du lemme est la cas d'égalité dans l'inégalité de Jensen : puisque h est strictement convexe et $\mathcal{E}[Z] \geq 0$, si $K(\mathbb{P}, \mathbb{Q}) = 0$, alors $Z = 1$ $\mathbb{Q}$ presque-sûrement et ceci entraîne $\mathbb{P} = \mathbb{Q}$. □

Dans le contexte d'un modèle régulier, on a, pour $\vartheta, \widetilde{\vartheta} \in \Theta$

$$K\big(\mathbb{P}_\vartheta, \mathbb{P}_{\widetilde{\vartheta}}\big) = \mathcal{F}(\widetilde{\vartheta}, \widetilde{\vartheta}) - \mathcal{F}(\vartheta, \widetilde{\vartheta}) = \int_{\mathbb{R}} \log \frac{f(\widetilde{\vartheta}, x)}{f(\vartheta, x)} f(\widetilde{\vartheta}, x)\mu(dx).$$

C'est une mesure de divergence disymétrique entre $\mathbb{P}_\vartheta$ et $\mathbb{P}_{\widetilde{\vartheta}}$. Son interprétation est similaire à celle de l'information de Fisher, comme le montre la représentation ci-dessus. L'avantage immédiat de la divergence de Kullback-Leibler sur l'information de Fisher est qu'elle est toujours définie, sans hypothèse de régularité sur la famille $\big\{\mathbb{P}_\vartheta,\, \vartheta \in \Theta\big\}$ sous-jacente.

Définition 6.15. *La valeur*

$$-\mathcal{F}(\vartheta, \vartheta) = -\int_{\mathbb{R}} f(\vartheta, x) \log f(\vartheta, x)\mu(dx)$$

est appelée entropie de Shannon associée la densité $f(\vartheta, \bullet)$.

L'entropie de Shannon peut-être utilisée comme mesure de dispersion lorsque, par exemple, la variance par rapport à la loi $f(\vartheta, x)\mu(dx)$ n'existe pas. Elle a un lien avec la théorie de l'information.

6.3.5 Le cas multidimensionnel

Si $\Theta \subset \mathbb{R}^d$ avec $d > 1$, tous les résultats de la Section précédente s'étendent de manière naturelle en remplaçant dérivation par rapportà ϑ par différentiabilité dans $\mathbb{R}^d$. L'information de Fisher devient la matrice d'information de Fisher.

Définition 6.16. *La matrice d'information de Fisher* $\mathbb{I}(\vartheta) = \big(\mathbb{I}(\vartheta)_{\ell,\ell'}\big)_{1\leq \ell,\ell' \leq d}$ *associée à la famille de densités* $\{f(\vartheta),\, \vartheta \in \Theta\}$ *avec* $\vartheta \in \Theta \subset \mathbb{R}^d$ *est définie au point* ϑ *par*

$$\mathbb{I}(\vartheta)_{\ell,\ell'} = \mathbb{E}_\vartheta\big[\partial_{\vartheta_\ell} \log f(\vartheta, X) \partial_{\vartheta_{\ell'}} \log f(\vartheta, X)\big], \quad 1 \leq \ell, \ell' \leq d.$$

pour peu que cette quantité soit bien définie, avec $\vartheta = (\vartheta_1, \ldots, \vartheta_d)^T$. *C'est une matrice symétrique positive.*

[5] voir, par exemple, Jacod et Protter [3], Chapitre 28.

Nous ne développerons pas la théorie en dimension plus grande que 1. Une référence avec des exemples détaillés est Borovkov [1].

6.4 Théorie asymptotique

6.4.1 Normalité asymptotique du maximum de vraisemblance

Le cas de la dimension 1

On considère l'expérience $\mathcal{E}^n$ engendré par un n-échantillon de loi $\mathbb{P}_\vartheta$, où la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ est dominée par une mesure μ sur $\mathbb{R}$ σ-finie, et on suppose $\Theta \subset \mathbb{R}$. Le résultat suivant donne le comportement asymptotiue de l'estimateur du maximum de vraisemblance. Le jeu d'hypothèses très large que nous faisons nous permet de nous ramener aux résultats du Chapitre 4 pour les Z- et M-estimateurs.

Proposition 6.4 (Normalité asymptotique de l'EMV)**.** *Si l'expérience $\mathcal{E}^n$ est régulière au sens de la Définition 6.14, alors l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$ est bien défini et asymptotiquement normal, et on a*

$$\sqrt{n}\big(\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} - \vartheta\big) \xrightarrow{\;d\;} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\vartheta)}\Big)$$

en loi sous $\mathbb{P}_\vartheta$, et $0 < \mathbb{I}(\vartheta) < +\infty$ est l'information de Fisher du modèle au point ϑ.

Démonstration. En interprétant l'estimateur du maximum de vraisemblance comme un M-estimateur, nous appliquons la Proposition 4.6 du Chapitre 4 pour la fonction de constraste $\psi(a, x) = \log f(a, x)$, ce qui nous conduit à vérifier les conditions de l'Hypothèse 4.2 pour appliquer la Proposition 4.5 à la fonction $\phi(a, x) = \partial_a \log f(a, x)$ pour nous ramener au comportement asymptotique des Z-estimateurs. □

Le cas multidimensionnel

La Proposition 6.4 s'étend au cas multdimensionnel, en remplaçant l'information de Fisher par la matrice d'information de Fisher définie dans la Section 6.3.5, en étendant l'Hypothèse 6.1 par une version multidimensionelle (la dérivée première par rapport à ϑ de la fonction $\vartheta \rightsquigarrow f(\vartheta, \bullet)$ devenant le gradient et la dérivée seconde la matrice hessienne). Nous ne développerons pas la théorie en dimension plus grande que 1. Une référence avec des exemples détaillés est Borovkov [1].

6.4.2 Comparaison d'estimateurs : efficacité asymptotique

Nous nous plaçons dans cette section dans le cas de la dimension 1, avec $\Theta \subset \mathbb{R}$ pour simplifier. Les extensions au cas multidimensionnel se font de la même manière que

pour la Section 6.3.5. On se restreint ici à la classe des estimateurs asymptotiquement normaux, c'est-à-dire les estimateurs $\widehat{\vartheta}_n$ pour lesquels

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v(\vartheta)\big)$$

pour $\vartheta \in \Theta$. On fait l'hypothèse supplémentaire

Hypothèse 6.2. *L'application $\vartheta \rightsquigarrow v(\vartheta)$ est continue et strictement positive sur Θ.*

Sous des hypothèses de régularité, on a vu que les M-estimateurs sont asymptotiquement normaux et vérifient 6.2. En particulier, pour l'estimateur du maximum de vraisemblance,

$$v(\vartheta) = \frac{1}{\mathbb{I}(\vartheta)}.$$

On a la règle de comparaison suivante

Définition 6.17. *Si $\widehat{\vartheta}_{n,1}$ et $\widehat{\vartheta}_{n,2}$ sont deux (suites d')estimateurs asymptotiquement normaux de variances asymptotiques respectives $v_1(\vartheta)$ et $v_2(\vartheta)$ et vérifiant l'Hypothèse 6.2, on dit que $\widehat{\vartheta}_{n,1}$ est plus efficace que $\widehat{\vartheta}_{n,2}$ si*

$$\forall \vartheta \in \Theta, \;\; v_1(\vartheta) \leq v_2(\vartheta)$$

et si de plus, il existe un point $\widetilde{\vartheta} \in \Theta$ tel que

$$v_1(\widetilde{\vartheta}) < v_2(\widetilde{\vartheta}).$$

Une suite d'estimateurs $\widehat{\vartheta}_n$ est asymptotiquement efficace s'il n'existe pas d'autre estimateurs (dans la classe considérée) plus efficace que $\widehat{\vartheta}_n$.

Remarque 6.4. L'hypothèse de normalité asymptotique en tout point $\vartheta \in \Theta$ permet en particulier d'exclure les estimateurs artificiels de la forme $\widehat{\vartheta}_n = \vartheta_0$ pour un point $\vartheta_0 \in \Theta$ arbitraire, qui sont catastrophiques pour le risque quadratique en dehors d'un « petit » voisinage de ϑ_0 mais qui ont un risque nul en ϑ_0.

Efficacité asymptotique du maximum de vraisemblance

Dans cette section, on considère une expérience statistique régulière et on suppose l'espace des paramètres $\Theta \subset \mathbb{R}$ pour simplifier. On se restreint en fait à la classe des Z-estimateurs, qui contient en particulier les M-estimateurs réguliers.

Un tel estimateur $\widehat{\vartheta}_n$ est obtenu comme solution d'une équation de type

$$\sum_{i=1}^{n} \phi(\widehat{\vartheta}_n, X_i) = 0 \tag{6.13}$$

où $\phi : \Theta \times \mathbb{R}$ est une fonction à choisie par le statisticien, qui détermine la méthode. En particulier, si

$$\phi(\vartheta, x) = \partial_\vartheta \log f(\vartheta, x) = \partial_\vartheta \ell(\vartheta, x)$$

dans le cas d'une famille de probabilités $\{\mathbb{P}_\vartheta(dx) = f(\vartheta, x)\mu(dx), \vartheta \in \Theta\}$ dominée par une mesure σ-finie μ, on retrouve l'estimateur du maximum de vraisemblance.

On considère une expérience statistique régulière engendrée par l'observation d'un n-échantillon

Théorème 6.1 (Efficacité asymptotique du maximum de vraisemblance parmi la classe des Z-estimateurs)**.** *Si $\widehat{\vartheta}_n$ est un Z-estimateur régulier associé à la fonction ϕ via* (6.13)*, alors $\widehat{\vartheta}_n$ est asymptotiquement normal de variance asymptotique*

$$v_\phi(\vartheta) = \frac{\mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right]}{\Big(\mathbb{E}_\vartheta\left[\partial_\vartheta \phi(\vartheta, X)\right]\Big)^2}.$$

De plus, pour tout choix de fonction ϕ, on a

$$v_\phi(\vartheta) \geq \frac{1}{\mathbb{I}(\vartheta)}. \tag{6.14}$$

Corollaire 6.1. *Dans un modèle régulier, l'estimateur du maximum de vraisemblance est asymptotiquement efficace parmi les Z-estimateurs réguliers.*

Démonstration. La première partie du théorème a déjà été montrée en 4.6. Montrons (6.14). On note $\phi'(\vartheta, x) = \partial_\vartheta \phi(\vartheta, x)$. Par construction la fonction ϕ vérifie

$$\partial_a \, \mathbb{E}_\vartheta\left[\phi(a, X)\right]_{\big|a=\vartheta} = 0,$$

ce qui s'écrit encore

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \phi'(\vartheta, x) f(\vartheta, x)\mu(dx) + \int_{\mathbb{R}} \phi(\vartheta, x)\partial_\vartheta f(\vartheta, x)\mu(dx) \\ &= \int_{\mathbb{R}} \phi'(\vartheta, x) f(\vartheta, x)\mu(dx) + \int_{\mathbb{R}} \phi(\vartheta, x)\partial_\vartheta \, \ell(\vartheta, x) f(\vartheta, x)\mu(dx), \end{aligned}$$

c'est-à-dire

$$\mathbb{E}_\vartheta\left[\phi'(\vartheta, X)\right] = -\,\mathbb{E}_\vartheta\left[\phi(\vartheta, X)\partial_\vartheta \ell(\vartheta, X)\right].$$

En appliquant l'inégalité de Cauchy-Schwarz, on obtient

$$\Big(\mathbb{E}_\vartheta\left[\phi'(\vartheta, X)\right]\Big)^2 \leq \mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right] \mathbb{E}_\vartheta\left[\big(\partial_\vartheta \ell(\vartheta, X)\big)^2\right],$$

c'est-à-dire

$$v_\phi(\vartheta) = \frac{\mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right]}{\Big(\mathbb{E}_\vartheta\left[\partial_\vartheta \phi(\vartheta, X)\right]\Big)^2} \leq \mathbb{E}_\vartheta\left[\big(\partial_\vartheta \, \ell(\vartheta, X)\big)^2\right] = \mathbb{I}(\vartheta).$$

□

Efficacité à un pas*

Dans un modèle régulier, l'estimateur du maximum de vraisemblance est « meilleur » que n'importe quel autre M-estimateur au sens de l'efficacité asymptotique. Pourtant, il est parfois plus facile de mettre en oeuvre un M-estimateur donné (ou d'ailleurs un Z-estimateur) plutôt que l'estimateur du maximum de vraisemblance, voir l'Exemple 4.4 du modèle de Cauchy.

On peut modifier un estimateur $\widehat{\vartheta}_n$ consistant et asymptotiquement normal de sorte qu'il ait asymptotiquement le même comportement que l'estimateur du maximum de vraisemblance. On note $\ell_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n \log f(\vartheta, X_i)$.

Proposition 6.5 (Efficacité à un pas)**.** *Si le modèle est régulier et si $\widehat{\vartheta}_n$ est un estimateur asymptotiquement normal, alors l'estimateur modifié*[6]

$$\widetilde{\vartheta}_n = \widehat{\vartheta}_n - \frac{\ell_n'(\widehat{\vartheta}_n)}{\ell_n''(\widehat{\vartheta}_n)}$$

vérifie

$$\sqrt{n}\big(\widetilde{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\vartheta)}\Big)$$

en loi sous $\mathbb{P}_\vartheta$ et est donc asymptotiquement efficace.

Le choix initial pourra donc être un M- ou Z-estimateur consistant et asymptotiquement normal, sans que l'on ait besoin de se soucier (asymptotiquement) de sa variance asymptotique.

Esquisse de démonstration. On écrit

$$\begin{aligned}
\sqrt{n}\big(\widetilde{\vartheta}_n - \vartheta\big) &= \sqrt{n}\big(\,\widehat{\vartheta}_n - \vartheta\big) - \frac{\sqrt{n}\ell_n'(\widehat{\vartheta}_n)}{\ell_n''(\widehat{\vartheta}_n)} \\
&= \sqrt{n}\big(\,\widehat{\vartheta}_n - \vartheta\big) - \frac{\sqrt{n}\ell_n'(\vartheta) + \sqrt{n}\big(\ell_n'(\widehat{\vartheta}_n) - \ell_n'(\vartheta)\big)}{\ell_n''(\vartheta) + \big(\ell_n''(\widehat{\vartheta}_n) - \ell_n''(\vartheta)\big)} \\
&= \sqrt{n}\big(\,\widehat{\vartheta}_n - \vartheta\big) - \frac{\sqrt{n}\ell_n'(\vartheta) + \sqrt{n}(\widehat{\vartheta}_n - \vartheta)\ell_n''(\vartheta) + u_n}{\ell_n''(\vartheta) + v_n}.
\end{aligned}$$

La seule difficulté consiste à montrer que $u_n \xrightarrow{\mathbb{P}_\vartheta} 0$ et $v_n \xrightarrow{\mathbb{P}_\vartheta} 0$. Cela se fait de la même manière que pour la preuve de la Proposition 4.5 ou 4.6. Alors $\sqrt{n}\big(\widetilde{\vartheta}_n - \vartheta\big)$ a le même comportement asymptotique que

$$\sqrt{n}\big(\,\widehat{\vartheta}_n - \vartheta\big) - \frac{\sqrt{n}\ell_n'(\vartheta) + \sqrt{n}(\widehat{\vartheta}_n - \vartheta)\ell_n''(\vartheta)}{\ell\text{”}_n(\vartheta)} = \sqrt{n}\frac{\ell'(\vartheta)}{\ell_n''(\vartheta)}$$

[6] Il faut bien sûr que le dénominateur du terme de correction soit non nul. L'événement sur lequel il est bien défini a une $\mathbb{P}_\vartheta$-probabilité qui tend vers 1 si le modèle est régulier. Nous omettons ces aspects techniques.

qui converge en loi sous $\mathbb{P}_\vartheta$ vers la loi $\mathcal{N}\big(0, \frac{1}{\mathbb{I}(\vartheta)}\big)$ de la même manière qu'à la Section 6.3.1. □

Exemple 6.2. Une source émet des particules de type A avec probabilité ϑ et de type B avec probabilité $1-\vartheta$, où $\vartheta \in \Theta = (0,1)$. On mesure l'énergie des particules, qui est distribuée selon une densité f_1 connue pour les particules de type A et f_2 pour les particules de type B. Si l'on détecte n particules avec des énergies $X_1, \ldots, X_n$, quelle est la valeur de ϑ ? En postulant que l'observation est un n-échantillon, la fonction de vraisemblance de l'expérience statistique engendrée par l'observation s'écrit

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^n \big(\vartheta f_1(X_i) + (1-\vartheta) f_2(X_i)\big),$$

de sorte que

$$\partial_\vartheta \log \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \sum_{i=1}^n \frac{f_1(X_i) - f_2(X_i)}{\vartheta f_1(X_i) + (1-\vartheta) f_2(X_i)}.$$

La résolution de l'équation de vraisemblance associée est d'autant plus difficile que n est grand. Supposons que $\int_{\mathbb{R}} \big(F_1(x) - F_2(x)\big)^2 dx < +\infty$, où $F_i(x) = \int_{-\infty}^x f_i(t)dt$, $i = 1, 2$. Soit $\widehat{\vartheta}_n$ l'estimateur qui minimise

$$a \rightsquigarrow \int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_a(x)\big)^2 dx,$$

avec

$$F_a(x) = aF_1(x) + (1-a)F_2(x),$$

et $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{X_i \le x}$ désigne la fonction de répartition empirique de F étudiée au Chapitre 3. En dérivant par rapport à la variable a, on obtient

$$\int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_a(x)\big)\big(F_1(x) - F_2(x)\big)dx = 0,$$

d'où

$$\widehat{\vartheta}_n = \frac{\int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_2(x)\big)\big(F_1(x) - F_2(x)\big)dx}{\int_{\mathbb{R}} \big(F_1(x) - F_2(x)\big)^2 dx}.$$

En s'appuyant sur le Chapitre 3, on peut montrer que $\widehat{\vartheta}_n$ est asymptotiquement normal. Alors l'estimateur modifié

$$\widetilde{\vartheta}_n = \widehat{\vartheta}_n - \frac{\partial_\vartheta \log \mathcal{L}_n(\widehat{\vartheta}_n, X_1, \ldots, X_n)}{\partial^2_\vartheta \log \mathcal{L}_n(\widehat{\vartheta}_n, X_1, \ldots, X_n)}$$

où

$$\partial^2_\vartheta \log \mathcal{L}_n(\widehat{\vartheta}_n, X_1, \ldots, X_n) = -\sum_{i=1}^n \frac{\big(f_1(X_i) - f_2(X_i)\big)^2}{\big(\vartheta f_1(X_i) + (1-\vartheta) f_2(X_i)\big)^2}$$

est asymptotiquement efficace, et sa variance asymptotique est l'information de Fisher du modèle

$$\mathbb{I}(\vartheta) = \int_{\mathbb{R}} \frac{\big(f_1(x) - f_2(x)\big)^2}{\vartheta f_1(x) + (1-\vartheta) f_2(x)} dx.$$

Remarque 6.5. Il existe une extension multimensionnelle lorsque $\Theta \subset \mathbb{R}^d$ avec $d \geq 1$, obtenue de la même manière par un développement de Taylor à l'ordre 2. La dérivée de $\vartheta \rightsquigarrow \ell_n(\vartheta)$ est remplacée par son gradient, et la dérivée seconde par sa matrice hessienne, supposée définie positive.

6.4.3 Le programme de Fisher et ses limites

En 1922, Fisher conjectura que pour un modèle régulier (dans un sens comparable avec celui de la Section 6.3.2),

(i) L'estimateur du maximum de vraisemblance converge et a pour variance asymptotique $\frac{1}{\mathbb{I}(\vartheta)}$.

(ii) Si, pour une suite d'estimateurs $\widehat{\vartheta}_n$, on a

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v(\vartheta)\big),$$

alors, nécessairement

$$v(\vartheta) \leq \frac{1}{\mathbb{I}(\vartheta)}.$$

Le programme de Fisher aurait permis, parmi une classe d'estimateurs raisonnables, de clore le débat sur l'optimalité asymptotique. On a vu que le point (i) de la conjecture de Fisher est vrai. On a montré que le point (ii) est vrai parmi la classe restreinte des Z-estimateurs réguliers.

Mais la conjecture de Fisher est fausse en général : pour tout estimateur asymptotiquement normal, on peut construire un estimateur modifié plus efficace. Une construction classique, le contre-exemple de Hodge-Lehmann, est étudiée dans l'Exercice 6.4.

Conclusion

1. Concernant la notion de modèle régulier, par souci de simplicité, nous nous sommes restreints à un jeu d'hypothèses assez fortes. On peut étendre significativement les hypothèses de régularité.
2. La comparaison asymptotique d'estimateurs reste une notion fragile et *ad-hoc*. Un point de vue alternatif est la recherche d'uniformité en le paramètre (approche minimax).

6.4.4 Modèles non-réguliers

Nous traitons le cas des modèles non-réguliers sur un exemple incontournable : la loi uniforme.

Calcul de la vraisemblance

Considérons l'expérience engendrée par un n-échantillon de loi uniforme sur $[0, \vartheta]$, où $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$. La famille de lois $\left\{ \mathbb{P}_\vartheta,\ \vartheta \in \Theta \right\}$ associée est dominée par la mesure de Lebesgue sur $\mathbb{R}_+$, et la densité $f(\vartheta, x)$ s'écrit

$$f(\vartheta, x) = \vartheta^{-1} 1_{[0,\vartheta]}(x).$$

La fonction $\vartheta \rightsquigarrow f(\vartheta, x)$ n'est pas régulière au sens de la Définition 6.14, puisqu'elle est discontinue en $\vartheta = x$. On ne peut pas définir d'information de Fisher, et la théorie asymptotique ne s'applique pas. La vraisemblance s'écrit

$$\begin{aligned} \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) &= \prod_{i=1}^n f(\vartheta, X_i) \\ &= \vartheta^{-n} \prod_{i=1}^n 1_{[0,\vartheta]}(X_i) \\ &= \vartheta^{-n} 1_{\max_{i=1,\ldots,n} X_i \leq \vartheta}. \end{aligned}$$

La fonction

$$\vartheta \rightsquigarrow \vartheta^{-n} 1_{\max_{i=1,\ldots,n} X_i \leq \vartheta}$$

atteint son maximum unique en $\vartheta = \max_{i=1,\ldots,n} X_i$ qui est donc l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$.

Comportement asymptotique de l'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance n'est pas asymptotiquement normal, et la précision d'estimation de $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$ est meilleure que la vitesse $1/\sqrt{n}$ des modèles réguliers.

On peut préciser son comportement asymptotique. Pour $t \in \mathbb{R}$, on a

$$\begin{aligned} \mathbb{P}_\vartheta \left[\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} \leq t \right] &= \mathbb{P}_\vartheta \left[\bigcap_{i=1}^n (X_i \leq t) \right] \\ &= \prod_{i=1}^n \mathbb{P}_\vartheta \left[X_i \leq t \right] \\ &= (\vartheta^{-1} t)^n 1_{[0,\vartheta]}(t) + 1_{t > \vartheta} \end{aligned}$$

par indépendance des X_i. Il vient

$$\begin{aligned}\mathbb{P}_\vartheta\left[n(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}-\vartheta)\le t\right] &= \mathbb{P}_\vartheta\left[\,\widehat{\vartheta}_n \le \vartheta+\tfrac{t}{n}\right]\\ &= \big(1+\vartheta^{-1}\frac{t}{n}\big)^n 1_{[-n\vartheta,0]}(t)+1_{t>0}\\ &\to e^{\vartheta^{-1}t}1_{t\le 0}+1_{t>0}.\end{aligned}$$

Donc $n(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}-\vartheta)$ converge en loi sous $\mathbb{P}_\vartheta$ vers une loi de fonction de répartition

$$F(t)=e^{\vartheta^{-1}t}1_{t\le 0}+1_{t>0},$$

dérivable presque-partout, et de densité $t\rightsquigarrow \vartheta^{-1}e^{-\vartheta t}1_{t\le 0}$, qui peut s'écrire comme $-Z$, où Z est une variable aléatoire exponentielle de paramètre ϑ^{-1}. On notera que dans ce modèle, la vitesse d'estimation est $1/n$ et non $1/\sqrt{n}$ comme dans les modèles réguliers.

