Statistique mathématique pour le Master 1
Cours de l'ENS Cachan Bretagne

Benoît Cadre

4 juin 2010

Table des matières

Chapitre 1

Modélisation statistique

1.1 Un exemple

Une pièce a une probabilité $p_0 \in]0,1[$ de tomber sur "pile". Sur les 1000 lancers réalisés indépendamment les uns des autres, on compte 520 "pile" et 480 "face". On est donc tenté de conclure que $p_0 \approx 0.52$. Cependant, de la même manière qu'il est sans intérêt de donner une valeur approchée d'une intégrale sans préciser l'erreur d'approximation, ce résultat n'a que peu de valeur, car il ne nous renseigne pas sur l'erreur commise.

Nous allons examiner de quelle manière la construction d'un modèle permet de combler cette lacune. On note $x_1, \cdots, x_n$ les résultats des $n = 1000$ lancers de pièce, avec la convention suivante : $x_i = 1$ si le i-ème lancer a donné "pile", et 0 dans le cas contraire. Le principe de base de l'estimation statistique est de considérer que $x_1, \cdots, x_n$ est une *réalisation* de la loi $\mathscr{B}(p_0)^{\otimes n}$, si pour chaque $p \in [0,1]$, $\mathscr{B}(p)$ désigne la loi de Bernouilli de paramètre p (i.e. $\mathscr{B}(p) = p\delta_1 + (1-p)\delta_0$, avec δ_0 et δ_1 les mesures de Dirac en 0 et 1). En l'absence d'informations sur la valeur de p_0, on ne peut en fait que supposer que $x_1, \cdots, x_n$ est une réalisation de l'une des lois $\{\mathscr{B}(p)^{\otimes n}, p \in]0,1[\}$.

De cet ensemble de probabilités, appelé *modèle statistique*, on cherche à déduire la valeur de p qui s'ajuste le mieux aux *observations* $x_1, \cdots, x_n$. Une réponse raisonnable est basée sur l'intuition suivante : compte tenu des informations dont on dispose, la meilleure approximation de p_0 que l'on puisse donner est une valeur

qui maximise la fonction

$$p \mapsto \mathscr{B}(p)^{\otimes n}(\{x_1,\cdots,x_n\}) = \prod_{i=1}^{n} \mathscr{B}(p)(\{x_i\}) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

C'est le principe de construction d'une valeur approchée -on parlera d'*estimateur*- de p_0 par *maximisation de la vraisemblance*. Selon ce principe, la valeur qui s'ajuste le mieux aux observations est la *moyenne empirique* des observations :

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

On retrouve ainsi la valeur $\bar{x}_n = 0.52$ du début.

L'introduction d'un modèle nous permet en plus de donner une erreur dans l'approximation. Soit $p \in]0,1[$, et $X_1,\cdots,X_n$ des v.a. i.i.d. sur l'espace probabilisé $(\Omega,\mathscr{F},\mathbb{P})$ de loi commune $\mathscr{B}(p)$. On peut calculer le *risque quadratique*, c'est-à-dire le carré de la distance L^2 entre la cible p et l'estimateur $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ obtenu par le principe de maximisation de la vraisemblance :

$$\mathbb{E}\,(\bar{X}_n - p)^2 = \frac{1}{n}\mathbb{E}X_1(1-\mathbb{E}X_1) = \frac{1}{n}p(1-p).$$

Comme $p(1-p) \leq 1/4$, l'erreur quadratique moyenne commise est donc majorée par $1/(2\sqrt{n}) \approx 0.016$. Cependant, si le résultat donne des informations sur la qualité de l'approximation, ce n'est qu'une évaluation en moyenne, qui ne dépend donc pas des observations.