6.5 Perte d'information⋆

6.5.1 Sous-expérience statistique

On considère une expérience statistique $\mathcal{E}$ arbitraire, engendrée par une observation Z à valeurs dans $(\mathfrak{Z},\mathcal{Z})$.

Dans l'expérience $\mathcal{E}$, un estimateur $\widehat{\vartheta}_n$ est la donnée d'une fonction : $\varphi_n:\mathfrak{Z}\to\Theta$ appliquée à l'observation, c'est-à-dire

$$\widehat{\vartheta}_n=\varphi_n(Z).$$

Considérons maintenant une application mesurable

$$T:(\mathfrak{Z},\mathcal{Z})\longrightarrow(\mathfrak{Y},\mathcal{Y})$$

où $(\mathfrak{Y},\mathcal{Y})$ est un espace mesurable donné, et posons $Y=T(Z)$. Alors Y apparaît comme une « sous-observation » de Z et un estimateur de la forme $\widetilde{\vartheta}_n=\varphi_n(Y)=\varphi_n\big(T(Z)\big)$ sera en général moins performant qu'un estimateur de la forme $\widehat{\vartheta}_n=\varphi_n(Z)$.

A l'application T est attachée une notion de perte d'information, ou de compression d'information, que nous allons un peu formaliser.

Définition 6.18. *On appelle sous-expérience de $\mathcal{E}$ associée à T est on note $\mathcal{E}^T$ l'expérience engendrée par l'observation $T(Z)$.*

Si

$$\mathcal{E}=\big(\mathbb{R}^n,\mathcal{B}^n,(\mathbb{P}_\vartheta,\vartheta\in\Theta)\big),$$

on a

$$\mathcal{E}^T=\big(T(\mathbb{R}^n),\mathcal{Y},(\mathbb{P}_\vartheta^T,\vartheta\in\Theta)\big),$$

où $\mathbb{P}_\vartheta^T$ est la mesure image de $\mathbb{P}_\vartheta$ par T. C'est une mesure de probabilité définie sur $\big(T(\mathbb{R}^n), \mathcal{Y}\big)$ par

$$\mathbb{P}_\vartheta^T\left[A\right] = \mathbb{P}_\vartheta\left[T^{-1}(A)\right], \quad A \in \mathcal{Y}.$$

Un premier résultat très intuitif est que l'on perd de l'information en passant de $\mathcal{E}$ à $\mathcal{E}^T$.

Proposition 6.6. *Si $\mathcal{E}$ et $\mathcal{E}^T$ sont régulières, alors, pour tout $\vartheta \in \Theta$*

$$\mathbb{I}\big(\vartheta \,\big|\mathcal{E}^T\big) \leq \mathbb{I}\big(\vartheta \,|\mathcal{E}\big),$$

où $\mathbb{I}(\vartheta\,|\mathcal{E})$ désigne l'information de Fisher pour l'expérience statistique $\mathcal{E}$ au point ϑ.

Notons tout d'abord que si μ domine $\mathcal{E}$, alors la mesure image μ^T de μ par T domine[7] $\mathcal{E}^T$. Posons

$$f^T(\vartheta, z) = \frac{d\,\mathbb{P}_\vartheta^T}{d\mu^T}(z), \quad z \in \mathfrak{Z},\ \vartheta \in \Theta.$$

On démontre cette proposition en deux étapes. Une première étape est un résultat intéressant en lui-même que nous énonçons sous forme de lemme

Lemme 6.5.1. *On a, pour tout $\vartheta \in \Theta$,*

$$\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, Z)\,|\,T(Z)\right] = \partial_\vartheta \log f^T\big(\vartheta, T(Z)\big) \quad \mathbb{P}_\vartheta - \textit{presque sûrement.}$$

Démonstration. Soit $A \in \mathcal{Y}$. D'une part, par caractérisation de l'espérance conditionnelle s'écrit

$$\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, Z) 1_{T(Z)\in A}\right] = \mathbb{E}_\vartheta\left[\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, Z)\,|T\right] 1_{T(Z)\in A}\right].$$

D'autre part, puisque $\mathbb{P}_\vartheta$ est la loi de Z, on a par formule de la mesure image (1.1)

$$\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, Z) 1_{T(Z)\in A}\right] = \int_{T^{-1}(A)} \partial_\vartheta \log f(\vartheta, z)\,\mathbb{P}_\vartheta(dz).$$

[7] En effet, si $\mathbb{P}_\vartheta^T\left[A\right] = 0$, alors $\mathbb{P}_\vartheta\left[T^{-1}(A)\right] = 0$ et donc $\mu\big[T^{-1}(A)\big] = 0 = \mu^T\left[A\right]$.

Puisque $\mathcal{E}$ est régulière, il vient

$$\begin{aligned}
&\int_{T^{-1}(A)} \partial_\vartheta \log f(\vartheta, z)\, \mathbb{P}_\vartheta(dz) \\
&= \int_{T^{-1}(A)} \partial_\vartheta f(\vartheta, z) \mu(dz) \\
&= \partial_\vartheta \int_{T^{-1}(A)} f(\vartheta, z) \mu(dz) \\
&= \partial_\vartheta \int_{T^{-1}(A)} \mathbb{P}_\vartheta(dz) \\
&= \partial_\vartheta \int_A \mathbb{P}_\vartheta^T(dz) \ \text{ (formule de la mesure image (1.1))} \\
&= \partial_\vartheta \int_A f^T(\vartheta, z) \mu^T(dz) \\
&= \int_A \partial_\vartheta f^T(\vartheta, z) \mu^T(dz) \\
&= \int_A \partial_\vartheta \log f^T(\vartheta, z)\, \mathbb{P}_\vartheta^T(dz) \\
&= \mathbb{E}_\vartheta \left[\partial_\vartheta \log f^T\big(\vartheta, T(Z)\big) 1_{T(Z)\in A} \right] \ \text{ (formule de la mesure image (1.1)).}
\end{aligned}$$

Comme A est arbitraire, on conclut par identification. □

Passons à la preuve de la Proposition 6.6 proprement dite On a

$$\mathbb{E}_\vartheta \left[\big(\partial_\vartheta \log f(\vartheta, Z) - \partial_\vartheta \log f^T\big(\vartheta, T(Z)\big) \big)^2 \right] \geq 0.$$

En développant le carré, on obtient

$$\mathbb{I}\big(\vartheta \,|\mathcal{E}\big) + \mathbb{I}\big(\vartheta \,\big|\mathcal{E}^T\big) - 2\, \mathbb{E}_\vartheta \left[\partial_\vartheta \log f(\vartheta, Z) \partial_\vartheta \log f^T\big(\vartheta, T(Z)\big) \right] \geq 0.$$

D'autre part,

$$\begin{aligned}
&\mathbb{E}_\vartheta \left[\partial_\vartheta \log f(\vartheta, Z) \partial_\vartheta \log f^T\big(\vartheta, T(Z)\big) \right] \\
&= \mathbb{E}_\vartheta \left[\mathbb{E}_\vartheta \left[\partial_\vartheta \log f(\vartheta, Z) \,|T \right] \partial_\vartheta \log f^T\big(\vartheta, T(Z)\big) \right] \\
&= \mathbb{E}_\vartheta \left[\big(\partial_\vartheta \log f^T\big(\vartheta, T(Z)\big) \big)^2 \right].
\end{aligned}$$

la dernière égalité étant obtenue en appliquant le Lemme 6.5.1. Cette dernière quantité est précisément $\mathbb{I}\big(\vartheta \,\big|\mathcal{E}^T\big)$, ce qui achève la démontration de la Proposition 6.6.

6.5.2 Statistique exhaustive

Absence de perte d'information

Nous nous intéressons à une classe particulière de fonctions T, celle qui ne font pas perdre d'information. Ecrites sous la forme $Y = T(Z)$ on appelle ces fonctions des « statistiques exhaustives ».

Définition 6.19 (Statistique exhaustive)**.** *On dit que la statistique T est exhaustive (ou plutôt $Y = T(Z)$) pour l'expérience régulière $\mathcal{E}$ si $\mathcal{E}^T$ est régulière et*

$$\mathbb{I}\big(\vartheta \,\big|\mathcal{E}^T\big) = \mathbb{I}\big(\vartheta \,|\mathcal{E}\big),$$

Pour de telles sous-expériences, il n'y a pas de perte d'information, et la théorie de l'efficacité asymptotique reste inchangée.

Remarque 6.6. Il existe une définition plus large qui permet de définir l'exhausitivité (l'absence de perte d'information), même lorsque l'information de Fisher n'est pas définie, que nous ne donnons pas ici. Nous utiliserons la notion d'exhaustivité au Chapitre 7 dans la cadre de modèles régulier, et nous pouvons nous contenter de la Définition 6.19 dans ce cours.

Remarque 6.7. Nous avons traité le cas d'un paramètre unidimensionnel $\vartheta \in \Theta \subset \mathbb{R}$ par souci de simplicité. On a des résultats analogues pour un paramètre $\vartheta \in \Theta \subset \mathbb{R}^d$ avec $d > 1$ en remplaçant l'information de Fisher par la matrice d'information de Fisher, pour des hypothèses de régularité suffisantes. Nous ne développerons pas ces aspects ici (voir tout de même l'Exemple 6.5).

Critère de factorisation

La notion d'exhaustivité, c'est-à-dire d'absence de perte d'information pour une sous-expérience n'est pas facile à manipuler à partir de la Définition 6.19. Nous donnons un critère très simple pour montrer qu'une statistique est exhaustive.

Théorème 6.2 (Critère de Factorisation)**.** *Si l'expérience $\mathcal{E}$ est dominée par μ, une statistique T est exhaustive si et seulement si la vraisemblance $f(\vartheta, Z) = \frac{d\,\mathbb{P}_\vartheta}{d\mu}(Z)$ s'écrit*

$$f(\vartheta, Z) = p\big(T(Z), \vartheta\big)h(Z) \quad \mu \textit{ presque-partout}, \tag{6.15}$$

où les fonctions $z \rightsquigarrow p(\bullet, z)$ et $z \rightsquigarrow h(z)$ sont mesurables et positives.

Nous donnons une preuve très simple dans notre cadre où nous supposons les expériences statistiques $\mathcal{E}$ et $\mathcal{E}^T$ régulières, et où nous supposons de plus que $f(\vartheta, \bullet)$ et strictement positive pour tout $\vartheta \in \Theta$ pour simplifier. Pour le cas général évoqué dans la Remarque 6.6, on trouvera une démonstation du théorème de factorisation dans Borovkov, [1], pp. 117–120.

Démonstration. Si $f(\vartheta, Z) = p\big(T(Z), \vartheta\big)h(Z)$ μ presque-partout, alors la mesure

$$\widetilde{\mu}(dz) = h(z)\mu(dz)$$

domine la famille $\big\{\, \mathbb{P}_\vartheta,\, \vartheta \in \Theta\big\}$. Puisque h est strictement positive, les ensembles de μ ou $\widetilde{\mu}$ mesure nulle coïncident. D'après l'Exercice 6.2, l'information de Fisher ne dépend pas du choix de la mesure dominante, que l'on calcule avec $\widetilde{f}(\vartheta, z) = \frac{d\,\mathbb{P}_\vartheta}{d\widetilde{\mu}}(z)$. On a d'une part,

$$\mathbb{E}_\vartheta\left[\partial_\vartheta \log \widetilde{f}(\vartheta, Z)\,|\,T(Z)\right] = \partial_\vartheta \log \widetilde{f}(\vartheta, Z)$$

$\widetilde{\mu}$ presque partout, puisque $\widetilde{f}(\vartheta, Z) = p\big(T(Z), \vartheta\big)$ est une fonction mesurable de $T(Z)$. D'autre part, d'après le Lemme 6.5.1 et avec les mêmes notations, on a

$$\mathbb{E}_\vartheta\left[\partial_\vartheta \log \widetilde{f}(\vartheta, Z)\,|\,T(Z)\right] = \partial_\vartheta \log \widetilde{f}^T(\vartheta, Z)$$

$\widetilde{\mu}$ presque partout. On en déduit

$$\partial_\vartheta \log \widetilde{f}(\vartheta, Z) = \partial_\vartheta \log \widetilde{f}^T(\vartheta, Z)$$

à un ensemble de $\widetilde{\mu}$ mesure nulle près. Le résultat en découle en passant au carré et en intégrant par rapport à $\mathbb{P}_\vartheta$.

Réciproquement, on a montré dans la Proposition 6.6 que

$$\mathbb{E}_\vartheta\left[\big(\partial_\vartheta \log f(\vartheta, Z) - \partial_\vartheta \log f^T\big(\vartheta, T(Z)\big)\big)^2\right] = \mathbb{I}\big(\vartheta\,|\,\mathcal{E}\big) - \mathbb{I}\big(\vartheta\,|\,\mathcal{E}^T\big) \geq 0.$$

En conséquence, si $\mathbb{I}\big(\vartheta\,|\,\mathcal{E}\big) = \mathbb{I}\big(\vartheta\,|\,\mathcal{E}^T\big)$, alors

$$\partial_\vartheta \log f(\vartheta, Z) = \partial_\vartheta \log f^T\big(\vartheta, T(Z)\big), \tag{6.16}$$

l'égalité ayant lieu $\mathbb{P}_\vartheta$ presque-sûrement, et aussi μ presque-partout en utilisant le fait que $f(\vartheta, \bullet)$ est strictement positive. En intégrant (6.16), on obtient la représentation (6.15). □

6.5.3 Exemples de statistiques exhaustives

Exemple 6.3 (Modèle de Bernoulli)**.** Dans l'exemple 1 du Chapitre 2, nous avons introduit deux expériences statistiques pour traiter le problème du sondage. D'une part, l'expérience $\mathcal{E}^n$, engendrée par l'observation d'un n-échantillon $X_1, \ldots, X_n$ de variables aléatoires de Bernoulli de paramètre $\vartheta \in \Theta = [0, 1]$, qui s'écrit

$$\mathcal{E}^n = \Big(\{0,1\}^n, \text{parties de}\big(\{0,1\}^n\big),\, \big\{\, \mathbb{P}_\vartheta^n,\, \vartheta \in \Theta\big\}\Big),$$

où $\mathbb{P}_\vartheta^n = \mathbb{P}_\vartheta \otimes \ldots \otimes \mathbb{P}_\vartheta$ (n-fois), avec

$$\mathbb{P}_\vartheta\left[X = 1\right] = \vartheta = 1 - \mathbb{P}_\vartheta\left[X = 0\right],$$

et qui correspond à l'observation du résultat de chaque votant. D'autre part, l'expérience $\widetilde{\mathcal{E}}^n$ engendrée par l'observation de la somme[8] $\sum_{i=1}^n X_i$, notée

$$\widetilde{\mathcal{E}}^n = \Big(\{0,\dots,n\}, \text{parties de}\big(\{0,\dots,n\}\big),\ \big\{\mathbb{Q}_\vartheta^n, \vartheta\in\Theta\big\}\Big),$$

où $\mathbb{Q}_\vartheta^n$ est la loi binômiale de paramètres (n,ϑ), et qui correspond à l'observation du nombre total de voix pour le candidat A. Intuitivement, les deux points de vue contiennent la même information sur le paramètre ϑ. La notion d'exhaustivité permet de formaliser cette intuition. L'expérience $\widetilde{\mathcal{E}}^n = \big(\mathcal{E}^n\big)^T$ est une sous-expérience de $\mathcal{E}^n$ pour l'application

$$\begin{aligned} T : \{0,1\}^n &\to \{0,\dots,n\} \\ (x_1,\dots,x_n) &\rightsquigarrow T(x_1,\dots,x_n) = \sum_{i=1}^n x_i. \end{aligned}$$

Ecrivons maintenant la vraisemblance dans $\mathcal{E}^n$ en prenant comme mesure dominante la mesure de comptage sur $\{0,1\}^n$: on a

$$\begin{aligned} \mathcal{L}\big(\vartheta, X_1,\dots,X_n\big) &= \prod_{i=1}^n \vartheta^{X_i}(1-\vartheta)^{1-X_i} \\ &= \vartheta^{T(X_1,\dots,X_n)}(1-\vartheta)^{n-T(X_1,\dots,X_n)}, \end{aligned}$$

et le critère de factorisation nous dit que la statistique $T(X_1,\dots,X_n)$ est exhaustive. Il n'y a donc pas de perte d'information si l'on considère $\widetilde{\mathcal{E}}^n$ plutôt que $\mathcal{E}^n$.

Exemple 6.4 (Loi exponentielle)**.** on considère l'expérience statistique engendrée par un n-échantillon de loi exponentielle de paramètre $\vartheta\in\Theta=\mathbb{R}_+\setminus\{0\}$. La vraisemblance s'écrit

$$\begin{aligned} \mathcal{L}_n(\vartheta, X_1,\dots,X_n) &= \vartheta^n \exp\Big(-\vartheta\sum_{i=1}^n X_i\Big) \\ &= \vartheta^n \exp\big(-\vartheta n\,\overline{X}_n\big) \\ &= p\big(T(X_1,\dots,X_n),\vartheta\big)h(X_1,\dots,X_n) \end{aligned}$$

avec $p(x,\vartheta)=\vartheta^n\exp\big(-\vartheta x\big)$ et $h=1$. Donc $T(X_1,\dots,X_n)=\overline{X}_n$ est une statistique exhaustive d'après le théorème de factorisation.

Exemple 6.5 (Un exemple en dimension $d=2$)**.** On considère l'expérience statistique engendrée par un n-échantillon de loi $\mathcal{N}(\mu,\sigma^2)$, avec comme paramètre $\vartheta=(\mu,\sigma^2)\in\Theta=\mathbb{R}\times\mathbb{R}_+\setminus\{0\}$. La vraisemblance s'écrit

$$\begin{aligned} \mathcal{L}_n(\vartheta, X_1,\dots,X_n) &= (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2\Big) \\ &= (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{n}{2\sigma^2}(\tfrac{1}{n}\sum_{i=1}^n X_i^2 - 2\,\overline{X}_n + \mu^2\Big), \end{aligned}$$

[8] notée n_A dans l'exemple du Chapitre 2.

ce qui montre que la statistique $T(X_1,\dots,X_n) = (\overline{X}_n, \frac{1}{n}\sum_{i=1}^n X_i^2)$ est exhaustive d'après le théorème de factorisation. Si l'on suppose $\sigma^2 = 1$ connu, alors le paramètre devient $\vartheta = \mu$ et la vraisemblance s'écrit :

$$\begin{aligned}\mathcal{L}_n(\vartheta, X_1,\dots,X_n) &= (2\pi)^{-n/2}\exp\Big(-\frac{1}{2}\sum_{i=1}^n (X_i-\mu)^2\Big)\\ &= (2\pi)^{-n/2}\exp\big(n\,\overline{X}_n - \tfrac{n\mu^2}{2}\big)\exp\big(-\tfrac{1}{2}\sum_{i=1}^n X_i^2\big)\end{aligned}$$

et on conclut que dans ce cas $T(X_1,\dots,X_n) = \overline{X}_n$ est exhaustive d'après le critère de factorisation.

6.6 Exercices

Exercice 6.1. On suppose que $\Theta = \{\vartheta_0, \vartheta\} \subset \mathbb{R}$ avec $\vartheta_0 \neq \vartheta_1$ est réduit à deux points et que les mesures $\mathbb{P}_{\vartheta_0}$ et $\mathbb{P}_{\vartheta_1}$ sont mutuellement absolument continues (c'est-à-dire $\mathbb{P}_{\vartheta_0} \ll \mathbb{P}_{\vartheta_1}$ et $\mathbb{P}_{\vartheta_1} \ll \mathbb{P}_{\vartheta_0}$). Montrer qu'il n'existe pas d'estimateur $\vartheta^\star$ tel que

$$\forall \vartheta \in \Theta,\ \ \mathcal{R}(\vartheta^\star, \vartheta) \leq \inf_{\widehat{\vartheta}_n} \mathcal{R}(\widehat{\vartheta}_n, \vartheta),$$

où l'infimum est pris sur l'ensemble de tous les estimateurs, où $\mathcal{R}(\widehat{\vartheta}_n, \vartheta) = \mathbb{E}_\vartheta\left[(\widehat{\vartheta}_n - \vartheta)^2\right]$ désigne le risque quadratique de l'estimateur $\widehat{\vartheta}_n$ au point ϑ.

Exercice 6.2. Soit $\big\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\big\}$, avec $\Theta \subset \mathbb{R}$ une famille de probabilités sur $\mathbb{R}$ régulière au sens de la Définition 6.14. On suppose que pour tout $\vartheta \in \Theta$, on a

$$f(\vartheta, x) > 0, \quad \mu(dx) - \text{presque partout},$$

où μ est une mesure dominante. Montrer que l'information de Fisher $\mathbb{I}(\vartheta)$ ne dépend pas du choix de μ.

Exercice 6.3 (Inégalité de Cramer-Rao)**.** On considère l'expérience engendrée par un n-échantillon de loi appartenant à la famille régulière $\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\}$, où $\Theta \subset \mathbb{R}$. Si $\widehat{\vartheta}_n$ est une estimateur de ϑ (de carré intégrable), on a, pour tout $\vartheta \in \Theta$

$$\mathbb{E}_\vartheta\left[(\widehat{\vartheta}_n - \vartheta)^2\right] \geq \frac{1 + b'(\vartheta)}{n\mathbb{I}(\vartheta)} + b(\vartheta)^2, \tag{6.17}$$

où $b(\vartheta) = \mathbb{E}_\vartheta\left[\,\widehat{\vartheta}_n\,\right] - \vartheta$ est le biais de l'estimateur $\widehat{\vartheta}_n$.

– En partant de l'identité $1 = \int_{\mathbb{R}} f(\vartheta, x)\mu(dx)$, montrer que

$$0 = \int_{\mathbb{R}} \partial_\vartheta f(\vartheta, x)\mu(dx).$$

– En déduire

$$\mathbb{E}_\vartheta \left[(\widehat{\vartheta}_n - \vartheta) \partial_\vartheta f(\vartheta, X) \right] = 1,$$

et par l'inégalité de Cauchy-Schwarz, montrer l'inégalité de Cramer-Rao (6.17).

Exercice 6.4 (Super-efficacité et contre-exemple de Hodge-Lehmann*)**.** Dans un modèle régulier, soit $\widehat{\vartheta}_n$ un estimateur asymptotiquement normal, de variance $v(\vartheta)$, pour $\vartheta \in \Theta \subset \mathbb{R}$. On suppose de plus que pour un point $\vartheta_0 \in \Theta$, il existe $\varepsilon > 0$ tel que

$$\sup_n \mathbb{E}_{\vartheta_0} \left[n^2 (\widehat{\vartheta}_n - \vartheta_0)^{2+\varepsilon} \right] < +\infty. \tag{6.18}$$

1. Donner un exemple de modèle vérifiant (6.18)
2. On pose

$$\widetilde{\vartheta}_n = \widehat{\vartheta}_n \, 1_{\{|\,\widehat{\vartheta}_n - \vartheta_0| > n^{-1/4}\}} + \vartheta_0 1_{\{|\,\widehat{\vartheta}_n - \vartheta_0| \leq n^{-1/4}\}}.$$

Montrer que si $\vartheta \neq \vartheta_0$, on a

$$\sqrt{n}(\widetilde{\vartheta}_n - \vartheta) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{v(\vartheta)}\Big).$$

3. Montrer que $n \, \mathbb{E}_{\vartheta_0} \left[(\widetilde{\vartheta}_n - \vartheta_0)^2 \right] \to 0$.

En conclusion, on a construit un estimateur $\widetilde{\vartheta}_n$ asymptotiquement normal en tout point $\vartheta \neq \vartheta_0$, de même variance asymptotique que $\widehat{\vartheta}_n$, et strictement meilleur que $\widehat{\vartheta}_n$ en ϑ_0. En déduire que cet estimateur infirme la seconde conjecture de Fisher.

Troisième partie

Tests d'hypothèses

Chapitre 7

Tests et régions de confiance

Nous avons déja rencontré la notion de test statistique dans le Chapitre 3. Dans ce chapitre, nous systématisons cette approche et donnons quelques résultats incontournables de construction de test et nous abordons la notion d'optimalité. Nous allons voir que si on accepte de hiérarchiser les erreurs de décision lorsque l'on procède à un test (principe de Neyman), alors il est possible de définir une notion d'optimalité plus satisfaisante que pour l'estimation.

7.1 Problématique des tests d'hypothèse

7.1.1 Test et erreur de test

Situation

On considère une expérience statistique engendrée par une observation Z à valeurs dans $(\mathfrak{Z}, \mathcal{Z})$ et associée à la famille de lois de probabilités

$$\big\{ \mathbb{P}_\vartheta,\ \vartheta \in \Theta \big\},$$

L'ensemble des paramètres Θ est un sous-ensemble de $\mathbb{R}^d$, avec $d \geq 1$.

Dans le modèle de la densité, $Z = (X_1, \ldots, X_n)$ est un n-échantilllon où les variables aléatoires réelles X_i sont indépendantes et de même loi, et $\mathbb{P}_\vartheta$ est la loi du n-échantillon définie sur $(\mathfrak{Z}, \mathcal{Z}) = (\mathbb{R}^n, \mathcal{B}^n)$.

Dans le modèle de la régression à « design » déterministe on peut écrire l'observation comme $Z = (Y_1, \ldots, Y_n)$, où les $Y_i = f(\vartheta, \boldsymbol{x}_i) + \xi_i$ sont indépendantes et le « design » $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ est donné une bonne fois pour toutes. Dans ce cas, $\mathbb{P}_\vartheta$ est la loi jointe des Y_i définie sur $(\mathfrak{Z}, \mathcal{Z}) = (\mathbb{R}^n, \mathcal{B}^n)$.

Principe du test statistique

On veut « décider » à partir de l'observation de Z si une propriété de la loi de Z est vérifiée ou non. Cette propriété se traduit mathématiquement par un sous-ensemble $\Theta_0 \subset \Theta$ de l'ensemble des paramètres, et la propriété signifie que $\vartheta \in \Theta_0$.

Définition 7.1 (Terminologie de test)**.** *On teste « l'hypothèse nulle »*

$$H_0 \ : \ \vartheta \in \Theta_0 \subset \Theta$$

contre « l'alternative »

$$H_1 \ : \ \vartheta \in \Theta_1 \subset \Theta,$$

avec $\Theta_0 \cap \Theta_1 = \emptyset$. Construire un test signifie construire une procédure $\varphi = \varphi(Z)$ de la forme

$$\varphi(Z) = 1_{\{Z \in \mathcal{R}\}} = \begin{cases} 0 & \text{si } Z \notin \mathcal{R}. \quad \text{« on accepte l'hypothèse nulle »} \\ \\ 1 & \text{si } Z \in \mathcal{R}. \quad \text{« on rejette l'hypothèse nulle »} \end{cases} \tag{7.1}$$

On dit que φ est un test simple.

Il est naturel de prendre $\Theta_1 = \Theta \setminus \Theta_0$ et c'est ce que l'on fera la plupart du temps. On verra toutefois que ce choix ne s'impose pas toujours et dépend des propriétés que l'on souhaite obtenir pour φ. Pour le moment, on supposera $\Theta_1 = \Theta \setminus \Theta_0$.

Définition 7.2. *Toute procédure statistique de la forme* (7.1) *est appelée test simple. On désigne indifféremment l'ensemble $\mathcal{R} \subset \mathfrak{Z}$ ou bien l'événement $\{Z \in \mathcal{R}\}$ comme zone de rejet ou encore zone critique du test φ.*

Remarque 7.1. Dans la définition 7.1, on parle de test simple car on n'autorise que deux réponses (accepter ou rejeter). On pourrait imaginer des situations plus générales, où l'on se refuse à décider, ou bien où l'on renvoie une valeur entre 0 et 1 qui indique un « degré de suspicion » de l'hypothèse.

Erreur de test

Lorsque l'on effectue un test simple, il y a quatre possibilités. Deux sont anecdotiques et correspondent à une bonne décision :

- Accepter l'hypothèse H_0 alors que $\vartheta \in \Theta_0$ (c'est-à-dire l'hypothèse H_0 est vraie).
- Rejeter l'hypothèse H_0 alors que $\vartheta \in \Theta_1$ (c'est-à-dire l'hypothèse H_0 est fausse).

Les deux autres possibilités sont celles qui vont nous occuper, et correspondent à une erreur de décision :

- Rejeter l'hypothèse H_0 alors que $\vartheta \in \Theta_0$ (c'est-à-dire l'hypothèse H_0 est vraie).

– Accepter l'hypothèse H_0 alors que $\vartheta \in \Theta_1$ (c'est-à-dure l'hypothèse H_0 est fausse).

Définition 7.3 (Erreur de première et seconde espèce)**.** *L'erreur de première espèce correspond à la probabilité maximale de rejeter l'hypothèse alors qu'elle est vraie :*

$$\sup_{\vartheta\in\Theta_0} \mathbb{E}_\vartheta \left[\varphi(Z)\right] = \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta \left[Z \in \mathcal{R}\right].$$

L'erreur de seconde espèce correspond à la probabilité maximale d'accepter l'hypothèse alors qu'elle est fausse :

$$\sup_{\vartheta\in\Theta_1} \mathbb{E}_\vartheta \left[1-\varphi(Z)\right] = \sup_{\vartheta\in\Theta_1} \mathbb{P}_\vartheta \left[Z \notin \mathcal{R}\right]. \tag{7.2}$$

Remarque 7.2. D'après cette terminologie, l'erreur de première espèce mesure la probabilité (maximale) de rejeter à tort, et l'erreur de seconde espèce d'accepter à tort. Dans le langage courant, commettre une erreur de première espèce revient à faire un « faux négatif », et commettre un erreur de seconde espèce revient à faire un « faux positif ».

Dans la plupart des situations, Θ_0 est « plus petit » que Θ_1 et le contrôle de l'erreur de seconde espèce (7.2) est difficile, surtout si Θ_1 contient des points « très proches » de Θ_0. C'est pour cela que l'on introduit la fonction de fonction de puissance d'un test, qui mesure sa performance locale sur l'alternative.

Définition 7.4. *La fonction de puissance du test simple φ est l'application*

$$\pi_\bullet(\varphi) : \Theta_1 \to [0,1]$$

définie par

$$\vartheta \in \Theta_1 \rightsquigarrow \pi_\vartheta(\varphi) = \mathbb{P}_\vartheta \left[Z \in \mathcal{R}\right].$$

Hypothèse simple, hypothèse composite

On utilise souvent la terminologie suivante dans le cas réel, où $\Theta \subset \mathbb{R}$. Soit $\vartheta_0 \in \Theta$.
– Tester $H_0 : \ \vartheta = \vartheta_0$ contre $H_1 : \ \vartheta = \vartheta_1$ avec $\vartheta_1 \neq \vartheta_0$. On parle de test d'une hypothèse simple contre une alternative simple.
– Tester $H_0 : \ \vartheta = \vartheta_0$ contre $H_1 : \ \vartheta \neq \vartheta_0$. On parle de test d'une hypothèse simple contre une alternative composite.
– Tester $H_0 : \ \vartheta > \vartheta_0$ contre $H_1 : \ \vartheta \leq \vartheta_0$. On parle de test d'une hypothèse composite contre une alternative composite.
– Tester $H_0 : \ \vartheta > \vartheta_0$ contre $H_1 : \ \vartheta = \vartheta_0$. On parle de test d'une hypothèse composite contre une alternative simple.

7.1.2 Comparaison de test, principe de Neyman

Idéalement, on souhaite que l'erreur de première espèce et l'erreur de seconde espèce soient toutes deux simultanément petites. Les deux tests triviaux

$$\varphi_1 = 1_\emptyset, \quad \text{et} \quad \varphi_2 = 1_\mathfrak{Z}$$

qui consistent respectivement à accepter systématiquement l'hypothèse et à la rejeter systématiquement, sans utiliser l'observation Z ont respectivement une erreur de première espèce nulle et une erreur de seconde espèce nulle. Malheureusement la puissance de φ_1 est catastrophique : $\pi_\vartheta(\varphi_1) = 0$ en tout point ϑ de toute alternative Θ_1. De même l'erreur de première espèce de φ_2 est égale à 1, même si l'hypothèse est réduite à un point, quelle que soit l'hypothèse.

Une méthodologie, proposée historiquement par Neyman, consiste à imposer une dissymétrie dans la problématique de test : on décide que le contrôle de l'erreur de première espèce est cruciale. La démarche de construction de test sera alors, parmi les tests qui ont une erreur de première espèce contrôlée, de choisir le (ou les) tests les plus puissants, c'est-à-dire ayant une erreur de seconde espèce la plus petite possible.

Définition 7.5. *Soit $\alpha \in [0,1]$ un niveau de risque. Un test simple φ est de niveau α si son erreur de première espèce est inférieure ou égale à α.*

Définition 7.6 (Principe de Neyman)**.** *Soit $\alpha \in [0,1]$ un niveau de risque. Le test $\varphi^\star$ est optimal (uniformément plus puissant, ou UPP) pour tester*

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

si $\varphi^\star$ est de niveau α et, pour tout test φ de niveau α, on a

$$\forall \vartheta \in \Theta_1, \ \ \pi_\vartheta(\varphi) \leq \pi_\vartheta(\varphi^\star).$$

7.2 Hypothèse simple contre alternative simple

7.2.1 Principe de Neyman et décision à deux points

Dans le cas d'une hypothèse simple contre une alternative simple, on sait résoudre de façon optimale le principe de Neyman. Il s'agit d'une situation remarquable, qui ne se généralise pas facilement – mis à part des cas particuliers comme les familles à rapport de vraisemblance monotone, voir Section 7.3.1 – dans un cadre non-asymptotique.

On suppose l'ensemble des paramètre réduit à deux points : $\Theta = \{\vartheta_0, \vartheta_1\}$. A partir de l'observation Z, on teste

$$H_0 : \vartheta = \vartheta_0 \quad \text{contre} \quad H_1 : \ \vartheta = \vartheta_1.$$

Définition 7.7 (Optimalité)**.** *Soit $\alpha \in [0,1]$ un niveau de risque. Un test $\varphi^\star$ de niveau α est optimal ou PP (Plus Puissant) si*

$$\pi(\varphi^\star) = \sup_{\varphi} \pi(\varphi)$$

où le supremum est pris parmi tous les tests de niveau α.

Dans le cas d'une hypothèse simple contre une alternative simple, estimation et test se confondent. En effet, un estimateur « raisonnable[1] » se représente sous la forme

$$\widehat{\vartheta}_n = \vartheta_0 1_{Z\in\mathcal{A}} + \vartheta_1 1_{Z\notin\mathcal{A}}$$

pour un certain ensemble $\mathcal{A} \subset \mathcal{Z}$, et peut se mettre en correspondance avec le test simple de l'hypothèse $H_0 : \vartheta = \vartheta_0$ contre $H_1 : \vartheta = \vartheta_1$ défini par

$$\varphi_n = 1_{\{Z\notin\mathcal{A}\}}.$$

Si $(\vartheta, \widetilde{\vartheta}) \rightsquigarrow \ell(\vartheta, \widetilde{\vartheta})$ est une fonction de perte[2] donnée, et si $\mathcal{R}\big(\widehat{\vartheta}_n, \vartheta\big) = \mathbb{E}_\vartheta\left[\ell(\widehat{\vartheta}_n, \vartheta)\right]$ désigne le risque de l'estimateur $\widehat{\vartheta}_n$ pour la perte $\ell(\bullet,\bullet)$ au point ϑ, on a

$$\begin{aligned}\mathcal{R}\big(\widehat{\vartheta}_n, \vartheta\big) &= \mathbb{E}_\vartheta\left[\ell(\vartheta_0,\vartheta)1_{Z\in\mathcal{A}} + \ell(\vartheta_1,\vartheta)1_{Z\notin\mathcal{A}}\right]\\ &= \ell(\vartheta_0,\vartheta)\,\mathbb{P}_\vartheta\left[\varphi = 0\right] + \ell(\vartheta_1,\vartheta)\,\mathbb{P}_\vartheta\left[\varphi = 1\right].\end{aligned}$$

Donc

$$\mathcal{R}(\widehat{\vartheta}_n, \vartheta_0) = \ell(\vartheta_1,\vartheta_0)\,\mathbb{P}_{\vartheta_0}\left[\varphi = 1\right]$$

soit l'erreur de première espèce du test φ, et

$$\mathcal{R}(\widehat{\vartheta}_n, \vartheta_1) = \ell(\vartheta_0,\vartheta_1)\big(1 - \pi(\varphi)\big),$$

soit l'erreur de econde espèce du test. Construire un estimateur ayant un risque « petit » en ϑ_0 et ϑ_1 est équivalent ici à construire un test ayant simultanément une erreur de première et de seconde espèce petite.

Le principe de Neyman au niveau α se traduit comme la recherche de φ qui minimise $\pi_{\vartheta_1}(\varphi)$, sous la contrainte $\mathbb{P}_{\vartheta_0}\left[\varphi = 1\right] \leq \alpha$.

7.2.2 Lemme de Neyman-Pearson

Dans le cas d'une hypothèse simple contre une alternative simple, un test optimal $\varphi^\star$ existe[3], et on sait le construire explicitement à l'aide du Lemme de Neyman-Pearson.

[1] c'est-à-dire contraint à prendre des valeurs dans l'espace des paramètres $\Theta = \{\vartheta_0, \vartheta_1\}$ ici.

[2] c'est-à-dire vérifiant les hypothèses minimales $\ell(\vartheta,\widetilde{\vartheta}) \geq 0$ pour tous $\vartheta, \widetilde{\vartheta}$ et $\ell(\vartheta,\widetilde{\vartheta}) = 0$ si et seulement si $\vartheta = \widetilde{\vartheta}$.

[3] Pour des raisons de simplicité, on fera dans ce cours une restriction technique, mais le résultat est vrai en toute généralité.

La famille $\{\mathbb{P}_{\vartheta_0}, \mathbb{P}_{\vartheta_1}\}$ est dominée, par exemple par $\mu = \mathbb{P}_{\vartheta_0} + \mathbb{P}_{\vartheta_1}$. Notons

$$f(\vartheta, z) = \frac{d\,\mathbb{P}_\vartheta}{d\mu}(z), \quad z \in \mathfrak{Z},\ \vartheta = \vartheta_0, \vartheta_1$$

les densités associées. Si l'on veut estimer ϑ dans ce contexte, alors l'estimateur du maximum de vraisemblance s'écrit

$$\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} = \vartheta_0 1_{f(\vartheta_1,Z)<f(\vartheta_0,Z)} + \vartheta_1 1_{f(\vartheta_0,Z)<f(\vartheta_1,Z)}$$

et il est bien défini sur l'événement $\{f(\vartheta_0, Z) \neq f(\vartheta_1, Z)\}$, sinon, on ne peut pas dire grand chose. La comparaison de $f(\vartheta_0, Z)$ et $f(\vartheta_1, Z)$ nous fournit donc une règle de décision naturelle. Mais on va un peu raffiner cette régle de décision, pour pouvoir « calibrer » l'erreur de première espèce. Soit $c = c(\alpha) > 0$ à choisir. On décide alors de rejeter H_0 si

$$f(\vartheta_1, Z) \geq c f(\vartheta_0, Z),$$

et on considère la famille des tests de région critique

$$\mathcal{R}_c = \big\{f(\vartheta_1, Z) \geq c f(\vartheta_0, Z)\big\}. \tag{7.3}$$

Le choix de c est réglé par le résultat suivant.

Théorème 7.1 (Lemme de Neyman-Pearson)**.** *Soit $\alpha \in [0, 1]$. S'il existe $c = c(\alpha)$ solution de*

$$\mathbb{P}_{\vartheta_0}\left[f(\vartheta_1, Z) \geq c f(\vartheta_0, Z)\right] = \alpha, \tag{7.4}$$

alors le test de région critique $\mathcal{R}^\star = \mathcal{R}_{c(\alpha)}$ est optimal.

Démonstration. Considérons un test simple de niveau α défini par la région critique $\mathcal{R}$. On a

$$\begin{aligned}\mathbb{P}_{\vartheta_1}\left[Z \in \mathcal{R}^\star\right] - \mathbb{P}_{\vartheta_1}\left[Z \in \mathcal{R}\right] &= \int_{\mathcal{R}^\star} f(\vartheta_1, z)\mu(dz) - \int_{\mathcal{R}} f(\vartheta_1, z)\mu(dz) \\ &= \int_{\mathcal{R}^\star \setminus \mathcal{R}} f(\vartheta_1, z)\mu(dz) - \int_{\mathcal{R} \setminus \mathcal{R}^\star} f(\vartheta_1, z)\mu(dz)\end{aligned}$$

car $f(\vartheta_1, z)\mu(dz) = \mathbb{P}_{\vartheta_1}(dz)$ est une mesure de probabilité. Puisque

$$\mathcal{R}^\star \setminus \mathcal{R} \subset \mathcal{R}^\star,$$

on a, sur cet ensemble

$$f(\vartheta_1, z) > c(\alpha) f(\vartheta_0, z).$$

De même, sur $\mathcal{R} \setminus \mathcal{R}^\star$,

$$f(\vartheta_1, z) \leq c(\alpha) f(\vartheta_0, z).$$

Il vient

$$\begin{aligned}\mathbb{P}_{\vartheta_1}\left[Z\in\mathcal{R}^\star\right]-\mathbb{P}_{\vartheta_1}\left[Z\in\mathcal{R}\right] &\geq c(\alpha)\Big(\int_{\mathcal{R}^\star\setminus\mathcal{R}} f(\vartheta_0,z)\mu(dz)-\int_{\mathcal{R}\setminus\mathcal{R}^\star} f(\vartheta_0,z)\mu(dz)\Big)\\ &= c(\alpha)\Big(\int_{\mathcal{R}^\star} f(\vartheta_0,z)\mu(dz)-\int_{\mathcal{R}} f(\vartheta_0,z)\mu(dz)\Big)\\ &= c(\alpha)\Big(\mathbb{P}_{\vartheta_0}\left[Z\in\mathcal{R}^\star\right]-\mathbb{P}_{\vartheta_0}\left[Z\in\mathcal{R}\right]\Big)\geq 0\end{aligned}$$

où l'on a utilisé cette fois-ci le fait que $f(\vartheta_0,z)\mu(dz)$ est une mesure de probabilité et finalement que $\mathcal{R}^\star$ est de la forme $\mathcal{R}_{c(\alpha)}$ donné par (7.3). □

Définition 7.8 (Test simple de Neyman-Pearson)**.** *Le test simple de l'hypothèse simple $H_0:\vartheta=\vartheta_0$ contre l'alternative simple $H_1:\vartheta=\vartheta_1$ défini*[4] *par la région critique $\mathcal{R}^\star=\mathcal{R}_{c(\alpha)}$ du Théorème 7.1 est appelé test de Neyman-Pearson.*

Corollaire 7.1. *Si $\varphi^\star$ est le test de Neyman-Pearson de niveau α de $H_0:\vartheta=\vartheta_0$ contre $H_1:\vartheta=\vartheta_1$, on a*

$$\pi(\varphi^\star)\geq\alpha.$$

Démonstration. Le test de Neyman-Pearson $\varphi^\star$ est plus puissant que tous les tests de niveau α, en particulier, il est plus puissant que le test artificiel $\varphi=1_{u\leq\alpha}$, où U est une variable aléatoire[5], indépendante de Z, de loi uniforme. En effet,

$$\mathbb{P}_{\vartheta_0}\left[\varphi=1\right]=\alpha$$

Donc φ est de niveau α et puisque $\varphi^\star$ est le test de Neyman-Pearson, on a

$$\pi(\varphi^\star)\geq\pi(\varphi)=\mathbb{P}_{\vartheta_1}\left[\varphi=1\right]=\alpha.$$

□

Remarque 7.3. Une condition suffisante pour que l'équation (7.4) ait une solution est que la variable aléatoire $f(\vartheta_1,Z)/f(\vartheta_0,Z)$ soit bien définie et ait une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$.

Exemple 7.1. Soit F la fonction de répartition d'une loi de probabilité donnée sur $\mathbb{R}$. On considère l'expérience statistique engendrée par un n-échantillon de loi $\mathbb{P}_\vartheta$ de fonction de répartition $F(\bullet-\vartheta)$, où $\vartheta\in\Theta=\{0,\vartheta_0\}$ pour un point $\vartheta_0\neq 0$ de $\mathbb{R}$. On teste $H_0:\vartheta=0$ contre $H_1:\vartheta=\vartheta_0$. Si $X_1,\ldots,X_n$ désigne l'échantillon observé, on a la représentation pour $\vartheta\in\Theta$

$$X_i=\vartheta+\zeta_i,\quad i=1,\ldots,n$$

[4]Cela suppose implicitement qu'une solution $c(\alpha)$ existe, ce qui sera vérifié dans tous nos exemples.

[5]quitte à considérer une bonne extension de l'espace de probabilité sur lequel sont définis les $\mathbb{P}_\vartheta$, on peut toujours faire « exister » une telle variable aléatoire

où les ζ_i sont des variables aléatoires indépendantes, identiquement distribuées, de loi F sous $\mathbb{P}_\vartheta$. Le problème consiste donc à tester l'absence d'un facteur additif $\vartheta = \vartheta_0$ s'ajoutant aux variables ζ_i ou non. Si l'on suppose que F est absolument continue, de densité f et que la variable aléatoire $f(X-\vartheta_0)/f(X)$ a une densité sous $\mathbb{P}_\vartheta$ avec $\vartheta \in \Theta$, alors (7.4) a une solution et le test de Neyman-Pearson a pour zone de rejet

$$\mathcal{R}_{n,\alpha} = \Big\{ \prod_{i=1}^{n} \frac{f(X_i-\vartheta_0)}{f(X_i)} > c(\alpha) \Big\},$$

où le choix de $c(\alpha) > 0$ est reglé par la condition de niveau α du test :

$$\mathbb{P}_0 \Big[\sum_{i=1}^{n} \log \frac{f(X_i-\vartheta_0)}{f(X_i)} > \log c(\alpha) \Big] = \alpha.$$

Lorsque n est grand, on peut calculer une valeur approchée de c_α à l'aide du théorème central-limite.

Exemple 7.2. Considérons une seule observation X de loi de Poisson de paramètre $\vartheta > 0$. on teste $H_0 : \vartheta = \vartheta_0$ contre $H_1 : \vartheta_1$, avec $\vartheta_0 \neq \vartheta_1$. Ici, le test de Neyman-Pearson a pour zone de rejet

$$\mathcal{R}_{n,\alpha} = \Big\{ \exp\big(-(\vartheta_1-\vartheta_0)\big)(\vartheta_1\vartheta_0^{-1})^X \geq c(\alpha) \Big\},$$

où le choix de $c(\alpha)$ garantit que le test est de niveau α. Ici,

$$\mathcal{R}_{n,\alpha} = \Big\{ X > \frac{\log c(\alpha) - (\vartheta_1-\vartheta_0)}{\log\vartheta_1 - \log\vartheta_0} \Big\}.$$

Pour trouver $c(\alpha)$, on doit en principe résoudre

$$\mathbb{P}_{\vartheta_0} \Big[X > \frac{\log c(\alpha) - (\vartheta_1-\vartheta_0)}{\log\vartheta_1 - \log\vartheta_0} \Big] = \alpha,$$

mais la loi de X n'est pas absolument continue, donc cette équation n'a pas de solution en général. On cherche alors le plus petit seuil $c(\alpha) > 0$ de sorte que

$$\mathbb{P}_{\vartheta_0} \Big[X > \frac{\log c(\alpha) - (\vartheta_1-\vartheta_0)}{\log\vartheta_1 - \log\vartheta_0} \Big] \geq \alpha.$$

En pratique, on procède de la manière suivante : par exemple, pour $\vartheta_0 = 5$ et $\alpha = 5\%$, on trouve

$$\mathbb{P}_{\vartheta_0}\big[X > 9\big] = 0,032, \quad \text{et} \quad \mathbb{P}_{\vartheta_0}\big[X > 8\big] = 0,068,$$

et on rejette l'hypothèse si $\{X > 9\}$ et on l'accepte si $\{X \leq 9\}$. Ainsi, l'erreur de première espèce du test est plus petite que $\alpha = 5\%$, mais on ne peut plus garantir que le test est optimal au sens du théorème 7.1.