Bien d'autres principes peuvent être envisagés pour préciser la qualité de l'approximation. Par exemple, supposons que l'on veuille construire un intervalle dans lequel p_0 doit se trouver, avec une probabilité de 0.95 par exemple. Le principe de construction est le suivant : pour chaque $p \in]0,1[$, on cherche dans un premier temps un *intervalle de confiance par excès* $I(X_1,\cdots,X_n)$ construit avec la suite de v.a. $X_1,\cdots,X_n$ tel que

$$\mathbb{P}\,(p \in I(X_1,\cdots,X_n)) \geq 0.95.$$

On peut alors conclure, avec les observations $x_1,\cdots,x_n$, que $p_0 \in I(x_1,\cdots,x_n)$, avec une probabilité de 95% au moins. Dans l'exemple qui nous intéresse, l'inégalité de Bienaymé-Tchebytchev nous donne, pour tout $\varepsilon > 0$:

$$\mathbb{P}(|\bar{X}_n - p| \geq \varepsilon) \leq \frac{\mathrm{var}(\bar{X}_n)}{\varepsilon^2} = \frac{\mathrm{var}(X_1)}{n\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

De ce fait, $\mathbb{P}(|\bar{X}_n - p| \geq \varepsilon) \leq 0.05$ au moins si $1/(4n\varepsilon^2) \leq 0.05$ soit, tous calculs faits, si $\varepsilon \geq 0.07$. Par suite,

$$\mathbb{P}\left(p \in [\bar{X}_n - 0.07, \bar{X}_n + 0.07]\right) \geq 0.95.$$

En utilisant les observations $x_1, \cdots, x_n$ on a $\bar{x}_n = 0.52$, et donc $p_0 \in [0.45, 0.59]$ avec une probabilité 0.95 au moins. Le moins que l'on dire ici est que cette information est peu satisfaisante, eu égard au grand nombre d'observations !

Comment améliorer ces résultats ? Si la question posée est "la pièce est-elle équilibrée ?", l'intervalle ci-dessus ne permet pas de donner une réponse ; dès lors, quelle stratégie de décision envisager ? L'objet de ce cours est de donner quelques éléments de réponse à ces questions. Dans un premier temps, il convient de fixer les objectifs de l'inférence statistique, ainsi que le contexte mathématique.

1.2 Principe fondamental de la statistique

Le phénomène aléatoire fournit n observations $x_1, \cdots, x_n$ de l'espace topologique $\mathscr{H}$. Celles-ci sont supposées être les réalisations d'une loi Q_0 sur l'espace probabilisable $(\mathscr{H}, \mathscr{B}(\mathscr{H}))$. Le principe de base de l'inférence statistique est d'utiliser ces n observations pour en déduire des informations sur Q_0. Cette approche est-elle raisonnable ? De manière plus ambitieuse, est-il possible de reconstruire un approximation de Q_0 à partir des observations $x_1, \cdots, x_n$? Nous allons voir que la réponse est affirmative, au moins dans le cas où le phénomène aléatoire global consiste en n phénomènes indépendants et régis par la même loi.

Au préalable, rappelons que la suite de probabilités $(\nu_n)_n$ sur $\mathbb{R}^d$ converge *étroitement* vers ν si, pour chaque fonction $f : \mathbb{R}^d \to \mathbb{R}$ continue bornée, on a :

$$\int_{\mathbb{R}^d} f \mathrm{d}\nu_n \to \int_{\mathbb{R}^d} f \mathrm{d}\nu.$$

Le critère de convergence de Lévy nous affirme que $(\nu_n)_n$ converge étroitement vers ν si, pour chaque $t \in \mathbb{R}^d$, la suite $(\hat{\nu}_n(t))_n$ converge vers $\hat{\nu}(t)$, où $\hat{\nu}$ désigne la tranformée de Fourier de ν, i.e. la fonction

$$\hat{\nu} : t \mapsto \int_{\mathbb{R}^d} \exp(it^T x) \nu(\mathrm{d}x),$$

et idem pour $\hat{\nu}_n$.

Dans la suite, δ_x désigne la mesure de Dirac en $x \in \mathbb{R}^d$.

Théorème [VARADARAJAN] *Soient $X_1, X_2, \cdots$ des v.a.i.i.d. sur $(\Omega, \mathscr{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}^k$, de loi commune μ. On note μ_n la mesure empirique des n premières v.a., i.e.*

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Alors, $\mathbb{P}$-p.s., la suite de mesures $(\mu_n)_n$ converge étroitement vers μ.