Remarque 7.4. Il existe une version plus sophistiquée du test de Neyman-Pearson, qui permet de traiter le cas où l'equation (7.4) n'a pas de solution, comme dans l'exemple 7.2. Il faut alors considérer une classe plus large que les tests simples, appelée tests randomisés (voir par exemple [1]).

7.3 Tests d'hypothèses composites

7.3.1 Familles à rapport de vraisemblance monotone*

On fait la restriction – importante ici – $\Theta \subset \mathbb{R}$, et plus précisément, Θ est un intervalle ouvert. On suppose la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ dominée, et on note μ une mesure dominante. Comme d'habitude, on définit la famille de densités

$$f(\vartheta, z) = \frac{d\,\mathbb{P}_\vartheta}{d\mu}(z), \quad z \in \mathfrak{Z}, \quad \vartheta \in \Theta.$$

L'hypothèse de travail dans toutes cette section est

Hypothèse 7.1. *Pour tout $\vartheta \in \Theta$, on a $f(\vartheta, z) > 0$, $\mu(dz)$ presque-partout.*

Soit $\widetilde{\vartheta} \in \Theta$ un point arbitraire de l'ensemble des paramètres. On souhaite tester une hypothèse nulle de la forme

$$H_0 : \quad \vartheta \leq \widetilde{\vartheta}$$

contre l'alternative

$$H_1 : \quad \vartheta > \widetilde{\vartheta}.$$

Si l'on souhaite appliquer le résultat de Neyman-Pearson, il faut, d'une certaine manière, pouvoir traiter tous les tests de l'hypothèse simple $H_0 : \vartheta = \vartheta_0$ contre l'alternative $H_1 : \vartheta = \vartheta_1$ simultanément pour tous les $\vartheta_0 \leq \widetilde{\vartheta}$ et $\vartheta_1 \geq \widetilde{\vartheta}$.

L'hypothèse suivante va permettre de nous ramener à cette situation :

Définition 7.9. *Sous l'Hypothèse 7.1, la famille de densité $\{f(\vartheta, \bullet), \vartheta \in \Theta\}$, avec $\Theta \subset \mathbb{R}$, est dite à rapport de vraisemblance monotone s'il existe une application $T : \mathfrak{Z} \to \mathbb{R}$ mesurable, de sorte que pour tous $\vartheta_1 < \vartheta_2$,*

$$\frac{f(\vartheta_2, Z)}{f(\vartheta_1, Z)} \quad \textit{est une fonction monotone de } T(Z).$$

Remarque 7.5. Quitte à changer T en $-T$, on peut toujours supposer que cette fonction est croissante.

Théorème 7.2 (Lehmann)**.** *Soi $\alpha \in [0, 1]$ un niveau de risque. On suppose que $\Theta \subset \mathbb{R}$ est un intervalle ouvert et que la famille $\{f(\vartheta, \bullet), \vartheta \in \Theta\}$ satisfait l'Hypothèse 7.1 et est à rapport de vraisemblance monotone. S'il existe une solution $c = c(\widetilde{\vartheta}, \alpha) > 0$ à*

$$\mathbb{P}_{\widetilde{\vartheta}}\big[T(Z) > c\big] = \alpha,$$

alors le test de région de rejet

$$\mathcal{R}^\star = \big\{T(Z) > c(\widetilde{\vartheta}, \alpha)\big\}$$

est de niveau α pour tester $H_0 : \vartheta \leq \widetilde{\vartheta}$ contre $H_1 : \vartheta > \widetilde{\vartheta}$, et de puissance maximale parmi tous les tests de niveau α.

Démonstration. C'est une adaptation de la preuve du Lemme de Neyman-Pearson. L'hypothèse d'une famille à rapport de vraisemblance monotone se traduit par la propriété suivante : l'inéquation

$$\text{pour tous } \vartheta > \widetilde{\vartheta}, \quad \frac{f(\vartheta, Z)}{f(\widetilde{\vartheta}, Z)} > c$$

se résout sous la forme

$$T(Z) \geq \kappa(\widetilde{\vartheta}, \vartheta, c)$$

pour une certaine fonction κ. Notons $\varphi^\star$ le test simple de région critique $\mathcal{R}^\star$ et soit $\vartheta' > \widetilde{\vartheta}$ un point arbitraire de l'alternative. Montrons que la puissance $\pi_{\vartheta'}(\varphi^\star)$ est maximale parmi tous les tests de niveau α pour tester H_0 contre H_1.

Si l'on considère le test de l'hypothèse simple $\vartheta = \widetilde{\vartheta}$ contre l'alternative simple $\vartheta = \vartheta'$, on sait que le test de Neyman-Pearson

$$\varphi^{\mathrm{NP}} = 1_{\frac{f(\vartheta', Z)}{f(\widetilde{\vartheta}, Z)} > c(\alpha, \vartheta', \widetilde{\vartheta})},$$

où $c(\alpha, \vartheta', \widetilde{\vartheta})$ est la constante du Théorème 7.1, a la puissance maximale parmi tous les tests de niveau α. D'après notre remarque préliminaire, il s'écrit aussi sous la forme

$$\varphi^{\mathrm{NP}} = 1_{T(Z) \geq \kappa\left(\widetilde{\vartheta}, \vartheta, c(\alpha, \vartheta', \widetilde{\vartheta})\right)},$$

et $c(\alpha, \vartheta', \widetilde{\vartheta})$ est déterminée par la condition

$$\mathbb{P}_{\widetilde{\vartheta}}\left[T(Z) \geq \kappa\big(\widetilde{\vartheta}, \vartheta, c(\alpha, \vartheta', \widetilde{\vartheta})\big)\right] = \alpha,$$

s'il existe. C'est le cas, d'après les hypothèses, et on a aussi

$$\kappa\big(\widetilde{\vartheta}, \vartheta, c(\alpha, \vartheta', \widetilde{\vartheta})\big) = c(\alpha, \vartheta', \widetilde{\vartheta}).$$

d'après le Lemme de Neyman-Pearson. On en déduit que $\varphi^\star$ a une erreur de seconde espèce maximale au point ϑ' parmi tous les tests de niveau α, et donc uniformément sur l'alternative.

Il reste à montrer que $\varphi^\star$ est bien de niveau α. Soit $\vartheta'' \leq \widetilde{\vartheta}$ un point arbitraire de l'hypothèse nulle. Posons

$$\alpha' = \mathbb{P}_{\vartheta''}\left[\varphi^\star = 1\right].$$

Alors α' est le niveau du test $\varphi^\star$ utilisé pour tester l'hypothèse nulle $\vartheta = \vartheta''$ contre l'alternative $\vartheta = \widetilde{\vartheta}$. Alors, comme précédemment, le Lemme de Neyman-Pearson entraîne que $\varphi^\star$ est optimal pour tester $\vartheta = \vartheta''$ contre l'alternative $\vartheta = \widetilde{\vartheta}$ au niveau α'.

Finalement, le Corollaire 7.1 implique que la puissance de $\varphi^\star$ est plus grande que α', c'est-à-dire

$$\pi(\varphi^\star) = 1 - \mathbb{P}_{\widetilde{\vartheta}}\left[\varphi^\star = 0\right] \geq \mathbb{P}_{\vartheta''}\left[\varphi^\star = 1\right],$$

c'est-à-dire

$$\mathbb{P}_{\vartheta''}\left[\varphi^\star = 1\right] \leq \mathbb{P}_{\widetilde{\vartheta}}\left[\varphi^\star = 0\right] = \alpha.$$

Comme ϑ'' est arbitraire, le théorème est démontré. □

7.3.2 Exemples

Exemple 7.3. On observe $X_1, \dots, X_n$ indépendantes, de loi $\mathcal{N}(\vartheta, \sigma^2)$, où σ^2 est connu, et $\vartheta \in \Theta = \mathbb{R}$. On teste $H_0 = \vartheta = \vartheta_0$ contre $H_1 : \vartheta = \vartheta_1$, avec $\vartheta_0 < \vartheta_1$. On a $Z = (X_1, \dots, X_n)$, et on prend pour mesure dominante μ la mesure de Lebesgue sur $\mathbb{R}^n$. Si $g(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ désigne la densité de la loi gaussienne standard sur $\mathbb{R}$, on a

$$\begin{aligned} f(\vartheta, Z) &= \sum_{i=1}^{n} g(\vartheta - X_i) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \vartheta)^2 \big) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} X_i^2 + \frac{n}{\sigma^2} \overline{X}_n - \frac{n\vartheta^2}{2\sigma^2} \big) \end{aligned}$$

d'où

$$\frac{f(\vartheta_1, Z)}{f(\vartheta_0, Z)} = \exp\big(\frac{n}{\sigma^2} (\vartheta_1 - \vartheta_0)^2 \overline{X}_n \big) \exp\big(- \frac{n}{2\sigma^2} (\vartheta_1^2 - \vartheta_0^2) \big).$$

La zone de rejet du test de Neyman-Pearson s'écrit

$$\begin{aligned} \big\{ f(\vartheta_1, Z) > c f(\vartheta_0, Z) \big\} &= \big\{ \frac{n}{\sigma^2} (\vartheta_1 - \vartheta_0) \overline{X}_n - \frac{n}{2\sigma^2} (\vartheta_1^2 - \vartheta_0^2) > c \big\} \\ &= \big\{ \overline{X}_n > \frac{\vartheta_0 + \vartheta_1}{2} + \frac{\sigma^2 \log c}{n(\vartheta_0 - \vartheta_1)} \big\}. \end{aligned}$$

Le choix de c est réglé par l'équation

$$\mathbb{P}_{\vartheta_0} \big[\overline{X}_n > \frac{1}{2} (\vartheta_0 + \vartheta_1) + \frac{\sigma^2 \log c}{n(\vartheta_0 - \vartheta_1)} \big] = \alpha. \tag{7.5}$$

Sous $\mathbb{P}_{\vartheta_0}$, les X_i sont distribuées comme des variables aléatoires gaussiennes indépendantes, de moyenne ϑ_0 et de variance σ^2. Donc, sous $\mathbb{P}_{\vartheta_0}$, on peut écrire

$$\overline{X}_n = \vartheta_0 + \frac{\sigma}{\sqrt{n}} \xi^{(\vartheta_0)}, \tag{7.6}$$

où la loi de $\xi^{(\vartheta_0)}$ –sous $\mathbb{P}_{\vartheta_0}$– est la loi gaussienne standard $\mathcal{N}(0, 1)$. Donc l'équation (7.5) est équivalente à

$$\mathbb{P}_{\vartheta_0} \big[\xi^{(\vartheta_0)} > \frac{\sqrt{n}}{2\sigma} (\vartheta_1 - \vartheta_0) + \frac{\sigma}{\sqrt{n}} \frac{\log c}{\vartheta_0 - \vartheta_1} \big] = \alpha,$$

soit

$$\frac{\sqrt{n}}{2\sigma} (\vartheta_1 - \vartheta_0) + \frac{1}{\sqrt{n}} \frac{\sigma \log c}{\vartheta_0 - \vartheta_1} = \Phi^{-1}(1 - \alpha)$$

où $\Phi(x)$ désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$, d'où finalement

$$c = \exp\big(-\frac{(\vartheta_1-\vartheta_0)^2}{2} + \frac{\sqrt{n}}{\sigma}(\vartheta_0-\vartheta_1)\Phi^{-1}(1-\alpha)\big).$$

Exemple 7.4. Dans le même contexte, on a bien, pour $\vartheta > \widetilde{\vartheta}$

$$\frac{f(\vartheta_1,Z)}{f(\vartheta_0,Z)} = \exp\Big(\frac{n(\vartheta-\widetilde{\vartheta})}{\sigma^2}T(X_1,\ldots,X_n) - \frac{n}{2\sigma^2}(\vartheta^2-\widetilde{\vartheta}^2)\Big),$$

avec $T(X_1,\ldots,X_n) = \overline{X}_n$. La famille $\{f(\vartheta,\bullet), \vartheta\in\mathbb{R}\}$ est à rapport de vraisemblance monotone, et un test optimal (uniformément plus puissant) de $H_0 : \vartheta \leq \widetilde{\vartheta}$ contre $H_1 : \vartheta > \widetilde{\vartheta}$ est donné par la région critique

$$\mathcal{R} = \big\{\overline{X}_n > c\big\},$$

où $c = c(\widetilde{\vartheta},\alpha)$ est calibré par l'équation

$$\mathbb{P}_{\widetilde{\vartheta}}\big[\overline{X}_n > c\big] = \alpha,$$

soit, d'après 7.6 en remplaçant ϑ_0 par $\widetilde{\vartheta}$,

$$\mathbb{P}_{\widetilde{\vartheta}}\Big[\xi^{(\widetilde{\vartheta})} > \frac{\sqrt{n}}{\sigma}(c-\widetilde{\vartheta})\Big] = \alpha,$$

où la loi de $\xi^{(\widetilde{\vartheta})}$ sous $\mathbb{P}_{\widetilde{\vartheta}}$ est la loi $\mathcal{N}(0,1)$. D'où

$$c = c(\widetilde{\vartheta},\alpha) = \widetilde{\vartheta} + \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

On peut expliciter sur cet exemple la puissance du test optimal

$$\varphi^\star = 1_{\overline{X}_n > \widetilde{\vartheta} + \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}}.$$

On a, pour tout point de l'alternative $\vartheta > \widetilde{\vartheta}$, en utilisant une fois de plus la représentation $\overline{X}_n = \vartheta + \frac{\sigma}{\sqrt{n}}\xi^\vartheta$, où la loi de $\xi^{(\vartheta)}$ sous $\mathbb{P}_\vartheta$ est la loi $\mathcal{N}(0,1)$

$$\begin{aligned}
\pi_\vartheta(\varphi^\star) &= \mathbb{P}_\vartheta\big[\vartheta + \frac{\sigma}{\sqrt{n}}\xi^{(\vartheta)} > \widetilde{\vartheta} + \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}\big] \\
&= \mathbb{P}_\vartheta\Big[\xi^{(\vartheta)} > \frac{\sqrt{n}}{\sigma}(\widetilde{\vartheta}-\vartheta) + \sigma\Phi^{-1}(1-\alpha)\Big] \\
&= 1-\Phi\Big(\frac{\sqrt{n}}{\sigma}(\widetilde{\vartheta}-\vartheta) + \sigma\Phi^{-1}(1-\alpha)\Big) \\
&= \Phi\Big(\frac{\sqrt{n}}{\sigma}(\vartheta-\widetilde{\vartheta}) - \sigma\Phi^{-1}(1-\alpha)\Big)
\end{aligned}$$

en utilisant l'identité $1-\Phi(x) = \Phi(-x)$ (qui traduit simplement le fait que la loi gaussienne standard est symétrique).

Remarque 7.6. Hormis quelques cas particuliers comme les familles à rapport de vraisemblance monotone[6] on ne sait pas en général exhiber de tests optimaux au sens

[6]et le cas des échantillons gaussiens étudiés dans le Section 7.6.1 plus loin.

de Neyman lorsque l'hypothèse nulle ou l'alternative sont composites. Pour développer une théorie générale, nous nous placerons – comme pour l'estimation – dans un cadre asymptotique dès le Chapitre 8.

7.4 $p-$valeur

7.4.1 Notion de $p-$valeur

Intrdocution sur un exemple

Reprenons l'Exemple 7.4 avec $\widetilde{\vartheta} = 0$, où l'on teste au niveau α l'hypothèse nulle $H_0 : \vartheta \leq 0$ contre l'alternative $H_1 : \vartheta > 0$. La règle de décision (optimale) prend la forme

$$\text{« On rejette l'hypothèse } H_0 \quad \text{si} \quad \overline{X}_n > \sigma \frac{\Phi^{-1}(1-\alpha)}{\sqrt{n}} \quad \text{ ».}$$

Si les observations X_i sont indépendantes, ont un moment d'ordre 2, et si n est grand, alors cette approche est plausible. Toutefois, on ne connaît pas σ en général, mais on peut l'estimer par $\widehat{\sigma}_n$, de sorte qu' en pratique, on va rejeter l'hypothèse si

$$\overline{X}_n > \widehat{\sigma}_n \frac{\Phi^{-1}(1-\alpha)}{\sqrt{n}}. \tag{7.7}$$

On se donne sa valeur de α favorite, par exemple 5%, et on effectue le test : on accepte ou on rejette, en fonction du nombre de données n, des valeurs calculées à partir des observations $\overline{X}_n$, $\widehat{\sigma}$ et de la valeur α choisie, selon la règle de décision (7.7).

Imaginons que l'on rejette l'hypothèse. Qu'aurions nous fait pour la choix de $\alpha = 1\%$? ou $\alpha = 1/1000$, etc. ? En prenant α est de plus en plus petit, il y a fatalement un seuil α à partir duquel on va systématiquement accepter l'hypothèse : pour se garder contre l'erreur de première espèce, on est prêt à augmenter les faux positifs[7].

Définition de la p-valeur d'un test

En pratique, accepter ou rejeter l'hypothèse n'a que peu de signification scientifique, surtout si α est proche du seuil limite ou la décision va basculer : en baissant α on accepte l'hypothèse (ou bien en augmentant α on rejette l'hypothèse). Par contre, le seuil de basculement de la décision (qui dépend des observations) a une signification et une interprétation : c'est ce que l'on appelle la p-valeur du test.

[7]Dans le cas limite $\alpha = 0$, on ne peut pas se permettre de rejeter l'hypothèse à tort, et ceci « oblige » le test à accepter systématiquement l'hypothèse

Définition 7.10 (p–valeur)**.** *Soit, pour tout $\alpha \in [0,1]$, une famille de tests simples φ_α de niveau α pour tester l'hypothèse H_0 contre l'alternative H_1. On note $\mathcal{R}_\alpha$ la zone de rejet de φ_α. On appelle p-valeur du test la quantité*

$$p - \text{valeur}(Z) = \inf\{\alpha,\ Z \in \mathcal{R}_\alpha\}$$

La p-valeur d'un test (de la famille de tests indicée par le niveau α) est le plus petit niveau pour lequel on rejette H_0.

Règle d'interprétation

On est confiant vis-à-vis de la décision de ne pas rejeter H_0 lorsque la p-valeur du test est grande. Voic quelques interprétations courantes qui sévissent dans les applications (extait du livre de Wasserman [8]) de l'inteprétation des ordres de grandeur des p-valeur

p–valeur	suspicion de rejet
< 0.01	suspicion très forte contre H_0
$0.01 - 0.05$	suspicion forte contre H_0
$0.05 - 0.1$	suspicion faible contre H_0
> 0.1	peu ou pas de suspicion contre H_0

Attention ! Une p-valeur grande n'est pas un indicateur en faveur de l'acceptation de l'hypothèse H_0, mais plutôt en faveur du non-rejet (suggérant en pratique d'envisager d'autres tests plus précis ou plus coûteux). Une p-valeur peut être grande pour deux raisons :

- effectivement, l'hypothèse H_0 est vraie
- l'hypothèse H_0 n'est pas vraie, mais le test est très peu puissant (beaucoup de faux positifs) et son erreur de seconde espèce est grande.

Concernant la seconde raison, prenons par exemple le test trivial $\varphi = 1_\emptyset$. Sa p-valeur vaut 1 et prend donc la plus grande valeur possible.

7.4.2 Propriétés de la p-valeur

On peut préciser –un peu– le sens mathématique des remarques précédentes. On se retreint au cas où l'hypothèse nulle est simple : on teste $H_0 : \vartheta = \vartheta_0$ contre $H_1 = \vartheta \neq \vartheta_0$.

Proposition 7.1. *Si $\{\varphi_\alpha, 0 \leq \alpha \leq 1\}$ est une famille de tests exactement*[8] *de niveau α dont la zone de rejet est de la forme*

$$\mathcal{R}_\alpha = \big\{T(Z) \geq c_\alpha\big\}$$

[8] au sens où l'erreur de première espèce vaut exactement α.

pour une certaine application $T : \mathfrak{Z} \to \mathbb{R}$ *mesurable. Alors, si* $\widetilde{Z}$ *désigne une copie indépendante de* Z*, on a*

$$p - \text{valeur}(Z) = \mathbb{P}_{\vartheta_0}\left[T(\widetilde{Z}) \geq T(Z) \,|\, Z\right].$$

De plus, si la loi de $T(Z)$ *est absolument continue sous* $\mathbb{P}_{\vartheta_0}$*, alors la loi de* $p - \text{valeur}(Z)$ *est uniforme sous* $\mathbb{P}_{\vartheta_0}$*.*

Le premier résultat de la Proposition 7.1 s'interprète de la façon suivante : la p-valeur est la probabilité sous $\mathbb{P}_{\vartheta_0}$ qu'une observation $T(\widetilde{Z})$ d'une expérience « copie » soit supérieure à ce que l'on a observé, c'est-à-dire $T(Z)$.

Démonstration. l'application $c_\bullet : [0,1] \to \overline{\mathbb{R}}$ est décroissante et $c_0 = +\infty$ et $c_1 = +\infty$. On a l'identité

$$c_{p-\text{valeur}(Z)} = T(Z).$$

Il vient

$$\begin{aligned}\mathbb{P}_{\vartheta_0}\left[T(\widetilde{Z}) \geq T(Z) \,|\, Z\right] &= \mathbb{P}_{\vartheta_0}\left[T(\widetilde{Z}) \geq c_{p-\text{valeur}(Z)} \,|\, Z\right] \\ &= p - \text{valeur}(Z)\end{aligned}$$

par définition de l'erreur de première espèce, en utilisant l'hypothèse que le test φ_α est exactement de niveau α

La seconde partie de la proposition est standard. Si F désigne la fonction de répartition de $T(\widetilde{Z})$, posons

$$Y = \mathbb{P}_{\vartheta_0}\left[T(\widetilde{Z}) \leq T(Z) \,|\, Z\right] = F\big(T(\widetilde{Z})\big).$$

Alors, pour tout réel x, on a

$$\begin{aligned}\mathbb{P}_{\vartheta_0}\left[Y \leq x\right] &= \mathbb{P}_{\vartheta_0}\left[F\big(T(Z)\big) \leq x\right] \\ &= \mathbb{P}_{\vartheta_0}\left[T(Z) \leq F^{-1}(x)\right] \\ &= F\big(F^{-1}(x)\big) = x\end{aligned}$$

si $x \in [0,1]$, et où $F^{-1}(x) = \inf\{t \in \mathbb{R},\, F(t) \geq x\}$ (Méléard [4], paragraphe 4.2.4 p. 78). Si $x \leq 0$, la probabilité ci-dessus vaut 0 et si $x > 1$, elle vaut 1. Donc la loi de Y sous $\mathbb{P}_{\vartheta_0}$ est uniforme sur $[0,1]$, ce qui achève la démonstration. □

7.5 Régions de confiance

On a déjà construit des intervalles de confiance dans le contexte de la précision d'estimation pour le modèle d'échantillonnage général du Chapitre 3. On formalise –un peu– dans cette section la notion et le lien naturel avec les tests d'hypothèse, que nous avons déja utilisé au Chapitre 3.

Situation

On considère l'expérience statistique engendrée par l'observation d'un n-échantillon $X_1, \ldots, X_n$ où la variable aléatoire réelle X_i suit la loi $\mathbb{P}_\vartheta$, avec $\Theta \subset \mathbb{R}^d$, $d \geq 1$. On peut immédiatement généraliser ce qui va suivre à une experience statistique arbitraire, avec un simple coût notationnel.

7.5.1 Région de confiance

Définition 7.11. *Soit $\alpha \in [0,1]$. Une région de confiance de niveau $1-\alpha$ pour le paramètre $\vartheta \in \Theta$ est un ensemble*

$$\mathcal{C} = \mathcal{C}_\alpha(X_1, \ldots, X_n) \subset \mathbb{R}^d,$$

tel que,

$$\forall \vartheta \in \Theta, \quad \mathbb{P}_\vartheta \left[\vartheta \in \mathcal{C}(X_1, \ldots, X_n)\right] \geq 1 - \alpha. \tag{7.8}$$

La propriété (7.8) est appelée « propriété de couverture » de la région $\mathcal{C}_\alpha(X_1, \ldots, X_n)$. Bien qu'en principe arbitraire, on construit en pratique des régions de confiance très particulières. Si $\Theta \subset \mathbb{R}$, on utilise le plus souvent des intervalles. Construire un intervalle de confiance de niveau $1-\alpha$ revient alors à se donner deux statistiques $g_\alpha(X_1, \ldots, X_n)$ et $d_\alpha(X_1, \ldots, X_n)$ avec

$$g_\alpha(X_1, \ldots, X_n) \leq d_\alpha(X_1, \ldots, X_n)$$

telles que, pour tout $\vartheta \in \Theta$,

$$\mathbb{P}_\vartheta \left[g_\alpha(X_1, \ldots, X_n) \leq \vartheta \leq d_\alpha(X_1, \ldots, X_n)\right] \geq 1 - \alpha.$$

Posé comme cela, la construction des statistiques $g_\alpha(X_1, \ldots, X_n)$ et $d_\alpha(X_1, \ldots, X_n)$ n' pas d'intérêt : n'imprte quel intervalle qui contient Θ conviendra. La qualité d'un intervalle de confiance de niveau $1-\alpha$ se mesurera à sa longueur (en générale aléatoire) que l'on cherche à rendre la plus petite possible, sous la contrainte de la propriété de couverture. Dans ce sens, la problématique des tests et des intervalles de confiance est similaire.

7.5.2 Fonctions pivotales : le cas non-asymptotique

Dans le cas particulier où l'ensemble des paramètres Θ est de dimension 1, nous examinons une méthode de construction de régions de confiance, très particulière, mais qui sera mise en oeuvre de manière plus systématique dans le cadre asymptotique (voir 8.2). Elle est fortement apparentée à la construction des tests.

Supposons que l'on dispose d'une variable aléatoire[9] $S(\vartheta, X_1, \dots, X_n)$ à valeurs dans $\mathbb{R}$ dont la loi sous $\mathbb{P}_\vartheta$ ne dépende pas de ϑ. En particulier, pour tout intervalle I de $\mathbb{R}$, la probabilité

$$\mathbb{P}_\vartheta \left[S(\vartheta, X_1, \dots, X_n) \in I \right]$$

ne dépend pas de ϑ.

Définition 7.12. *On appelle pivot toute variables aléatoire $S(\vartheta, X_1, \dots, X_n)$ dont la loi ne dépend pas de ϑ.*

Exemple 7.5.

1. Si $X_1, \dots, X_n$ sont indépendantes, de même loi $\mathcal{N}(\vartheta, \sigma^2)$, où σ^2 est connu et $\vartheta \in \Theta = \mathbb{R}$ est le paramètre inconnu, alors

$$S(\vartheta, X_1, \dots, X_n) = \frac{\overline{X}_n - \vartheta}{\sigma}$$

   est pivotale.

2. Si $X_1, \dots, X_n$ sont indépendantes, de même loi exponentielle de paramètre ϑ, où $\vartheta \in \mathbb{R}_+ \setminus \{0\}$ est le paramètre, alors $S(\vartheta, X_1, \dots, X_n) = \vartheta \overline{X}_n$ est pivotale : en effet,

$$S(\vartheta, X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \vartheta X_i.$$

   La loi de X sous $\mathbb{P}_\vartheta$ est exponentielle de paramètre ϑ. Sa densité par rapport à la mesure de Lebesgue s'écrit $\vartheta g(\vartheta \bullet)$, où $g(x) = \exp(-x) 1_{\{x \in \mathbb{R}_+\}}$ est la densité de la loi exponentielle de paramètre 1. De manière générale, si X a pour densité f par rapport à la mesure de Lebesgue, alors ϑX a pour densité $\vartheta^{-1} f(\vartheta^{-1} \bullet)$ si $\vartheta \neq 0$. Donc ϑX a pour densité $g(\bullet)$ qui ne dépend pas de ϑ. Par suite, la loi de $S(\vartheta, X_1, \dots, X_n)$ ne dépend pas de ϑ.

Une méthode de construction est la suivante. Soit ξ une variable aléatoire de même loi que le pivot. Pour $\alpha \in [0, 1]$, on considère la classe des intervalles $I_\alpha \subset \mathbb{R}$ vérifiant

$$\mathbb{P}_\vartheta \left[S(\vartheta, X_1, \dots, X_n) \in I_\alpha \right] = \mathbb{P}_\vartheta \left[\xi \in I_\alpha \right] \geq 1 - \alpha \tag{7.9}$$

Alors la région

$$\mathcal{I}_\alpha = \left\{ \vartheta \in \Theta, S(\vartheta, X_1, \dots, X_n) \in I_\alpha \right\}$$

est une région de confiance pour ϑ de niveau $1 - \alpha$. On est alors ramené à choisir dans la classe des intervalles I_α satisfaisant (7.9) de sorte que le diamètre de $\mathcal{I}_\alpha$ soit le plus petit possible.

[9] Attention : $S(\vartheta, X_1, \dots, X_n)$ dépend de ϑ : elle n'est pas observable et ce n'est pas une statistique.

Méthode générique de construction d'un pivot

Dans les deux exemples précédents, les pivots se basent sur des estimateurs préliminaires du paramètre ϑ. Si $\widehat{\vartheta}_n$ est un estimateur de ϑ, une méthode générique de construction d'un pivot est la suivante.

On note $x \rightsquigarrow \Gamma_\vartheta(x) = \mathbb{P}_\vartheta\left[\,\widehat{\vartheta}_n \leq x\right]$, la fonction de répartition de $\widehat{\vartheta}_n$ au point ϑ.

Proposition 7.2. *Si*
(i) $\vartheta \rightsquigarrow \Gamma_\vartheta(x)$ *est monotone pour tout* $x \in \mathbb{R}$,
(ii) $x \rightsquigarrow \Gamma_\vartheta(x)$ *est continue pour tout* $\vartheta \in \Theta$,
alors

$$S(\vartheta, X_1, \ldots, X_n) = \Gamma_\vartheta(\widehat{\vartheta}_n)$$

est un pivot de loi uniforme sur $[0,1]$. *En particulier, pour tout* $\alpha \in [0,1]$

$$\mathbb{P}_\vartheta\left[\frac{\alpha}{2} \leq \Gamma_\vartheta(\widehat{\vartheta}_n) \leq 1 - \frac{\alpha}{2}\right] = 1 - \alpha$$

et

$$\mathcal{I}_\alpha = \left[\Gamma^{-1}_{\alpha/2}, \Gamma^{-1}_{1-\alpha/2}\right]$$

est un intervalle de confiance pour ϑ *de niveau* $1-\alpha$.

Remarque 7.7. De même, pour tout $\rho \in [0,1]$,

$$\mathcal{I}^{(\rho)}_\alpha = \left[\Gamma^{-1}_{\rho\alpha}, \Gamma^{-1}_{(1-\rho)\alpha}\right]$$

et on peut chercher la valeur ρ qui minimise $\Gamma^{-1}_{(1-\rho)\alpha} - \Gamma^{-1}_{\rho\alpha}$ pour trouver le meilleur intervalle de confiance parmi la classe des estimateurs donnés par le pivot.

7.5.3 Dualité tests – régions de confiance

Il existe un lien naturel entre intervalles de confiances et tests que nous avons déjà mis en évidence au Chapitre 3.

Un exemple illustratif

Considérons l'expérience statistique engendrée par l'observation de $X_1, \ldots, X_n$, indépendantes et de même loi $\mathcal{N}(\vartheta, \sigma^2)$, où $\sigma^2 > 0$ est connu et $\vartheta \in \Theta = \mathbb{R}$ est le paramètre inconnu. Soit $\alpha \in [0,1]$. Posons, pour $\vartheta_0 \in \Theta$,

$$\mathcal{A}_\alpha(\vartheta_0) = \left\{ \left|\vartheta_0 - \overline{X}_n\right| \leq \frac{\sigma}{\sqrt{n}} \Phi^{-1}\big(1 - \frac{\alpha}{2}\big) \right\}$$

et

$$\mathcal{R}_\alpha(\vartheta_0) = \left\{ \left|\vartheta_0 - \overline{X}_n\right| > \frac{\sigma}{\sqrt{n}} \Phi^{-1}\big(1 - \frac{\alpha}{2}\big) \right\}.$$

Alors l'ensemble $\mathcal{R}_\alpha(\vartheta_0)$ s'interprète naturellement comme la zone de rejet d'un test de niveau α pour l'hypothèse

$$H_0 : \vartheta = \vartheta_0, \quad \text{contre} \quad H_1 : \vartheta \neq \vartheta_0.$$

De plus, $\mathcal{A}_\alpha(\vartheta_0) = \mathcal{R}_\alpha(\vartheta_0)^c$ correspond à la zone où l'on accepte l'hypothèse.

Proposition 7.3. *Si, pour tout $\vartheta_0 \in \Theta$, il existe une test de niveau α et de zone de rejet $\mathcal{R}_\alpha(\vartheta_0)$ de l'hypohèse nulle $H_0 : \vartheta = \vartheta_0$ contre l'alternative $H_1 : \vartheta \neq \vartheta_0$, alors, pour tout $\vartheta \in \Theta$*

$$\mathcal{C} = \mathcal{C}_\alpha(X_1, \dots, X_n) = \left\{ \vartheta \in \Theta,\ (X_1, \dots, X_n) \in \mathcal{R}_\alpha(\vartheta)^c \right\}$$

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

Réciproquement, si $\mathcal{C}_\alpha(X_1, \dots, X_n)$ est une région de confiance de niveau $1 - \alpha$ pour le paramètre $\vartheta \in \Theta$, alors, le test de l'hypothèse nulle $H_0 : \vartheta = \vartheta_0$ contre l'alternative $\vartheta \neq \vartheta_0$ de région critique

$$\mathcal{R}_\alpha(\vartheta_0) = \left\{ (X_1, \dots, X_n) \in \mathcal{C}_\alpha^c \right\}$$

est de niveau α.

Démonstration. On a

$$\begin{aligned} \mathbb{P}_\vartheta \left[\vartheta \in \mathcal{C}(X_1, \dots, X_n) \right] &= \mathbb{P}_\vartheta \left[(X_1, \dots, X_n) \in \mathcal{R}(\vartheta_0)^c \right] \\ &= 1 - \mathbb{P}_\vartheta \left[(X_1, \dots, X_n) \in \mathcal{R}(\vartheta_0) \right] \\ &= \geq 1 - \alpha. \end{aligned}$$

Réciproquement, il suffit de noter que pour tout $\vartheta_0 \in \Theta$, on a

$$\begin{aligned} \mathbb{P}_{\vartheta_0} \left[(X_1, \dots, X_n) \in \mathcal{R}(\vartheta_0) \right] &= 1 - \mathbb{P}_{\vartheta_0} \left[(X_1, \dots, X_n) \in \mathcal{R}^c \right] \\ &= 1 - \mathbb{P}_{\vartheta_0} \left[(X_1, \dots, X_n) \in \mathcal{C} \right] \leq \alpha. \end{aligned}$$

□

Remarque 7.8. Ce résultat, relativement immédiat, ne nous dit rien sur la puissance du test d'une part, ni sur la qualité (le diamètre) de la région de confiance d'autre part. Ces deux notions sont évidemment étroitement liées.

7.6 Tests dans le modèle de régression linéaire

7.6.1 Echantillons gaussiens

Situation

Dans toute cette section, on considère l'expérience statistique engendrée par un n-échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$, où $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \subset \{0\}$. Il y a coïncidence

dans ce cas très simple avec le modèle de régression linéaire à « design » déterministe : les observations sont $\mathbf{Y} = (Y_1, \dots, Y_n)$ et on a a représentation

$$\mathbf{Y} = \mathbb{M}\,\mu + \sigma\boldsymbol{\xi}, \tag{7.10}$$

où

$$\mathbb{M} = (1 \dots 1)^T \ \ (n \text{ fois}) \text{ et } \ \ \boldsymbol{\xi} = (\xi_1 \dots \xi_n)^T,$$

les ξ_i étant sous $\mathbb{P}_\vartheta$ des variables gaussiennes standard. L'estimateur du maximum de vraisemblance, est

$$\begin{aligned}\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}} &= \big(\widehat{\mu}_n^{\,\mathtt{mv}}, (\widehat{\sigma}_n^2)^{\,\mathtt{mv}}\big) \\ &= \Big(\overline{Y}_n, \tfrac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2\Big),\end{aligned}$$

voir Chapitre 5, Proposition 5.5. Une autre manière –peut-être plus naturelle dans ce contexte– est de maximiser directement la log-vraisemblance

$$\ell_n\big((\mu,\sigma^2), Y_1, \dots, Y_n\big) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2.$$

On a

$$\begin{cases} \partial_\mu \ell_n\big((\mu,\sigma^2), Y_1, \dots, Y_n\big) &= \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu) \\ \\ \partial_{\sigma^2} \ell_n\big((\mu,\sigma^2), Y_1, \dots, Y_n\big) &= -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n (Y_i - \mu)^2, \end{cases}$$

ce qui nous fournit le point critique

$$\widehat{\vartheta}_n = \big(\overline{Y}_n, \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2\big).$$

On vérifie ensuite que le point critique est l'unique maximum global et donc $\widehat{\vartheta}_n = \widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$. Un estimateur sans biais de σ^2 est

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2 = \frac{n}{n-1}(\widehat{\sigma}_n^2)^{\,\mathtt{mv}}.$$

Les propriétés vecteurs gaussiens et de lois dérivées étudiées au Chapitre 1 nous donnent gratuitement la loi jointe de $(\overline{Y}_n, s_n^2)$.

Lemme 7.6.1. *Sous $\mathbb{P}_\vartheta$, les variables $\overline{Y}_n$ et s_n^2 sont indépendantes. De plus, $\overline{Y}_n$ suit la loi $\mathcal{N}\big(\mu, \frac{\sigma^2}{n}\big)$ et $(n-1)\frac{s_n^2}{\sigma^2}$ suit la loi du χ^2 à $n-1$ degrés de liberté.*

Démonstration. C'est une application de la Proposition 5.10 qui repose sur la Proposition 1.1 (Cochran) du Chapitre 1. □

Batterie de tests classiques

Soit $\mu_0 \in \mathbb{R}$ et $\sigma_0^2 > 0$ donnés.

1. On teste
$$H_0 : \mu \leq \mu_0 \quad \text{contre} \quad H_1 : \mu > \mu_0.$$
Un test de niveau α est donné par la zone de rejet
$$\mathcal{R}_\alpha = \left\{ T(\mathbf{Y}) > q^{\mathsf{t}}_{1-\alpha,n-1} \right\},$$
où
$$T(\mathbf{Y}) = \frac{\sqrt{n}(\overline{Y}_n - \mu_0)}{\left(\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2\right)^{1/2}},$$
où $q^{\mathsf{t}}_{1-\alpha,n-1}$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de libertés.
Si l'on veut tester
$$H_0 : \mu \geq \mu_0 \quad \text{contre} \quad H_1 : \mu < \mu_0,$$
on prend la zone de rejet définie par
$$\mathcal{R}_\alpha = \left\{ T(\mathbf{Y}) < q^{\mathsf{t}}_{1-\alpha,n-1} \right\}.$$

2. On teste
$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu \neq \mu_0.$$
Un test de niveau α est par exemple, le test défini par la zone de rejet
$$\mathcal{R}_\alpha = \left\{ \left|T(\mathbf{Y})\right| > q^{\mathsf{t}}_{1-\alpha/2,n-1} \right\}.$$
Il n'est pas optimal.

3. On teste
$$H_0 : \sigma^2 \leq \sigma \quad \text{contre} \quad H_1 : \sigma^2 > \sigma_0^2.$$
Un test de niveau α est défini par la zone de rejet
$$\mathcal{R}_\alpha = \left\{ V(\mathbf{Y}) > q^{\chi^2}_{1-\alpha,n-1} \right\},$$
où
$$V(\mathbf{Y}) = \frac{1}{\sigma_0^2}\sum_{i=1}^n (Y_i - \overline{Y})^2$$
et $q^{\chi^2}_{1-\alpha,n-1}$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $n-1$ degrés de liberté. Si l'on veut tester
$$H_0 : \sigma \geq \sigma_0 \quad \text{contre} \quad H_1 : \sigma < \sigma_0,$$
on prend la zone de rejet définie par
$$\mathcal{R}_\alpha = \left\{ V(\mathbf{Y}) < q^{\chi^2}_{1-\alpha,n-1} \right\},$$

4. Finalement, si l'on teste

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

on construit un test de niveau α en définissant le test de zone de rejet comme

$$\mathcal{R}_\alpha = \big\{V(\mathbf{Y}) < c_1(\alpha) \ \text{ ou } \ V(\mathbf{Y}) > c_2(\alpha)\big\},$$

où les constantes $c_i(\alpha), i = 1, 2$ sont définies par les conditions

$$\forall \mu \in \mathbb{R}, \ \ \mathbb{P}_{(\mu,\varsigma_0)}\left[\mathcal{R}_\alpha\right] = \alpha$$

et

$$\forall \mu \in \mathbb{R}, \ \ \mathbb{E}_{\mu,\sigma_0}\left[V(\mathbf{Y}) 1_{[c_1(\alpha),c_2(\alpha)]}\big(V(\mathbf{Y})\big)\right] = (n-1)(1-\alpha).$$

Un type de tests couramment rencontrés en pratique sont les tests relatifs à deux échantillons gaussiens. C'est l'objet de l'exercice 7.1.

Sur l'optimalité des tests dans le cas gaussien

Nous avons affirmé l'optimalité de certains des tests présentés dans le paragraphe précédent. Pour la démontrer, on prouve d'abord qu'un test optimal peut être construit par la statistique de test annoncée (la moyenne empirique, la variance empirique, la statistque $\mathfrak{t}$ de Student, et ainsi de suite), puis on optimise les paramètres de sorte de garantir le niveau voulu pour une erreur de seconde espèce minimale, et on retrouve ainsi les tests présentés ci-dessus. Le premier point est délicat et utilise la notion de statistique exhaustive définie au Chapitre 6 et le fait que les modèles gaussiens considérés appartiennent à une famille remarquable de modèles statististique, les modèles exponentiels, dont l'étude dépasse le cadre de ce cours.

7.6.2 Test d'appartenance à un sous-espace linéaire

Situation

On se place dans le cadre du Chapitre 5, sous l'Hypothèse de la Proposition 5.6 et dans la cadre de la régression multiple gaussienne : on observe

$$\mathbf{Y} = \mathbb{M}\,\vartheta + \boldsymbol{\xi}, \ \ \vartheta \in \Theta = \mathbb{R}^d$$

et on suppose

$$\mathbb{M}^T\,\mathbb{M} > 0.$$

On suppose de plus que $\boldsymbol{\xi}$ suit la loi normale sur $\mathbb{R}^n$ de matrice de variance-covariance σ^2 fois l'identité (c'est-à-dire les ξ_i sont indépendantes, de loi $\mathcal{N}(0, \sigma^2)$.)

Un premier cas simple

Soit $a \in \mathbb{R}$. On veut tester $H_0 : \vartheta_j = a$ contre $H_1 : \vartheta \neq a$, pour la composante ϑ_j du vecteur $\vartheta = (\vartheta_1, \ldots, \vartheta_d)^T$, où la direction j est fixée à l'avance.

Un corollaire de la Proposition 5.10 du Chapitre 5 est le résultat suivant

Lemme 7.6.2. *On a, pour tout $\vartheta \in \Theta$, l'égalité en loi sous $\mathbb{P}_\vartheta$*

$$\frac{(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})_j - \vartheta_j}{\sigma\sqrt{(\mathbb{M}^T\mathbb{M})^{-1}_{jj}}} \overset{d}{=} \mathcal{N}(0,1).$$

Démonstration. On a, d'après la Proposition 5,

$$\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta_j \overset{d}{=} \mathcal{N}\big(0, \sigma^2(\mathbb{M}^T\mathbb{M})^{-1}\big)$$

en loi sous $\mathbb{P}_\vartheta$, donc, en posant $v_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ où le terme non-nul est à la j-ième place, la variable aléatoire $(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})_j - \vartheta_j = (\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)^T v_j$ est gaussienne, de moyenne

$$\mathbb{E}_\vartheta\left[(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)^T v_j\right] = 0$$

et de variance

$$\begin{aligned}\mathbb{E}_\vartheta\left[\big((\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)^T v_j\big)^2\right] &= v_j^T\,\mathbb{E}_\vartheta\left[(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \vartheta)^T\right] v_j \\ &= \sigma^2 v_j^T(\mathbb{M}^T\mathbb{M})^{-1} v_j \\ &= \sigma^2(\mathbb{M}^T\mathbb{M})^{-1}_{jj}.\end{aligned}$$

□

Si σ est inconnu, alors, en introduisant l'estimateur s_n^2, le Lemme 7.6.2 devient

Lemme 7.6.3. *On a, pour tout $\vartheta \in \Theta$, l'égalité en loi sous $\mathbb{P}_\vartheta$,*

$$\frac{(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})_j - \vartheta_j}{\widehat{s}_n\sqrt{(\mathbb{M}^T\mathbb{M})^{-1}_{jj}}} \overset{d}{=} \mathfrak{t}(n-d),$$

où $\mathfrak{t}(n-p)$ est la loi de Student de paramètre $n-d$.

Démonstration. Posons $\eta = \sigma(\mathbb{M}^T\mathbb{M})^{-1}_{jj}(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}{}_j - \vartheta_j)$ et

$$\mathfrak{K} = (n-d)\frac{s_n^2}{\sigma^2} = \frac{\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\|^2}{\sigma^2}$$

d'après la Proposition 5.10. Alors sous $\mathbb{P}_\vartheta$, la variable η est gaussienne centrée réduite, et $\mathfrak{K}$ suit la loi du χ^2 à $n-d$ degrés de liberté d'après la Propostion 1.1 (Cochran), et est indépendante de $\mathbf{Y}$ donc de η. □

En conséquence, le test défini par la région critique

$$\mathcal{R}_\alpha = \left\{ \Big| \frac{\big(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\big)_j - a}{\widehat{\sigma}_n (\mathbb{M}^T \mathbb{M})_{jj}^{-1/2}} \Big| > q^{\mathfrak{T}}_{1-\alpha/2,n-d} \right\},$$

où $q^{\mathfrak{T}}_{1-\alpha/2,n-d}$ désigne le quantile d'ordre $1-\alpha$ de la loi de Student à $n-d$ degrés de liberté est de niveau α pour tester $H_0 : \vartheta_j = a$ contre $H_1 : \vartheta_j \neq a$.

Remarque 7.9. Avec ce résultat, on n'a pas d'information sur l'erreur de seconde espèce (la puissance du test), que l'on doit étudier séparément.

Une hypothèse plus générale

Soit $(a_1, \ldots, a_m) \in \mathbb{R}^m$, avec $m < d$ et soit

$$1 \leq j_1 < j_2 < \ldots < j_m \leq d$$

une direction donnée. On souhaite tester

$$H_0 : \vartheta_{j_1} = a_1, \ldots, \vartheta_{j_m} = a_m$$

contre l'alternative

$$H_1 : \quad \text{il existe un indice } \; k \in \{1, \ldots, m\}, \;\; \text{tel que } \; \vartheta_{j_k} \neq a_k.$$

Le cas le plus utile : la sélection de variables

C'est un cas particulier de la situation précédente utile dans de nombreuses situations pratiques. On se place dans le modèle linéaire

$$\mathbf{Y} = \mathbb{M}\ \vartheta + \boldsymbol{\xi},$$

où chaque observation Y_i s'écrit

$$Y_i = \vartheta^T \boldsymbol{x}_i + \xi_i = \sum_{i=1}^{d} \vartheta_i x_i + \xi_i, \;\; i = 1, \ldots, n.$$

(On peut poser $x_1 = 1$ si l'on souhaite incorporer une « ordonnée à l'origine »). Dans le cas de la sélection de variables, on teste si les k premières variables influencent Y, les $d-k$ suivantes ne jouant pas de rôle, ce qui se traduit par l'hypothèse nulle

$$H_0 : \vartheta_{k+\ell} = 0, \;\; \ell = 1, \ldots, \ell = d-k,$$

contre l'alternative

$$H_1 : \text{il existe } \; 1 \leq \ell \leq d-k, \;\; \vartheta_{k+\ell} \neq 0.$$

La sélection de variables est un problème vaste et très important en pratique. On présente quelques compléments sur ce sujet dans l'Exercice 7.2.