Preuve Pour simplifier la preuve, on suppose que X_1 est intégrable. D'après le critére de Lévy, il suffit de montrer que

$$\mathbb{P}\left(\forall t \in \mathbb{R}^d : \hat{\mu}_n(t) \longrightarrow \hat{\mu}(t)\right) = 1,$$

si $\hat{\mu}_n$ et $\hat{\mu}$ désignent les transformées de Fourier de μ_n et μ. Or, la loi forte des grands nombres nous montre que pour tout $t \in \mathbb{R}^d$, l'événement

$$\Omega(t) = \{\hat{\mu}_n(t) \longrightarrow \hat{\mu}(t)\}$$

est de probabilité. Soit $T \subset \mathbb{R}^d$ un ensemble dénombrable dense, et

$$\Omega_0 = \bigcap_{t \in T} \Omega(t) \bigcap \bigcap \left\{ \frac{1}{n} \sum_{j=1}^{n} \|X_j\| \longrightarrow \mathbb{E}\|X_1\| \right\},$$

où $\|.\|$ désigne la norme euclidienne de $\mathbb{R}^d$. Comme X_1 est intégrable et T est dénombrable, on a $\mathbb{P}(\Omega_0) = 1$ d'après la loi forte des grands nombres et car $\mathbb{P}(\Omega(t)) = 1$ pour tout t. Fixons $t \in \mathbb{R}^d$ et $\omega \in \Omega_0$. On choisit une suite $(t_p)_p$ de T telle que $t_p \to t$, et on note pour tout $s \in \mathbb{R}^d$, $\hat{\mu}_n^{\omega}(s)$ la réalisation en ω de $\hat{\mu}_n(s)$, i.e.

$$\hat{\mu}_n^{\omega}(s) = \frac{1}{n} \sum_{j=1}^{n} \exp(is^T X_j(\omega)).$$

Soit p fixé. On a :

$$\begin{aligned} |\hat{\mu}_n^{\omega}(t) - \hat{\mu}(t)| &\leq |\hat{\mu}_n^{\omega}(t) - \hat{\mu}_n^{\omega}(t_p)| + |\hat{\mu}_n^{\omega}(t_p) - \hat{\mu}(t_p)| + |\hat{\mu}(t_p) - \hat{\mu}(t)| \\ &\leq \|t - t_p\| \left(\frac{1}{n} \sum_{j=1}^{n} \|X_j(\omega)\| + \mathbb{E}\|X_1\| \right) + |\hat{\mu}_n^{\omega}(t_p) - \hat{\mu}(t_p)| \end{aligned}$$

Par suite, pour tout p :

$$\limsup_n |\hat{\mu}_n^{\omega}(t) - \hat{\mu}(t)| \leq 2\|t - t_p\|\mathbb{E}\|X_1\|.$$

En faisant enfin tendre p vers l'infini, on peut en déduire que pour tout $\omega \in \Omega_0$ et tout $t \in \mathbb{R}^d$, $\hat{\mu}_n^{\omega}(t) \to \hat{\mu}(t)$. Comme $\mathbb{P}(\Omega_0) = 1$, le résultat est démontré. □

Reprenons le contexte où les observations $(x_1, \cdots, x_n) \in \mathscr{H}^n$ sont issues de n phénomènes aléatoires indépendants et régis par la même loi Q_0 sur $\mathscr{H} = \mathbb{R}^d$. Le théorème de Varadarajan montre que si $(X_1, \cdots, X_n)$ suit la loi $Q_0^{\otimes n}$ alors la *mesure empirique*

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

est proche de la mesure Q_0, lorsque n est assez grand. Autrement dit, en multipliant les expériences, la mesure discrète

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

est proche de la mesure Q_0. Ce résultat donne un appui théorique à la démarche statistique consistant à tenter de reconstruire la mesure théorique Q_0 à l'aide des observations $x_1, \cdots, x_n$. Toute démarche en statistique inférentielle asymptotique est basée sur ce théorème, qui mérite donc son titre de "*Théorème fondamental de la statistique"*.