Les F-tests

C'est la cadre le plus général, qui inclut les situations décrites précédemment dans cette section.

Soit $\mathbb{G}$ la matrice d'une application linéaire de $\mathbb{R}^d$ dans $\mathbb{R}^m$, avec $m \leq d$, et soit $\mathbf{b} = (a_1, \ldots, a_m)^T$ un vecteur de $\mathbb{R}^m$ arbitraire. On veut tester l'hypothèse nulle

$$H_0 : \mathbb{G}\vartheta = \mathbf{b}$$

contre l'alternative

$$H_1 : \mathbb{G}\vartheta \neq \mathbf{b}.$$

On suppose que $\mathbb{G}$ est de la forme

$$\mathbb{G} = \begin{pmatrix} 0 & \ldots & 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 1 \end{pmatrix},$$

où le premier bloc de 0 a d lignes et $d-m$ colonnes, alors que le second bloc est la matrice identité à m lignes et m colonnes.

Proposition 7.4. *Sous l'hypothèse, c'est-à-dire sous $\mathbb{P}_\vartheta$ avec $\mathbb{G}\vartheta = \mathbf{b}$, on a l'égalité en loi*

$$\mathbb{G}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} \sim \mathcal{N}\big(\mathbf{b}, \sigma^2\mathbb{G}(\mathbb{M}^T\,\mathbb{M})^{-1}\mathbb{G}^T\big).$$

Démonstration. C'est une application de la Proposition 1.1 (Cochran). □

Notons qu'ici, la matrice de variance-covariance est de dimension m. Donc, pour tout point de l'hypothèse ϑ, c'est-à-dire vérifiant $\mathbb{G}\vartheta = \mathbf{b}$, la vecteur m-dimensionnel $\mathbb{G}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ est gaussien, de moyenne $\mathbf{b}$ et de matrice de variance-covariance

$$\mathbf{U} = \sigma^2\mathbb{G}(\mathbb{M}^T\,\mathbb{M})^{-1}\mathbb{G}^T.$$

Notons que puisque $\mathbb{M}^T\,\mathbb{M}$ est inversible, la matrice $\mathbf{U}$ est définie positive. Posons

$$\eta = (\mathbb{G}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \mathbf{b})^T\mathbf{U}^{-1}(\mathbb{G}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \mathbf{b}).$$

Donc sous $\mathbb{P}_\vartheta$ avec $\mathbb{G}\vartheta = \mathbf{b}$, la variable aléatoire η suit la loi du χ^2 à m-degrés de libertés.

On sait alors construire un test de niveau α lorsque σ est connu.

Si σ est inconnu, on peut l'estimer comme précédemment, mais dans le contexte modèle linéaire gaussien général, où ϑ est de dimension $d \geq 1$, voir Proposition 5.10 du Chapitre 5. Alors

$$\widehat{\sigma}_n^2 = \frac{\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\|^2}{n-d},$$

et en posant

$$\widehat{\mathbf{U}} = \widehat{\sigma}_n^2 \mathbb{G}(\mathbb{M}^T \mathbb{M})^{-1} \mathbb{G}^T,$$

la statistique

$$F(\mathbf{Y}) = \frac{(\mathbb{G}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \mathbf{b})^T \widehat{\mathbf{U}}^{-1} (\mathbb{G}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} - \mathbf{b})}{m}$$

est pivotale sous $\mathbb{P}_\vartheta$ avec $\mathbb{G}\vartheta = \mathbf{b}$ et suit la loi de Fisher-Snedecor à $(m, n-d)$ degrés de libertés. Un test de niveau α est alors fourni par la région de rejet

$$\mathcal{R}_\alpha = \left\{ F(\mathbf{Y}) > q^{\mathtt{FS}}_{1-\alpha,m,n-d} \right\},$$

où $q^{\mathtt{FS}}_{1-\alpha,m,n-d}$ désigne le quantile d'ordre $1-\alpha$ de la loi de Fisher-Snedecor à $(m, n-d)$ degrés de liberté. Là encore, ceci ne nous fournit pas d'information sur l'erreur de seconde espèce du test que l'on doit étudier séparément.

7.7 Exercices

Exercice 7.1. Soient $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$ deux échantillons indépendants, de taille respective m et n, de loi respective $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$. On teste

$$H_0 : \mu_1 = \mu_1 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

Construire un test basé sur la statistique

$$T_n = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{(s_m^{(1)})^2 + (s_n^{(2)})^2}},$$

où $(s_m^{(1)})^2 = \frac{1}{m}\sum_{i=1}^m (X_i - \overline{X}_m)^2$ et $(s_n^{(2)})^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$, et étudier sa consistance.

Exercice 7.2 (Règle de Bonferroni en test multiple)**.** On souhaite faire m tests simultanéments. On teste

$$H_{0,i} \quad \text{contre} \quad H_{1,i}, \quad \text{pour } i = 1, \ldots, m$$

Etant donné m tests $\{\varphi_\alpha^{(i)},\ i = 1, \ldots, m\}$ où $\varphi_\alpha^{(i)}$ est un test de niveau α pour l'hypothèse $H_{0,i}$ contre l'alternative $H_{1,i}$, on construit les p-valeurs associées

$$p - \text{valeur}(\varphi_\bullet^{(i)}), \quad i = 1, \ldots, m.$$

La règle de Bonferroni consiste à rejeter l'hypothèse $H_{0,i}$ si $p - \text{valeur}(\varphi_\bullet^{(i)}) < \alpha/m$. Montrer que la probabilité de rejeter à tort une hypothèse nulle parmi les m hypothèses nulle est inférieure à α.

Chapitre 8

Tests asymptotiques

On a vu dans le chapitre précédent que, mis à part des cas relativement particuliers, on n'a pas de méthode de construction de test systématique. Dans ce chapitre, on se place dans le régime asymptoique $n \to \infty$, lorsque l'information de modèle est « grande ». Dans ce cas, dès que le modèle est suffisamment régulier au sens du Chapitre 6 et que l'on dispose d'estimateurs « raisonables », on sait construire des tests de façon un peu plus systématique.

Cependant, on ne pourra pas obtenir l'optimalité d'une suite de tests de niveau (asymptotique) donné aussi facilement qu'au chapitre précédent ; on se contentera d'une notion plus faible : la convergence ou consistance de la suite de tests.

8.1 Convergence d'une suite de tests

On se place dans la problématique du Chapitre 7. Etant donné une suite d'expériences statistiques $\mathcal{E}^n$ ayant pour ensemble de paramètres $\Theta \subset \mathbb{R}^d$ avec $d \geq 1$, on teste

$$H_0 : \vartheta \in \Theta_0 \quad \text{contre} \quad H_1 : \vartheta \in \Theta_1, \quad \text{avec} \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

On se donne un test ou plutôt une suite de tests [1] simples φ_n dans $\mathcal{E}^n$ de l'hypothèse nulle H_0 contre l'alternative H_1.

Définition 8.1 (Niveau asymptotique d'une suite de tests)**.** *Soit $\alpha \in [0,1]$. Le test φ_n est asymptotiquement de niveau α si son erreur de première espèce est asymptotiquement plus petite que α :*

$$\forall \vartheta \in \Theta_0, \quad \limsup_{n\to\infty} \mathbb{P}_\vartheta \left[\varphi_n = 0\right] \leq \alpha.$$

[1] de la même manière que l'on parle d'estimateur pour une suite d'estimateurs, on utilisera le terme test pour désigner une suite de tests.

Définition 8.2. *Le test φ_n est convergent ou consistant si sa puissance asymptotique vaut 1, c'est-à-dire si son erreur de seconde espèce est asymptotiquement nulle :*

$$\forall \vartheta \in \Theta_1, \quad \lim_{n\to\infty} \mathbb{P}_\vartheta\big[\varphi_n = 1\big] = 1 = 1 - \lim_{n\to\infty} \mathbb{P}_\vartheta\big[\varphi_n = 1\big].$$

8.2 Tests de Wald

8.2.1 Le cas d'une hypothèse nulle simple

Traitons d'abord le cas du test d'une hypothèse nulle simple $H_0 : \vartheta = \{\vartheta_0\}$ contre $H_1 : \vartheta \neq \vartheta_0$. Plaçons-nous en dimension $d = 1$ pour simplifier. Supposons que l'on dispose d'un estimateur $\widehat{\vartheta}_n$ asymptotiquement normal, c'est-à-dire pour lequel on a, pour tout $\vartheta \in \Theta$,

$$\sqrt{n}\big(\,\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v(\vartheta)\big),$$

où $v(\vartheta) > 0$, la convergence ayant lieu en loi sous $\mathbb{P}_\vartheta$. On suppose que la fonction $\vartheta \rightsquigarrow v(\vartheta)$ est régulière. Sous l'hypothèse, c'est-à-dire sous $\mathbb{P}_{\vartheta_0}$, on a la convergence

$$\sqrt{n}\frac{\widehat{\vartheta}_n - \vartheta_0}{\sqrt{v(\vartheta_0)}} \xrightarrow{d} \mathcal{N}(0,1),$$

en loi sous $\mathbb{P}_{\vartheta_0}$, ou encore, en appliquant la Proposition 1.8 (Slutsky)

$$T_n = \sqrt{n}\frac{\widehat{\vartheta}_n - \vartheta_0}{\sqrt{v(\widehat{\vartheta}_n)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{8.1}$$

en loi sous $\mathbb{P}_{\vartheta_0}$. On en déduit « presque immédiatement » la construction suivante

Proposition 8.1. *Pour tout $\alpha \in (0,1)$, le test φ_n défini par la zone de rejet*

$$\mathcal{R}_{n,\alpha} = \big\{\big|T_n\big| \geq \Phi^{-1}(1-\alpha/2)\big\},$$

où $\Phi^{-1}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ de la loi normale standard, est asymptotiquement de niveau α et consistant.

Démonstration. Le contrôle du niveau asymptotique de φ_n est une conséquence immédiate de la convergence (8.1) :

$$\mathbb{P}_{\vartheta_0}\big[\varphi_n = 1\big] = \mathbb{P}_{\vartheta_0}\big[\big|T_n\big| \geq \Phi^{-1}(1-\alpha/2)\big] \to \alpha.$$

Montrons la consistance. Soit $\vartheta \neq \vartheta_0$ un point de l'alternative. On écrit

$$T_n = \sqrt{n}\frac{\widehat{\vartheta}_n - \vartheta}{\sqrt{v(\widehat{\vartheta}_n)}} + \sqrt{n}\frac{\vartheta - \vartheta_0}{\sqrt{v(\widehat{\vartheta}_n)}}. \tag{8.2}$$

Le premier terme tend en loi sous $\mathbb{P}_\vartheta$ vers la loi $\mathcal{N}(0,1)$, en appliquant la convergence (8.1) avec ϑ à la place de ϑ_0. Le dénominateur du second terme converge en probabilité sous $\mathbb{P}_\vartheta$ vers $v(\vartheta)$, et le numérateur diverge vers $\pm\infty$. Donc

$$|T_n| \xrightarrow{\mathbb{P}_\vartheta} +\infty$$

et donc $\varphi_n \xrightarrow{\mathbb{P}_\vartheta} 1$ pour tout $\vartheta \neq \vartheta_0$. On en déduit la consistance de φ_n (par exemple par convergence dominée). □

Remarque 8.1. Ici, le choix de la zone de rejet ne s'impose pas naturellement. Si $\mathcal{D}_\alpha \subset \mathbb{R}$ est tel que

$$\mathbb{P}\left[\xi \in \mathcal{D}_\alpha\right] = 1 - \alpha \tag{8.3}$$

où $\xi \sim \mathcal{N}(0,1)$, alors le test $\varphi_n(\mathcal{D}_\alpha)$ défini par la zone de rejet

$$\mathcal{R}_n(\mathcal{D}_\alpha) = \left\{T_n \notin \mathcal{D}_\alpha\right\}$$

est asymptotiquement de niveau α.

Remarque 8.2. Pour construire le test φ_n de la Proposition 8.1, on a choisi la zone d'acceptation

$$\mathcal{D}_\alpha = \left[-\Phi^{-1}(1-\alpha/2), \Phi^{-1}(1-\alpha/2)\right]$$

car elle est de longueur minimale parmi les zones $\mathcal{D}_\alpha$ satisfaisant (8.3) mais ce choix n'a pas d'importance si l'on n'étudie pas plus précisément la puissance du test. Si l'on se contente simplement de la consistance, il suffit d'imposer en plus que $\mathcal{D}_\alpha$ est borné. Dans ce cas, on a toujours $\varphi_n(\mathcal{D}_\alpha) \xrightarrow{\mathbb{P}_\vartheta} 1$ pour tout point $\vartheta \neq \vartheta_0$ l'alternative et $\varphi_n(\mathcal{D}_\alpha)$ est consistant.

Remarque 8.3. Le test φ_n basé sur la statistique T_n dépend de $v(\vartheta)$. Intuitivement, il sera d'autant meilleur (d'autant plus puissant) que $v(\vartheta)$ sera petit. Cela se voit immédiatement sur la décomposition (8.2) : le terme de droite diverge « d'autant mieux » que $v(\widehat{\vartheta}_n)$ et donc asymptotiquement $v(\vartheta)$ est petit, sans que cela affecte son erreur de première espèce.
Si on est dans un modèle d'échantillonnage régulier, on aura donc intérêt à prendre l'estimateur de variance asymptotique minimale, c'est-à-dire l'estimateur du maximum de vraisemblance, qui fournit $v(\vartheta) = \mathbb{I}(\vartheta)$.

Dans la convergence (8.1), on aurait pu, de manière équivalente, remplacer la statistique T_n par son carré, et obtenir

$$T_n^2 = n\frac{(\widehat{\vartheta}_n - \vartheta_0)^2}{v(\widehat{\vartheta}_n)} \xrightarrow{d} \chi^2(1)$$

en loi sous $\mathbb{P}_\vartheta$, où $\chi^2(1)$ désigne la loi du χ^2 à 1 degré de liberté. En construisant un test basé sur la statistique T_n avec comme loi limite, on obtient la zone de rejet

$$\widetilde{\mathcal{R}}_{n,\alpha} = \left\{ T_n^2 \geq q^{\chi^2}_{1-\alpha,1} \right\}$$

où $q^{\chi^2}_{1-\alpha,1}$ désigne le quantile d'ordre $1-\alpha$ de la loi du χ^2 à 1 degré de liberté. Sans surprise, $\widetilde{\mathcal{R}}_{n,\alpha} = \mathcal{R}_{n,\alpha}$!

8.2.2 Hypothèse nulle composite

On se place dans le cadre général $\Theta \subset \mathbb{R}^d$, et on suppose que Θ_0 peut s'écrire sous la forme

$$\Theta_0 = \big\{ \vartheta \in \Theta, \;\; g(\vartheta) = 0 \big\}$$

où l'application

$$g : \mathbb{R}^d \to \mathbb{R}^m$$

est régulière. Par exemple, l'hypothèse nulle simple $H_0 : \vartheta = \vartheta_0$ pour un point $\vartheta_0 \in \Theta$ donné peut toujours se ramener à la condition $g(\vartheta) = 0$, avec $g(\vartheta) = \vartheta - \vartheta_0$.

Remarque 8.4. En dimension $d = 1$, l'hypothèse composite $H_0 : \vartheta > \vartheta_0$ s'écrit bien sous la forme $g(\vartheta) = 0$ avec $g(\vartheta) = 1_{\vartheta \leq \vartheta_0}$, mais la fonction $\vartheta \rightsquigarrow g(\vartheta)$ n'est pas continue en ϑ_0.

Construction du test de Wald

Hypothèse 8.1. *L'application $g : \mathbb{R}^d \to \mathbb{R}^m$ est continûment différentiable. De plus, sa différentielle, en tant qu'élément de $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$, est de rang maximal m en tout point ϑ de (l'intérieur[2] de) Θ_0.*

On notera $J_g(\vartheta)$ la matrice de la différentielle de g au point ϑ. On suppose qu'il existe un estimateur $\widehat{\vartheta}_n$ de ϑ asymptotiquement normal, au sens suivant

Hypothèse 8.2.

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, V(\vartheta)\big),$$

en loi sous $\mathbb{P}_\vartheta$, où $V(\vartheta)$ est définie positive, et $\vartheta \rightsquigarrow V(\vartheta)$ est continue pour tout $\vartheta \in \Theta$.

Proposition 8.2. *Sous l'Hypothèse 8.1, en tout point $\vartheta \in \Theta_0$ de l'hypothèse, c'est-à-dire vérifiant $g(\vartheta) = 0$, on a*

$$\sqrt{n} g(\widehat{\vartheta}_n) \xrightarrow{d} \mathcal{N}\big(0, J_g(\vartheta) V(\vartheta) J_g(\vartheta)^T\big)$$

sous $\mathbb{P}_\vartheta$ lorsque $n \to \infty$.

[2] en ne tenant pas compte de cette restriction quand Θ_0 se réduit à un seul point.

Corollaire 8.1. *Posons $\Sigma_g(\vartheta) = J_g(\vartheta)V(\vartheta)J_g(\vartheta)^T$ dans la proposition précédente. On a la convergence*

$$T_n^2(g) = ng(\widehat{\vartheta}_n)^T\Sigma_g(\widehat{\vartheta}_n)^{-1}g(\widehat{\vartheta}_n) \xrightarrow{d} \chi^2(m) \tag{8.4}$$

sous $\mathbb{P}_\vartheta$, où $\chi^2(m)$ désigne la loi du χ^2 à m degrés de liberté. Pour tout $\alpha \in (0,1)$, le test défini par la région critique

$$\mathcal{R}_{n,\alpha} = \left\{T_n^2 \geq q^{\chi^2}_{1-\alpha,m}\right\} \tag{8.5}$$

où $q^{\chi^2}_{1-\alpha,m}$ désigne le quantile d'ordre $1-\alpha$ de la loi du χ^2 à m degrés de liberté est asymptotiquement de niveau α et consistant.

Définition 8.3 (Test de Wald)**.** *On appelle test de Wald de $H_0 : g(\vartheta) = 0$ contre $H_1 : g(\vartheta) \neq 0$ associé à l'estimateur asymptotiquement normal $\widehat{\vartheta}_n$ le test basé sur la statistique T_n^2 définie en* (8.4) *de région critique $\mathcal{R}_{n,\alpha}$ défini en* (8.5)*. La statistique T_n^2 s'appelle statistique de Wald (associée à l'estimateur $\widehat{\vartheta}_n$).*

Remarque 8.5. Le test de la Proposition 8.1 est un test de Wald, dans la cas très particulier où $g(\vartheta) = \vartheta - \vartheta_0$ en dimension 1. En particulier, $g'(\vartheta) = 1$ en tout point $\vartheta \in \Theta \subset \mathbb{R}$.

Démonstration de la Proposition 8.2 et de son Corollaire 8.1. La proposition est simplement la version multidimensionnelle de la « méthode delta », Proposition 1.11 appliquée à $g(\widehat{\vartheta}_n)$ d'après l'Hypothèse 8.2, en utilisant le fait que sous l'hypothèse nulle $g(\vartheta) = 0$. Pour son corollaire, on en déduit d'abord la convergence

$$\sqrt{n}\Sigma_g(\vartheta)^{-1}g(\widehat{\vartheta}_n) \xrightarrow{d} \mathcal{N}(0, \mathrm{Id}_m),$$

en loi sous $\mathbb{P}_\vartheta$, puis, par la Proposition 1.8 (Slutsky), par continuité de $\vartheta \rightsquigarrow \Sigma_g(\vartheta)$

$$\sqrt{n}\Sigma_g(\widehat{\vartheta}_n)^{-1}g(\widehat{\vartheta}_n) \xrightarrow{d} \mathcal{N}(0, \mathrm{Id}_m).$$

En passant à la norme au carré

$$\|\sqrt{n}\Sigma_g(\widehat{\vartheta}_n)^{-1}g(\widehat{\vartheta}_n)\|^2 = ng(\widehat{\vartheta}_n)^T\Sigma_g(\widehat{\vartheta}_n)^{-1}g(\widehat{\vartheta}_n) \xrightarrow{d} \|\mathcal{N}(0, \mathrm{Id}_m)\|^2 \sim \chi^2(m).$$

On en déduit que le test donné par la région de rejet $\mathcal{R}_{n,\alpha}$ est asymptotiquement de niveau α.

Montrons qu'il est consistant. On raisonne comme en dimension 1 : si $\vartheta \in \Theta_1$ est un point de l'alternative, on a $g(\vartheta) \neq 0$, on force le terme $g(\vartheta)$ dans T_n et on écrit

$$T_n^2 = T_{n,1}^2 + T_{n,2}^2,$$

avec

$$T_{n,1}^2 = n\big(g(\widehat{\vartheta}_n) - g(\vartheta)\big)^T\Sigma_g(\widehat{\vartheta}_n)^{-1}\big(g(\widehat{\vartheta}_n) - g(\vartheta)\big),$$

et un terme additionnel

$$T_{n,2}^2 = U_n + V_n,$$

qui se redécompose en

$$U_n = ng(\vartheta)^T \Sigma_g(\widehat{\vartheta}_n)^{-1} g(\vartheta)$$

et

$$V_n = n\big(g(\widehat{\vartheta}_n) - g(\vartheta)\big)^T \Sigma_g(\widehat{\vartheta}_n)^{-1} g(\vartheta) + ng(\vartheta)^T \Sigma_g(\widehat{\vartheta}_n)^{-1} \big(g(\widehat{\vartheta}_n) - g(\vartheta)\big).$$

Pour tout ϑ, le terme $T_{n,1}$ converge en loi sous $\mathbb{P}_\vartheta$ vers la loi du χ^2 à m degrés de liberté : c'est la « méthode delta » appliquée à $g(\widehat{\vartheta}_n)$ lorsque $g(\vartheta) \neq 0$. Il reste à démontrer que $T_{n,2}$ diverge. Par continuité, $V_g(\widehat{\vartheta}_n) \overset{\mathbb{P}_\vartheta}{\to} V_g(\vartheta)$, donc $U_n \overset{\mathbb{P}_\vartheta}{\to} +\infty$. Le terme V_n diverge de même, mais on ne peut pas contrôler son signe. Il reste à vérifier que V_n est petit devant U_n. Pour cela, on écrit $V_n = \sqrt{n}\tilde{V}_n$, avec

$$\tilde{V}_n = \sqrt{n}\big(g(\widehat{\vartheta}_n) - g(\vartheta)\big)^T \Sigma_g(\widehat{\vartheta}_n)^{-1} g(\vartheta) + \sqrt{n} g(\vartheta)^T \Sigma_g(\widehat{\vartheta}_n)^{-1} \big(g(\widehat{\vartheta}_n) - g(\vartheta)\big)$$

et chacun des termes converge séparement en loi sous $\mathbb{P}_\vartheta$ via la Proposition 8.2. Donc $V_n/U_n \overset{\mathbb{P}_\vartheta}{\to} 0$ et le corollaire est démontré. □

8.3 Test « sup sur sup »*

Situation et notations

On suppose pour simplifier que $\mathcal{E}^n$ est engendrée par un n-échantillon

$$X_1, \ldots, X_n$$

de variables aléatoires réelles, dont la loi appartient à la famille $\big\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\big\}$, avec $\Theta \subset \mathbb{R}^d$, $d \geq 1$, dominée par un mesure σ-finie μ sur $\mathbb{R}$. On note $\{f(\vartheta, \bullet),\ \vartheta \in \Theta\}$ la famille de densités associées. On teste $H_0 : \vartheta \in \Theta_0$ contre $H_1 : \vartheta \in \Theta_1$, avec $\Theta_0 \cap \Theta_1 = \emptyset$.

La statistique « sup sur sup »

Si les des deux hypothèses sont simples, c'est-à-dire $\Theta_0 = \{\vartheta_0\}$ et $\Theta_1 = \{\vartheta_1\}$, avec $\vartheta_0 \neq \vartheta_1$, alors l'approche de Neyman-Pearson de la Section 7.2.2 du chapitre précédent suggère de considérer le rapport des vraisemblances

$$\frac{\mathcal{L}_n(\vartheta_1, X_1, \ldots, X_n)}{\mathcal{L}_n(\vartheta_0, X_1, \ldots, X_n)} = \frac{\prod_{i=1}^n f(\vartheta_1, X_i)}{\prod_{i=1}^n f(\vartheta_0, X_i)},$$

ou son logarithme

$$\sum_{i=1}^n \log f(\vartheta_1, X_i) - \sum_{i=1}^n \log f(\vartheta_0, X_i),$$

et, suivant la règle de la construction du test du rapport de vraisemblance, on rejette l'hypothèse nulle $\vartheta = \vartheta_0$ si Λ_n dépasse un seuil, calibré pour contrôler l'erreur de première espèce.

Lorsque Θ_0 et Θ_1 ne sont pas réduits à un point, une règle conservative consiste à remplacer la quantité ci-dessus par

$$\widetilde{\Lambda}_n(X_1, \dots, X_n) = \sup_{\vartheta \in \Theta_1} \sum_{i=1}^{n} \log f(\vartheta, X_i) - \sup_{\vartheta \in \Theta_0} \sum_{i=1}^{n} \log f(\vartheta, X_i)$$

et donc de comparer la vraisemblance de « la valeur la plus vraisemblable » sur Θ_0 à « la valeur la plus vraisemblable » sur Θ_0. Malheureusement, le calcul de la loi de cette quantité est difficile, même asymptotiquement. On remplace alors $\widetilde{\Lambda}_n$ par

$$\begin{aligned}\Lambda_n &= \sup_{\vartheta \in \Theta} \sum_{i=1}^{n} \log f(\vartheta, X_i) - \sup_{\vartheta \in \Theta_0} \sum_{i=1}^{n} \log f(\vartheta_0, X_i) \\ &= \log \frac{\sup_{\vartheta \in \Theta} \mathcal{L}(\vartheta, X_1, \dots, X_n)}{\sup_{\vartheta \in \Theta_0} \mathcal{L}(\vartheta, X_1, \dots, X_n)},\end{aligned}$$

où le supremum au dénominateur est évalué sur tout l'espace des paramètres. On peut se convaincre – au moins heuristiquement – que cette approche est raisonnable si le modèle est suffisamment régulier. Dans ce cas, si $\vartheta \in \Theta_1$, sous $\mathbb{P}_\vartheta$, la quantité qui atteint le maximum pour le dénominateur est l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$ qui convergera vers $\vartheta \in \Theta_1$.

Définition 8.4. *On appelle Λ_n la « statistique du rapport de vraisemblance maximal ».*

Un résultat remarquable est que sous l'hypothèse nulle, la loi de la statistique du rapport de vraisemblance maximal est asymptotiquement la loi du χ^2 (à une constante multiplicative près) pour un nombre de degrés de liberté dépendant de la dimension de Θ_0, et ceci conduit à une méthode systématique de construction de tests.

8.3.1 Rapport de vraisemblance maximal asymptotique

On suppose le modèle régulier au sens du Chapitre 6. Notons $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$ l'estimateur du maximum de vraisemblance du Θ et $\widehat{\vartheta}_{\mathrm{n},0}^{\mathtt{mv}}$ l'estimateur du maximum de vraisemblance restreint à Θ_0 (c'est-à-dire obtenu lorsque l'on maximise la vraisemblance sur Θ_0).

En appliquant la formule de Taylor à l'ordre 2 à $\vartheta \rightsquigarrow \ell(\vartheta, x) = \log f(\vartheta, \bullet)$, on réécrit

Λ_n comme

$$\begin{aligned}
&-\sum_{i=1}^{n}\big(\ell(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0},X_i)-\ell(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}},X_i)\big)\\
&=-\Big(\sum_{i=1}^{n}\nabla\ell(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}},X_i)\Big)^T(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}})-\tfrac{1}{2}(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0})^T\Big(\sum_{i=1}^{n}H_{\ell(\bullet,X_i)}[\widetilde{\vartheta}_n]\Big)(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0})\\
&=-\tfrac{1}{2}(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0})^T\Big(\sum_{i=1}^{n}H_{\ell(\bullet,X_i)}[\widetilde{\vartheta}_n]\Big)(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}),
\end{aligned}$$

où $\widetilde{\vartheta}_n$ est un point entre $\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}$ et $\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}$ et $H_{\ell(\bullet,X_i)}(\vartheta)$ désigne la matrice hessienne de la fonction $\vartheta\rightsquigarrow\ell(\vartheta,X_i)$ au point ϑ. Le terme d'ordre 1 disparaît par définition du maximum de vraisemblance (dès que $\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}\in\Theta$). Sous les hypothèses de régularité sur le modèle $\big\{\mathbb{P}_\vartheta,\,\vartheta\in\Theta\big\}$, si $\vartheta\in\Theta_0$, on a les convergences

$$\sqrt{n}\big(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}-\vartheta\big)\xrightarrow{d}\mathcal{N}\big(0,\mathbb{I}^{-1}(\vartheta)\big)\ \text{ en loi sous }\ \mathbb{P}_\vartheta,\ \vartheta\in\Theta_0,\tag{8.6}$$

où $\mathbb{I}^{-1}(\vartheta)$ désigne l'inverse de la matrice d'information de Fisher du modèle $\big\{\mathbb{P}_\vartheta,\,\vartheta\in\Theta\big\}$, et on a toujours

$$\sqrt{n}\big(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\vartheta\big)\xrightarrow{d}\mathcal{N}\big(0,\mathbb{I}^{-1}(\vartheta)\big)\ \text{ en loi sous }\ \mathbb{P}_\vartheta\,.\tag{8.7}$$

Donc la suite de vecteurs $\sqrt{n}(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0})$ est bornée en probabilité sous $\mathbb{P}_\vartheta,\vartheta\in\Theta_0$. Par ailleurs, on a toujours la convergence

$$-\frac{1}{n}\sum_{i=1}^{n}H_{\ell(\bullet,X_i)}[\vartheta]\overset{\mathbb{P}_\vartheta}{\to}\mathbb{I}(\vartheta),\quad\vartheta\in\Theta_0\tag{8.8}$$

(composante par composante) par la loi des grands nombres. On en déduit la

Proposition 8.3. *Si l'expérience statistique est régulière au sens du Chapitre 6, pour tout $\vartheta\in\Theta_0$ (c'est-à-dire en se plaçant sous l'hypothèse H_0), on a les approximations suivantes*

$$\Lambda_n=\tfrac{1}{2}\sqrt{n}\big(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}-\vartheta\big)^T\mathbb{I}(\vartheta)\sqrt{n}\big(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}-\vartheta\big)^T+\varepsilon_n$$

et aussi

$$\Lambda_n=\tfrac{1}{2}\sqrt{n}\big(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}-\vartheta\big)^T\mathbb{I}(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}})\sqrt{n}\big(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}-\vartheta\big)^T+\varepsilon'_n$$

où ε_n et ε'_n sont deux suites qui tendent vers 0 en probabilité sous $\mathbb{P}_\vartheta$ pour tout $\vartheta\in\Theta_0$.

Démonstration. La première approximation est simplement une combinaison des estimations précédentes : on écrit

$$\begin{aligned}
&-(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0})^T\Big(\sum_{i=1}^{n}H_{\ell(\bullet,X_i)}[\widetilde{\vartheta}_n]\Big)(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0})\\
&=-\sqrt{n}(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0})^T\Big(\frac{1}{n}\sum_{i=1}^{n}H_{\ell(\bullet,X_i)}[\widetilde{\vartheta}_n]\Big)\sqrt{n}(\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n}}-\widehat{\vartheta}^{\mathtt{mv}}_{\mathrm{n},0}),
\end{aligned}$$

et on utilise d'une part le fait que le terme du milieu converge en probabilité vers $\mathbb{I}^{-1}(\vartheta)$ via (8.8) en utilisant le fait que $\widetilde{\vartheta}_n$ est proche de ϑ (nous omettons les détails), et d'autre part que la suite $\sqrt{n}(\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} - \widehat{\vartheta}_{\mathrm{n},0}^{\mathtt{mv}})$ est bornée en $\mathbb{P}_\vartheta$ probabilité pour $\vartheta \in \Theta_0$ par (8.6) et (8.7).

La seconde approximation est simplement une conséquence de la Proposition 1.8 (Slutsky). □

Remarque 8.6. Les estimateurs $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$ et $\widehat{\vartheta}_{\mathrm{n},0}^{\mathtt{mv}}$ ne sont pas les mêmes en général. Un exemple classique – rencontré aussi en régression – est celui de l'expérience statistique engendrée par un n-échantilllon de loi $\mathcal{N}(\mu, \sigma^2)$, avec $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$. Alors, si $\Theta_0 = \{\vartheta \in \Theta,\ \mu = 0\}$, on a

$$\widehat{\vartheta}_{\mathrm{n},0}^{\mathtt{mv}} = \big(0, \tfrac{1}{n}\sum_{i=1}^n X_i^2\big), \quad \text{alors que} \quad \widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} = \big(\overline{X}_n, \tfrac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2\big).$$

8.3.2 Lien avec la statistique de Wald

Plaçons-nous dans le cas d'une hypothèse nulle simple $\Theta_0 = \{\vartheta_0\}$ pour simplifier. La statistique T_n^2 du test de Wald définie dans le Corollaire 8.1 par (8.4) s'écrit à l'aide de la fonction $g(\vartheta) = \vartheta - \vartheta_0$, et $J_g = \mathrm{Id}_d$.

Si l'expérience sous-jacente est régulière, le choix de l'estimateur $\widehat{\vartheta}_n = \widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$ conduit à $V(\vartheta) = \mathbb{I}(\vartheta)$, où $\mathbb{I}(\vartheta)$ est l'information de Fisher du modèle. On a donc dans ce cas $\Sigma_g(\vartheta) = J_g(\vartheta)V(\vartheta)J_g(\vartheta)^T = \mathbb{I}(\vartheta)$ et finalement,

$$T_n^2 = \sqrt{n}\big(\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} - \vartheta_0\big)^T \mathbb{I}(\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}})\sqrt{n}\big(\widehat{\vartheta}_{\mathrm{n},0}^{\mathtt{mv}} - \vartheta\big)^T.$$

Par ailleurs, puisque l'hypothèse nulle H_0 est simple, on a $\widehat{\vartheta}_{\mathrm{n},0}^{\mathtt{mv}} = \vartheta_0$. D'après la Proposition 8.3, on déduit

$$T_n^2 = 2\Lambda_n + \varepsilon_n, \tag{8.9}$$

où ε_n tend vers 0 en probabilité sous $\mathbb{P}_{\vartheta_0}$.

En conclusion, dans le cas d'une hypothèse nulle simple, la statistique de Wald associée à l'estimateur du maximum de vraisemblance et la statistique du rapport de vraisemblance maximal sont asymptotiquement équivalentes. On en déduit immédiatement que – pour une hypothèse nulle simple – la statistique du rapport de vraisemblance maximale converge en loi vers la loi du χ^2 à d degrés de liberté

Remarque 8.7. Le lien que nous veons de montrer est très simplificateur. L'équivalence (8.9) s'étend au-delà d'une hypothèse simple. Nous nous contenterons de ce résultat particulier dans ce cours.

Remarque 8.8. Une autre statistique remarquable, la statistique du score (voir par exemple Wasserman, [8]), se déduit de ces approximations.

8.3.3 Résultat général pour le rapport de vraisemblance maximal*

Dans le cas d'une hypothèse nulle simple $\Theta_0 = \{\vartheta_0\}$, nous venons de voir – par l'équivalence asymptotique avec la statistique de Wald associée à l'estimateur du maximum de vraisemblance – que la statistique $2\Lambda_n$ suit asymptotiquement la loi du χ^2 à d degrés de liberté. Ici, grâce à la Propostion 8.1, le degré d doit être compris comme le rang de la diférentielle de $J_g(\vartheta)$, qui dans le cas trivial $g(\vartheta) = \vartheta - \vartheta_0$ est maximal.

Ce résultat se généralise. On suppose que Θ_0 peut s'écrire sous la forme

$$\Theta_0 = \big\{\vartheta \in \Theta, \;\; g(\vartheta) = 0\big\}$$

où l'application

$$g : \mathbb{R}^d \to \mathbb{R}^m$$

est régulière au sens de l'Hypothèse 8.1, c'est-à-dire continûment différentiable, sa différentielle étant de rang maximal m en tout point de (l'intérieur de) Θ_0.

Proposition 8.4. *Si l'expérience statistique est régulière au sens du Chapitre 6. Sous l'Hypothèse 8.1, pour tout point ϑ (dans l'intérieur) de Θ_0 (ou si Θ_0 est réduit à un point), c'est-à-dire tel que $g(\vartheta) = 0$, on a*

$$2\Lambda_n \xrightarrow{d} \chi^2(m).$$

Nous admettons ce résultat. On en déduit un test de asymptotiquement de niveau α défini par la région critique

$$\mathcal{R}_{n,\alpha} = \big\{2\Lambda_n \geq q^{\chi^2}_{1-\alpha,m}\big\},$$

où $q^{\chi^2}_{1-\alpha,m}$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à m degrés de liberté.

8.4 Tests du χ^2

Notation et préliminaire

Si X une variable qualitative pouvant prendre d valeurs distinctes, on note $\{1, \ldots, d\}$ l'ensemble de ses valeurs pour simplifier. En toute généralité, la loi de X s'écrit

$$\mathbb{P}\big[X = \ell\big] = p_\ell, \quad \ell = 1, \ldots, d$$

avec $0 \leq p_\ell \leq 1$ et $\sum_{\ell=1}^d p_\ell = 1$, et le vecteur $\boldsymbol{p} = (p_1, \ldots, p_d)^T$ caractérise la loi de X. Désormais, nous identifions les lois de probablités prenant d valeurs avec les vecteurs $\boldsymbol{p}$ de l'ensemble

$$\mathcal{M}_d = \Big\{ \boldsymbol{p} = (p_1, \ldots, p_\ell)^T, \quad 0 \leq p_\ell \leq 1, \quad \sum_{\ell=1}^d p_\ell = 1\Big\}$$

8.4.1 Test d'adéquation du χ^2

On observe un n-échantillon

$$X_1, \dots, X_n$$

de loi $\boldsymbol{p} \in \mathcal{M}_d$ inconnue et on teste l'hypothèse

$$H_0 : \boldsymbol{p} = \boldsymbol{q}, \quad \text{contre} \quad H_1 : \boldsymbol{p} \neq \boldsymbol{q}$$

où $\boldsymbol{q} \in \mathcal{M}_d$ est une loi donnée. L'expérience statistique associée à l'observation s'écrit

$$\mathcal{E}^n = \Big(\{1, \dots, d\}^n, \mathcal{P}(\{1, \dots, d\}^n),\ \big\{ \mathbb{P}_{\boldsymbol{p}}^n,\ \boldsymbol{p} \in \mathcal{M}_d \big\} \Big),$$

où $\mathbb{P}_{\boldsymbol{p}}^n$ est la loi[3] d'un n-échantillon de loi $\boldsymbol{p}$.

Pour construire un test, une idée immédiate est de comparer les fréquences empiriques

$$\widehat{p}_{n,\ell} = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i = \ell}, \quad \ell = 1, \dots, d \tag{8.10}$$

avec q_ℓ, $\ell = 1, \dots, d$. En effet, la loi des grands nombre garantit la convergence

$$(\widehat{p}_{n,1}, \dots, \widehat{p}_{n,d}) \xrightarrow{\mathbb{P}_{\boldsymbol{p}}} (p_1, \dots, p_d) = \boldsymbol{p} \tag{8.11}$$

en probabilité sous $\mathbb{P}_{\boldsymbol{p}}$. L'étape suivante consiste à établir une vitesse de convergence dans (8.11). En anticipant sur le théorème central-limite, on considère le vecteur

$$\boldsymbol{U}_n(\boldsymbol{p}) = \sqrt{n} \left(\frac{\widehat{p}_{n,1} - p_1}{\sqrt{p_1}}, \dots, \frac{\widehat{p}_{n,d} - p_d}{\sqrt{p_d}} \right)^T$$

qui est bien défini si toutes les composantes de $\boldsymbol{p}$ sont non nulles, ainsi que sa norme au carré

$$\|\boldsymbol{U}_n(\boldsymbol{p})\|^2 = n \sum_{\ell=1}^{d} \frac{\big(\widehat{p}_{n,\ell} - p_\ell\big)^2}{p_\ell}.$$

Par le théorème central limite, chaque composante de $\boldsymbol{U}_n$ converge en loi vers une gaussienne centrée réduite, mais ceci ne permet pas d'en déduire la convergence en loi vectorielle (et donc pas non plus celle de $\|\boldsymbol{U}_n\|^2$, utile pour construire un test), puisque les variables aléatoires $\widehat{p}_{\ell,n}$ ne sont pas indépendantes. Le résultat suivant précise la convergence

Proposition 8.5. *Si les composantes de $\boldsymbol{p}$ sont toutes non nulles, alors*

$$\boldsymbol{U}_n(\boldsymbol{p}) \xrightarrow{d} \mathcal{N}\big(0, V(\boldsymbol{p})\big), \tag{8.12}$$

[3]Dans cette section, $\boldsymbol{p} \in \mathcal{M}_d$ remplacera l'écriture habituelle $\vartheta \in \Theta$.

où $V(\boldsymbol{p}) = \mathrm{Id}_d - \sqrt{\boldsymbol{p}}(\sqrt{\boldsymbol{p}})^T$, *et* $\sqrt{\boldsymbol{p}} = (\sqrt{p_1}, \ldots, \sqrt{p_d})^T$. *De plus*

$$\|\boldsymbol{U}_n(\boldsymbol{p})\|^2 \xrightarrow{d} \chi^2(d-1), \tag{8.13}$$

où $\chi^2(d-1)$ *désigne la loi du* χ^2 *à* $d-1$ *degrés de liberté.*

Démonstration. Pour $i = 1, \ldots, n$ et $1 \leq \ell \leq d$, posons

$$Y_\ell^i = \frac{1}{\sqrt{p_\ell}}(1_{\{X_i=\ell\}} - p_\ell).$$

La suite de vecteurs $\boldsymbol{Y}_i = (Y_1^i, \ldots, Y_d^i)$ est indépendante et de même loi, car chaque terme $\boldsymbol{Y}_i$ ne fait intervenir que la variables X_i est les X_i sont indépendantes et de même loi. Notons que

$$\boldsymbol{U}_n(\boldsymbol{p}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Y}_i \,.$$

De plus,

$$\mathbb{E}\left[Y_\ell^i\right] = 0, \;\; \mathbb{E}\left[(Y_\ell^i)^2\right] = p_\ell^{-1}(p_\ell - 2p_\ell^2 + p_\ell^2) = 1 - p_\ell,$$

et pour $\ell \neq \ell'$,

$$\mathbb{E}\left[Y_\ell^i Y_{\ell'}^i\right] = (p_\ell p_{\ell'})^{-1/2}(0 - 2p_\ell p_{\ell'} + p_\ell p_{\ell'}) = -(p_\ell p_{\ell'})^{1/2}.$$

On applique alors le théorème central limite vectoriel 1.4 du Chapitre 1. On obtient la convergence (8.12).

Pour la convergence (8.13), par continuité du carré de la norme, on a

$$\|\boldsymbol{U}_n(\boldsymbol{p})\|^2 \xrightarrow{d} \left\|\mathcal{N}\big(0, V(\boldsymbol{p})\big)\right\|^2 \;\sim\; \chi^2\Big(\mathrm{Rang}\big(V(\boldsymbol{p})\big)\Big),$$

la dernière égalité en loi étant une application de la Proposition 1.1 (Cochran). En effet, la matrice $V(\boldsymbol{p}) = \mathrm{Id}_d - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}$ est la matrice de la projection orthogonale sur l'orthogonal de l'espace vectoriel de dimension 1 engendré par le vecteur $\sqrt{\boldsymbol{p}}$. On a aussi $\mathrm{Rang}\big(V(\boldsymbol{p})\big) = d-1$, d'où le résultat. □

Définition 8.5 (distance du χ^2)**.** *Si* $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{M}_d$ *et les coefficients* $\boldsymbol{q}$ *sont tous non nuls, on appelle distance du* χ^2 *entre les lois* $\boldsymbol{p}$ *et* $\boldsymbol{q}$ *la quantité*

$$\chi^2(\boldsymbol{p}, \boldsymbol{q}) = \sum_{\ell=1}^{d} \frac{(p_\ell - q_\ell)^2}{q_\ell}.$$

Notons $\widehat{\boldsymbol{p}}_n = (\widehat{p}_{n,1}, \ldots, \widehat{p}_{n,d})^T$. La Définition 8.5 est motivée par l'identité

$$\|\boldsymbol{U}_n(\boldsymbol{p})\|^2 = n\chi^2\big(\widehat{\boldsymbol{p}}_n, \boldsymbol{p}\,\big).$$

Remarque 8.9. Le terme « distance » est manifestement impropre, puisque qu'en général on a $\chi^2(\boldsymbol{p},\boldsymbol{q}) \neq \chi^2(\boldsymbol{q},\boldsymbol{p})$. Toutefois, on a la propriété essentielle

$$\chi^2(\boldsymbol{p},\boldsymbol{q}) = 0 \Longleftrightarrow \boldsymbol{p} = \boldsymbol{q}\,.$$

Avec ces notations et la Proposition 8.5, on en déduit le test suivant, appelé test d'adéquation du χ^2.

Proposition 8.6. *Soit $\boldsymbol{q} \in \mathcal{M}_d$ une loi donnée dont les coefficients sont tous non nuls.*

Pour tout $\alpha \in (0,1)$, le test défini par la zone de rejet

$$\mathcal{R}_{n,\alpha} = \left\{ n\chi^2\big(\widehat{\boldsymbol{p}}_n, \boldsymbol{q}\,\big) \geq q^{\chi^2}_{1-\alpha,d-1} \right\}$$

où $q^{\chi^2}_{1-\alpha,d-1}$ est le quantile de la loi du χ^2 à $d-1$ degrés de liberté, est asymptotiquement de niveau α et consistant.