1.3 Modèle statistique

Formalisons le concept de modèle statistique vu dans la section 1.1. Dans ce cadre, l'espace des observations était $\{0,1\}^n$.

Définitions *Un modèle statistique est un couple $(\mathscr{H}^n, \mathscr{P})$, où $\mathscr{H}$ est l'espace -supposé topologique- de chaque observation, et $\mathscr{P}$ est une famille de lois de probabilités sur $\mathscr{H}^n$ muni de sa tribu borélienne. Le modèle statistique $(\mathscr{H}^n, \mathscr{P})$ est paramétré par Θ si $\mathscr{P} = \{P_\theta\}_{\theta \in \Theta}$*

L'expérience aléatoire sous-jacente fournit n observations $(x_1, \cdots, x_n) \in \mathscr{H}^n$ du même phénomène aléatoire, qui est régit par la loi inconnue P_0. Le principe

de base de la statistique est de considérer que $(x_1, \cdots, x_n)$ est régit par l'une des lois d'un modèle $\mathscr{P}$, avec $P_0 \in \mathscr{P}$. Cette étape de modélisation étant achevée, il s'agira de chercher quelle loi de ce modèle s'ajuste le mieux aux observations.

Par exemple, lorsque les expériences ont été menées indépendamment les unes des autres, l'observation $(x_1, \cdots, x_n)$ est régie par la loi $P_0 = Q_0^{\otimes n}$, et le modèle statistique est un ensemble de probabilités sur $\mathscr{H}^n$ contenant $Q_0^{\otimes n}$.

A noter, donc : à l'inverse du probabiliste, le statisticien travaille plutôt sur l'espace des observations, qui constitue un cadre d'étude plus naturel. Par ailleurs, le statisticien ne suppose pas que la loi des observations est connue, à l'inverse du probabiliste.

Exemple En utilisant des observations indépendantes $x_1, \cdots, x_n$ de la durée de vie de n ampoules du même type, on veut connaître la loi suivie par la durée de vie de ce type d'ampoule. La 1ère étape consiste à définir le modèle statistique associé, dont l'espace des observations est $\mathbb{R}_+^n$. Du point de vue de la modélisation, il est raisonnable d'affirmer qu'une v.a. X sur $(\Omega, \mathscr{F}, \mathbb{P})$ qui représente la durée de vie de l'ampoule est *sans mémoire*, i.e.

$$\mathscr{L}(X - t | X \geq t) = \mathscr{L}(X), \ \forall t \geq 0.$$

Cette propriété signifie que l'ampoule "ne se souvient pas d'avoir vieilli". Par ailleurs, on peut aussi supposer que la loi de X est à densité par rapport à la mesure de Lebesgue. On sait alors qu'il existe $\lambda > 0$ tel que $X \sim \mathscr{E}(\lambda)$. Comme les observations des durées de vie sont indépendantes, $x_1, \cdots, x_n$ est une réalisation d'une loi $\mathscr{E}(\lambda_0)^{\otimes n}$, pour un certain $\lambda_0 > 0$ qu'il s'agira de trouver. Le modèle statistique associé à cette expérience aléatoire est donc $(\mathbb{R}_+^n, \{\mathscr{E}(\lambda)^{\otimes n}\}_{\lambda > 0})$. Nous verrons dans la suite comment trouver une valeur de λ qui s'ajuste aux observations.

Dans l'exemple de la section 1.1, comme les lancers de pièce sont indépendants, la loi dont sont issues les résultats de l'expérience est clairement l'une des lois du modèle $\mathscr{P} = \{\mathscr{B}(p)^{\otimes n}, p \in]0,1[\}$. Remarquons aussi que l'application $p \mapsto \mathscr{B}(p)^{\otimes n}$ est injective : cette propriété, appelée identifiabilité, ôte tout ambiguité dans le modèle, en permettant d'associer à des observations une, et une seule loi du modèle.