Démonstration. La première partie de la Proposition découle de la Proposition 8.5 : on a $\boldsymbol{p} = \boldsymbol{q}$ sous l'hypothèse, donc

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{p}}\left[(X_1,\ldots,X_n) \in \mathcal{R}_{n,\alpha}\right] &= \mathbb{P}_{\boldsymbol{q}}\left[n\chi^2\big(\widehat{\boldsymbol{p}}_n, \boldsymbol{q}\,\big) \geq q^{\chi^2}_{1-\alpha,d-1}\right] \\
&= \mathbb{P}_{\boldsymbol{q}}\left[\|\boldsymbol{U}_n(\boldsymbol{q})\|^2 \geq q^{\chi^2}_{1-\alpha,d-1}\right] \\
&\to \alpha.
\end{aligned}$$

Pour montrer la consistance, plaçons nous sous l'alternative H_1. Alors $\boldsymbol{p} \neq \boldsymbol{q}$ et $\chi^2(\boldsymbol{p},\boldsymbol{q}) \neq 0$. On a la convergence en probabilité sous $\mathbb{P}_{\boldsymbol{p}}$

$$\chi^2\big(\widehat{\boldsymbol{p}}_n, \boldsymbol{q}\,\big) \xrightarrow{\mathbb{P}_{\boldsymbol{p}}} \chi^2(\boldsymbol{p},\boldsymbol{q}) \neq 0.$$

Donc $n\chi^2\big(\widehat{\boldsymbol{p}}_n, \boldsymbol{q}\,\big)$ diverge vers $+\infty$ en probabilité sous $\mathbb{P}_{\boldsymbol{p}}$. La consistance de la suite de tests en découle par convergence dominée. □

Exemple 8.1 (Mendel)**.** Dans la célèbre expérience de Mendel à l'origine de la génétique, le croisement de pois donne lieu à quatre phénotypes identifiés (combinant couleur et forme). Selon la théorie de l'hérédité de Mendel, si les phénotypes de type I, II, III et IV sont distribués selon une loi multinomiale (voir Section 4.1.2, Chapitre 4) de paramètre

$$\boldsymbol{q} = \Big(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\Big).$$

Mendel rapporte les résultats suivants : pour $n = 556$ observations, la répartition observée entre les phénotypes de type I, II, III et IV est $(315, 101, 108, 32)$. On teste $H_0 : \boldsymbol{p} = \boldsymbol{q}$

contre $H_1 : \boldsymbol{p} \neq \boldsymbol{q}$, où $\boldsymbol{p} \in \mathcal{M}_4$ qui est l'ensemble des lois dont les coefficients sont tous non-nuls. On a ici

$$\begin{aligned} n\chi^2\big(\widehat{\boldsymbol{p}}_n, \boldsymbol{q}\,\big) &= 556\Big(\frac{(\frac{315}{556}-\frac{9}{16})^2}{\frac{9}{16}} + \frac{(\frac{101}{556}-\frac{3}{16})^2}{\frac{3}{16}} + \frac{(\frac{108}{556}-\frac{3}{16})^2}{\frac{3}{16}} + \frac{(\frac{32}{556}-\frac{1}{16})^2}{\frac{1}{16}}\Big) \\ &= 0,47. \end{aligned}$$

Pour le niveau $\alpha = 5\%$, la valeur critique de rejet du test est $q^{\chi^2}_{1-\alpha,3} = 0,7815$ et puisque $0,47 < 0,7815$, on accepte H_0. On peut aussi calculer la p-valeur du test[4]. Dans un cadre asymptotique, si $Z \sim \chi^2(3)$ est distribuée selon la loi du χ^2 avec 3 degrés de liberté, on a donc (voir Proposition 7.1)

$$p - -\text{valeur} = \mathbb{P}_{\boldsymbol{q}}\big[Z > 0,47\big] = 0,93,$$

ce qui ne nous incite pas à rejeter[5] H_0.

8.4.2 Test du χ^2 d'indépendance*

Test du χ^2 avec paramètres estimés

On observe un n-échantillon

$$X_1, \ldots, X_n$$

de loi $\boldsymbol{p} \in \mathcal{M}_d$ inconnue et on teste l'hypothèse nulle composite

$$H_0 : \boldsymbol{p} \in (\mathcal{M}_d)_0 \quad \text{contre} \quad H_1 : \boldsymbol{p} \in \mathcal{M}_d \setminus (\mathcal{M}_d)_0,$$

où $(\mathcal{M}_d)_0 \subset \mathcal{M}_d$. On suppose que $(\mathcal{M}_d)_0$ se représente sous la forme

$$(\mathcal{M}_d)_0 = \big\{\, \boldsymbol{p} = \boldsymbol{p}(\gamma), \;\; \gamma \in \Gamma \big\},$$

où $\Gamma \subset \mathbb{R}^d$ est un sous-ensemble régulier de $\mathbb{R}^d$ de dimension $m < d-1$ (une variété affine ou différentiable de dimension k). La famille $\{\boldsymbol{p},\, \boldsymbol{p} \in \mathcal{M}_d\}$ est régulière au sens du Chapitre 6 et il en va de même pour la famille $\{\boldsymbol{p},\, \boldsymbol{p} \in (\mathcal{M}_d)_0\}$ dès que $\gamma \rightsquigarrow \boldsymbol{p}(\gamma)$ est suffisamment régulière (voir Exercice 6.1). Sans être plus précis pour le moment, cela signifie que les estimateurs du maximum de vraisemblance pour la famille $\{\boldsymbol{p},\, \boldsymbol{p} \in \mathcal{M}_d$ et pour la famille restreinte $\{\boldsymbol{p},\, \boldsymbol{p} \in (\mathcal{M})_0\}$ sont bien définis et asymptotiquement normaux. On peut donc utiliser le test basé sur la statistique du rapport de vraisemblance maximal Λ_n de la Section 8.3.

Nous avons d'abord besoin du résultat auxiliaire suivant :

[4] il s'agit alors ici d'une notion de p-valeur asymptotique, voir Section 7.4 du Chapitre 7.

[5] Attention, rappelons que la signification de $0,93$ nous conduite à ne pas rejeter H_0, mais cela peut être aussi bien dû au fait que H_0 est vrai ou bien que la puissance du test est faible.

Lemme 8.4.1. *On a les estimateurs du maximum de vraisemblance suivants : pour la famille*[6] $\{\boldsymbol{p}, \boldsymbol{p} \in \mathcal{M}_d\}$ *:*

$$\widehat{\boldsymbol{p}}_n^{\mathtt{mv}} = \left(\widehat{p}_{n,1}, \ldots, \widehat{p}_{n,p}\right)^T \tag{8.14}$$

où le vecteur $\left(\widehat{p}_{n,1}, \ldots, \widehat{p}_{n,p}\right)^T$ *est le vecteur des fréquences empiriques défini par 8.10 dans la Section 8.4.1, et pour la famille restreinte* $\{\boldsymbol{p}, \boldsymbol{p} \in (\mathcal{M}_d)_0\}$ *:*

$$\boldsymbol{p}(\widehat{\gamma}_n^{\mathtt{mv}}) = \arg\max_{\gamma \in \Gamma} \sum_{\ell=1}^{d} n\widehat{p}_{n,\ell} \log p_\ell(\gamma).$$

Démonstration. Montrons d'abord (8.14). La loi de l'observation $X_1, \ldots, X_n$ est dominée par la mesure de comptage sur $\{1, \ldots d\}^n$. On a donc

$$\mathcal{L}_n(\boldsymbol{p}, X_1, \ldots, X_n) = \prod_{i=1}^{n} p_{X_i}, \quad \boldsymbol{p} = (p_1, \ldots, p_d)^T,$$

mais cette formule n'est pas très exploitable. En notant $N_\ell = \sum_{i=1}^n 1_{\{X_i=\ell\}}$, on a une correspondance univoque entre $(X_1, \ldots, X_n)$ et $(N_1, \ldots, N_d)$ puisque les X_i ne prennent qu'un nombre fini de valeurs. Ceci permet de réécrire la loi du vecteur $(X_1, \ldots, X_n)$ à l'aide de $(N_1, \ldots, N_d)$.

Pour tous $x_1, \ldots, x_n \in \{1 \ldots, d\}$, avec $\sum_{i=1}^n x_i = n$ et en notant $n_\ell = \sum_{i=1}^n 1_{\{x_i=\ell\}}$, on a

$$\begin{aligned}\mathbb{P}_{\boldsymbol{p}}\left[X_1 = x_1, \ldots, X_n = x_n\right] &= \mathbb{P}_{\boldsymbol{p}}\left[N_1 = n_1, \ldots, N_d = n_d\right] \\ &= \frac{n!}{n_1! \cdots n_d!} \prod_{\ell=1}^{d} p_\ell^{n_i}.\end{aligned}$$

On en déduit que le logarithme de la vraisemblance est

$$\mathcal{L}_n(\boldsymbol{p}, X_1, \ldots, X_n) = c(X_1, \ldots, X_n) + \sum_{\ell=1}^{d} N_\ell \log p_\ell, \tag{8.15}$$

où $c(X_1, \ldots, X_n)$ est une constante qui ne dépend pas de $\boldsymbol{p}$. Donc maximiser la log-vraisemblance revient à chercher le maximum de

$$(p_1, \ldots, p_d) \rightsquigarrow \sum_{i=1}^{d} N_i \log p_i, \text{ sous la contrainte } \sum_{i=1}^{d} p_i = 1.$$

On peut diviser cette fonction par n sans changer le problème. Alors, en notant μ la fonction de comptage sur $\{1, \ldots, d\}$ et $f(x) = N_x/n$ pour $x \in \{1, \ldots, d,$ on cherche à maximiser

$$g \rightsquigarrow \int f(x) \log g(x) \mu(dx)$$

[6] restreinte aux $\boldsymbol{p}$ dont toutes les composantes sont non nulles

avec f et g des densités par rapport à μ. Le Lemme 4.4.1 (inégalité d'entropie) donne la solution $g = f$, soit $p_\ell = N_\ell/n = \widehat{p}_{n,\ell}$. La deuxième partie du lemme découle de la représentation (8.15) de la log-vraisemblance. □

On a le résultat remarquable suivant

Proposition 8.7. *Si Λ_n désigne la statistique du rapport de vraisemblance maximal défini en* (8.4)*, on a, pour tout point $\boldsymbol{p} \in \mathcal{M}_d$*

$$2\Lambda_n = n\chi^2\big(\widehat{\boldsymbol{p}}_n^{\,\mathtt{mv}}, \boldsymbol{p}(\widehat{\gamma}_n^{\,\mathtt{mv}})\big) + \varepsilon_n,$$

où ε_n tend vers 0 en probabilité sous $\mathbb{P}_{\boldsymbol{p}}$ pour tout $\boldsymbol{p} \in \mathcal{M}_0$.

Démonstration. On reprend les notations de la preuve du Lemme 8.4.1. On a

$$2\Lambda_n = \sum_{\ell=1}^{d} N_\ell \Big(\log(N_\ell/n) - \log p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}}) \Big) = 2\sum_{\ell=1}^{d} N_\ell \log \frac{N_\ell}{n p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})}.$$

Sous l'hypothèse nulle, c'est-à-dire si $\boldsymbol{p} = \boldsymbol{p}(\gamma)$ pour un $\gamma \in \Gamma$, on a simultanément

$$\frac{N_\ell}{n} \overset{\mathbb{P}_{\boldsymbol{p}}}{\to} \boldsymbol{p}(\gamma), \quad \text{et} \quad \boldsymbol{p}(\widehat{\gamma}_n^{\,\mathtt{mv}})) \overset{\mathbb{P}_{\boldsymbol{p}}}{\to} \boldsymbol{p}(\gamma).$$

En posant $\varepsilon_{n,\ell} = \frac{N_\ell}{n} - \boldsymbol{p}(\widehat{\gamma})$, on écrit le développement de Taylor du logarithme à l'ordre 2 :

$$\begin{aligned}
2\Lambda_n =& 2n \sum_{\ell=1}^{d} \big(\varepsilon_{n,\ell} + p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})\big) \log\Big(1 + \frac{\varepsilon_{n,\ell}}{p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})}\Big) \\
=& 2n \sum_{\ell=1}^{d} \big(\varepsilon_{n,\ell} + p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})\big) \left(\frac{\varepsilon_{n,\ell}}{p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})} - \frac{1}{2}\left(\frac{\varepsilon_{n,\ell}}{p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})}\right)^2 \big(1 + o_{\boldsymbol{p}}(1)\big) \right) \\
=& 2n \sum_{\ell=1}^{d} \left(\varepsilon_{n,\ell} + \frac{1}{2}\frac{\varepsilon_{n,\ell}^2}{p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})}\big(1 + o_{\boldsymbol{p}}(1)\big) - \frac{1}{2}\frac{\varepsilon_{n,\ell}^3}{p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})^2}\big(1 + o_{\boldsymbol{p}}(1)\big) \right),
\end{aligned}$$

où $o_{\boldsymbol{p}}(1)$ désigne une suite de variables aléatoires qui tend vers 0 en probabilité sous $\mathbb{P}_{\boldsymbol{p}}$.

Les N_ℓ/n et les $p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})$ sont des fréquences empiriques, donc leur somme en ℓ vaut 1 pour chacun d'où $\sum_{\ell=1}^{d} \varepsilon_{n,\ell} = 0$. On en déduit

$$\begin{aligned}
2\Lambda_n &= n \sum_{\ell=1}^{d} \frac{\varepsilon_{n,\ell}^2}{p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})} + \varepsilon_n \\
&= n \sum_{\ell=1}^{d} \frac{\big(N_\ell/n - p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})\big)^2}{p_\ell(\widehat{\gamma}_n^{\,\mathtt{mv}})} + \varepsilon_n \\
&= n\chi^2\big(\widehat{\boldsymbol{p}}_n^{\,\mathtt{mv}}, \boldsymbol{p}(\widehat{\gamma}_n^{\,\mathtt{mv}})\big) + \varepsilon_n,
\end{aligned}$$

où ε_n est une suite de variables aléatoires qui tend vers 0 en probabilité sous $\mathbb{P}_{\boldsymbol{p}}$. □

Ce développement asymptotique permet de construire le test suivant

Proposition 8.8. *Si $\gamma \rightsquigarrow \boldsymbol{p}(\lambda)$ est régulière et Γ de dimension m, on a pour tout point de l'hypothèse $\boldsymbol{p} \in (\mathcal{M})_0$,*

$$n\chi^2\big(\widehat{\boldsymbol{p}}_n^{\mathtt{mv}}, \boldsymbol{p}(\widehat{\gamma}_n^{\mathtt{mv}})\big) \xrightarrow{d} \chi^2(d-m-1).$$

En particulier, le test défini par la zone de rejet

$$\mathcal{R}_{n,\alpha} = \big\{n\chi^2\big(\widehat{\boldsymbol{p}}_n^{\mathtt{mv}}, \boldsymbol{p}(\widehat{\gamma}_n^{\mathtt{mv}})\big) \geq q^{\chi^2}_{1-\alpha,d-m-1}\big\} \tag{8.16}$$

où $q^{\chi^2}_{1-\alpha,d-m-1}$ désigne le quantile de la loi du χ^2 à $d-m-1$ degrés de liberté est asymptotiquement de niveau α et consistant.

Nous admettons ce résultat. On pourra consulter van der Vaart [7] ou Borovkov [1] pour une preuve et des compléments.

Définition 8.6 (Test du χ^2 avec paramètres estimés)**.** *On appelle test du χ^2 avec paramètres estimés le test de zone de rejet définie par* (8.16).

Application au test d'indépendance

Un cas très classique du test du χ^2 avec paramètres estimés et celui du test d'indépendance. On observe un n-échantillon

$$(X_1, Y_1), \ldots, (X_n, Y_n) \tag{8.17}$$

où les variables X_i et Y_i sont qualitatives, prenant respectivement à d_1 et d_2 valeurs possibles. La loi $\boldsymbol{p}$ du couple (X, Y) est à valeurs dans

$$\mathcal{M}_{d_1,d_2} = \Big\{\boldsymbol{p} = (p_{\ell,\ell'})_{1\leq\ell\leq d_1, 1\leq\ell'\leq d_2},\ 0 \leq p_{\ell,\ell'} \leq 1,\ \sum_{\ell,\ell'} p_{\ell,\ell'} = 1\Big\}.$$

Notons les lois marginales du vecteur $(X, Y)^T$.

$$p_{\ell,\bullet} = \mathbb{P}\big[X = \ell\big], \quad \text{et} \quad p_{\bullet,\ell'} = \mathbb{P}\big[Y = \ell'\big]$$

pour $1 \leq \ell \leq d_1, 1 \leq \ell' \leq d_2$, et où on a

$$p_{\ell,\bullet} = \sum_{\ell'=1}^{d_2} p_{\ell,\ell'}, \quad p_{\bullet,\ell'} = \sum_{\ell=1}^{d_1} p_{\ell,\ell'}.$$

On teste l'indépendance des variables X et Y à partir de l'observation du n-échantillon (8.17). Cela se traduit par l'hypothèse nulle :

$$H_0 : \forall \ell, \ell' \quad p_{\ell,\ell'} = p_{\ell,\bullet} p_{\bullet,\ell'}$$

contre l'alternative

$$H_1 : \exists\, \ell, \ell', \quad p_{\ell,\ell'} \neq p_{\ell,\bullet} p_{\bullet,\ell'}.$$

Ici, l'hypothèse nulle s'écrit

$$H_0 : \boldsymbol{p} \in (\mathcal{M}_{d_1,d_2})_0 = \Big\{ \boldsymbol{p} = (p_{\ell,\ell'}),\ p_{\ell,\ell'} = p_{\ell,\bullet} p_{\bullet,\ell'} \Big\}$$

et donc $(\mathcal{M}_{d_1,d_2})_0$ est en correpondance avec $\mathcal{M}_{d_1} \times \mathcal{M}_{d_2}$. On applique alors les résultats de la section précédente avec $m = (d_1 - 1)(d_2 - 1) < d_1 d_2 - 1$. Il nous faut pour cela connaître l'estimateur du maximum de vraisemblance sur $(\mathcal{M}_{d_1,d_2})_0$.

Lemme 8.4.2. *Pour la famille* $\big\{ \boldsymbol{p},\, \boldsymbol{p} \in (\mathcal{M}_{d_1,d_2})_0 \big\}$*, l'estimateur du maximum de vraisemblance* $\widehat{\boldsymbol{p}}^{\,\mathtt{mv}}_{n,0}$ *s'écrit*

$$\big(\widehat{p}^{\,\mathtt{mv}}_{n,0}\big)_{\ell,\ell'} = \widehat{p}_{n,(\ell,\bullet)}\, \widehat{p}_{n,(\bullet,\ell')}$$

pour $1 \leq \ell \leq d_1$, $1 \leq \ell' \leq d_2$, *avec*

$$\widehat{p}_{n,(\ell,\bullet)} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i = \ell\}} \quad \textit{et} \quad \widehat{p}_{n,(\bullet,\ell')} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i = \ell'\}}$$

les fréquences empiriques marginales, qui sont aussi les estimateurs de maximum de vraisemblance correspondants aux familles des marginales d'après le Lemme 8.4.1.

Démonstration. C'est essentiellement la même preuve que celle du Lemme 8.4.1. Si $\boldsymbol{p} \in (\mathcal{M}_{d_1,d_2})_0$, les variables aléatoires X_i et Y_i sont indépendantes, et la vraisemblance s'écrit

$$\mathcal{L}_n\big(\boldsymbol{p}, (X_1, Y_1), \dots (X_n, Y_n)\big) = \prod_{i=1}^{n} p_{X_i,\bullet}\, p_{\bullet,Y_i} = \big(\prod_{i=1}^{n} p_{X_i,\bullet}\big)\big(\prod_{i=1}^{n} p_{\bullet,Y_i}\big).$$

En notant $N_\ell^X = \sum_{i=1}^{\ell} 1_{\{X_i=\ell\}}$ et $N_{\ell'} = \sum_{i=1}^{n} 1_{\{Y_i=\ell'\}}$ et en passant au logarithme, on obtient

$$\begin{aligned} &\log \mathcal{L}_n\big(\boldsymbol{p}, (X_1, Y_1), \dots (X_n, Y_n)\big) \\ &= c(X_1, \dots, X_n, Y_1, \dots, Y_n) + \sum_{\ell=1}^{d_1} N_\ell^X \log p_{\ell,\bullet} + \sum_{\ell'=1}^{d_2} N_\ell^Y \log p_{\bullet,\ell'}, \end{aligned}$$

où $c(X_1, \dots, X_n, Y_1, \dots, Y_n)$ ne dépend pas de $\boldsymbol{p}$, et on raisonne comme pour le Lemme 8.4.1 en remplaçant $\{1, \dots, d\}$ par $\{1, \dots, d_1 + d_2\}$. □

Par ailleurs, le Lemme 8.4.1 donne l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{p}}^{\,\mathtt{mv}}_n$ pour la famille globale $\big\{ \boldsymbol{p},\, \boldsymbol{p} \in \mathcal{M}_{d_1,d_2} \big\}$ qui est l'estimateur des fréquences empiriques

$$(\widehat{\boldsymbol{p}}_n)_{\ell,\ell'} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{(X_i,Y_i)=(\ell,\ell')\}}$$

pour $1 \leq \ell \leq d_1$, $1 \leq \ell' \leq d_2$.

Alors, comme précédemment, sous l'hypothèse nulle, c'est-à-dire pour $\boldsymbol{p} \in (\mathcal{M}_{d_1,d_2})_0$ on a la convergence

$$n\chi^2\Big(\widehat{\boldsymbol{p}}_n^{\mathtt{mv}}, \widehat{\boldsymbol{p}}_{n,0}^{\mathtt{mv}}\Big) \xrightarrow{d} \chi^2\Big((d_1-1)(d_2-1)\Big)$$

en loi sous $\mathbb{P}_{\boldsymbol{p}}$. En particulier, la statistique de test s'écrit

$$n\chi^2\Big(\widehat{\boldsymbol{p}}_n^{\mathtt{mv}}, \widehat{\boldsymbol{p}}_{n,0}^{\mathtt{mv}}\Big) = n\sum_{\ell,\ell'} \frac{\Big((\widehat{\boldsymbol{p}}_n)_{\ell,\ell'} - \widehat{p}_{n,(\ell,\bullet)}\widehat{p}_{n,(\bullet,\ell')}\Big)^2}{\widehat{p}_{n,(\ell,\bullet)}\widehat{p}_{n,(\bullet,\ell')}}.$$

Proposition 8.9 (Test d'indépendance du χ^2)**.** *Pour tout $\alpha \in (0,1)$, le test défini par la zone de rejet*

$$\mathcal{R}_{n,\alpha} = \Big\{n\chi^2\Big(\widehat{\boldsymbol{p}}_n^{\mathtt{mv}}, \widehat{\boldsymbol{p}}_{n,0}^{\mathtt{mv}}\Big) \geq q^{\chi^2}_{1-\alpha,(d_1-1)(d_2-1)}\Big\},$$

où $q^{\chi^2}_{1-\alpha,(d_1-1)(d_2-1)}$ est le quantile d'ordre α de la loi du χ^2 à $(d_1-1)(d_2-1)$ degrés de liberté est asymptotiquement de niveau α et consistant.

Nous admettons la démonstration de ce résultat est essentiellement une application de la Proposition 8.8.

Exemple 8.2. On test l'indépendance entre le nombre d'enfants d'un ménage et son revenu[7] sur une population de $n = 25263$ ménages en Suède au milieu du siècle passé. Les ménages sont classés en 4 catégories selon leur revenus la catégorie I correspond aux revenus les plus faibles et la catégorie IV aux revenus les plus élevés. Les résultats obtenus sont les suivants :

nb. enfants	I	II	III	IV	pop.
0	2161	3577	2184	1636	9558
1	2755	5081	2222	1052	11110
2	936	1753	640	306	3635
3	225	419	96	38	778
≥ 4	39	98	31	14	182
pop.	6116	10928	5173	3016	25263

Sans préjuger de la pertinence de la modélisation, on met en place un test du χ^2 d'indépendance pour la loi $\boldsymbol{p} \in \mathcal{M}_{4,5}$ de la variable (nombre d'enfants, revenu) à valeurs dans $\{0,1,2,3,\geq 4\} \times \{I, II, III, IV\}$ dont la distribution empirique est donnée par le tableau ci-dessus et dont les marginales empiriques se lisent sur la dernière colonne et la dernière ligne. On trouve

$$n\chi^2\Big(\widehat{\boldsymbol{p}}_n^{\mathtt{mv}}, \widehat{\boldsymbol{p}}_{n,0}^{\mathtt{mv}}\Big) = 568,5$$

[7] d'après [1], p 354

ce qui est significativement plus grand que le quantile d'ordre $1-\alpha$ pour une loi du χ^2 à $(5-1)(4-1) = 12$ degrés de liberté, même pour des petites valeurs de α. Dans ces conditions, on rejette l'hypothèse d'indépendance.

Bibliographie

[1] Borovkov, A. A. *Mathematical statistics* (traduit du russe). Gordon and Breach science publishers, 1998.

[2] Genon-Catalot, V., Picard, D. *Éléments de statistique asymptotique*. Mathématiques & Applications. Springer-Verlag, Paris, 1993.

[3] Jacod, J. et Protter, P. *Probability essentials*. Seconde édition. Universitext. Springer-Verlag, Berlin, 2003.

[4] Méléard, S. *Aléatoire*. Polycopié de l'École polytechnique.

[5] Monfort, A. *Statistique*. Polycopié de l'École polytechnique (version éditée par O. Cappé).

[6] Tsybakov, A. *Statistique Appliquée*. Polycopié de l'Université de Pierre et Marie Curie.

[7] van der Vaart, A. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge, 1998.

[8] Wasserman, L. *All of statistics. A concise course in statistical inference*. Springer Texts in Statistics. Springer-Verlag, New York, 2004.

Index