Définitions *Soit $\mathscr{P} = \{P_\theta\}_{\theta \in \Theta}$ un modèle statistique.*

1. *Il est dit identifiable si l'application $\theta \mapsto P_\theta$ définie sur Θ est injective.*
2. *Il est dit paramétrique si il existe $d \in \mathbb{N}$ tel que $\Theta \subset \mathbb{R}^d$. Sinon, il est non paramétrique.*

Le modèle statistique $\{N(m,\sigma^2); m \in \mathbb{R}, \sigma > 0\}$ est paramétrique et identifiable, mais $\{N(m,\sigma^2); m \in \mathbb{R}, \sigma \neq 0\}$, qui est aussi paramétrique, n'est pas identifiable car $N(m,\sigma^2) = N(m,(-\sigma)^2)$. Par ailleurs, le modèle constitué de toutes les lois à densité est non paramétrique.

On s'intéressera dans ce cours aux modèles paramétriques. Cette restriction confère au modèle un atout majeur : en limitant l'espace des probabilités à explorer, elle permet d'obtenir de meilleurs résultats quantitatifs.

1.4 Domination dans un modèle statistique

Soit le modèle statistique paramétrique $(\mathscr{H}^n, \mathscr{P})$, avec un espace d'observations individuelles $\mathscr{H} \subset \mathbb{R}^k$.

Rappelons que, pour 2 mesures σ-finies μ et ν sur $\mathbb{R}^p$, μ est absolument continue par rapport à ν, et on note $\mu \ll \nu$, si pour tout $A \in \mathscr{B}(\mathbb{R}^p)$ tel que $\nu(A) = 0$, on a $\mu(A) = 0$. Dans ce cas, le théorème de Radon-Nikodym nous donne l'existence d'une fonction mesurable f et ν-p.p. positive, appelée densité de μ par rapport à ν, telle que $\mathrm{d}\mu = f\mathrm{d}\nu$. Si ν est la mesure de Lebesgue, on parle plus simplement de la densité de μ. Enfin, si μ est bornée, f est ν-intégrable.

Définition *Le modèle statistique $(\mathscr{H}^n, \mathscr{P})$ est dit dominé si il existe une mesure σ-finie μ telle que $P \ll \mu$ pour chaque $P \in \mathscr{P}$. La mesure μ est appelée mesure dominante du modèle.*

Les modèles $\{N(m,\sigma^2); m \in \mathbb{R}, \sigma > 0\}$ et $\{\mathscr{B}(p)^{\otimes n}; p \in]0,1[\}$ sont dominés : une mesure dominante du premier est la mesure de Lebesgue sur $\mathbb{R}$, alors qu'une mesure dominante du second est $(\delta_0 + \delta_1)^{\otimes n}$. De manière plus générale, les exemples de modèles dominés que nous rencontrerons le seront soit par rapport à une mesure de comptage, soit par rapport à une mesure de Lebesgue.

Théorème *Supposons que $(\mathscr{H}^n, \mathscr{P})$ est dominé, et notons* conv$(\mathscr{P})$ *son convexi-*

fié, i.e.

$$\mathrm{conv}(\mathscr{P}) = \left\{ \sum_n a_n P_n, \text{ avec } P_k \in \mathscr{P}, a_k \geq 0 \text{ et } \sum_n a_n = 1 \right\}.$$

Alors, il existe une probabilité de $\mathrm{conv}(\mathscr{P})$ *qui domine* $\mathscr{P}$.

Preuve Soit μ une mesure dominante. On note $\mathscr{C}$ l'ensemble des événements C tels que $\mu(C) > 0$ et tels qu'il existe $P_C \in \mathrm{conv}(\mathscr{P})$ dont la densité f_C par rapport à μ vérifie $f_C > 0$ μ-p.p. sur C. Choisissons $(C_n)_{n\geq 1}$, une suite de $\mathscr{C}$, telle que :

$$\lim_{n\to\infty} \mu(C_n) = \sup_{C\in\mathscr{C}} \mu(C) \leq +\infty.$$

On note P_{C_n} la probabilité associée à chaque C_n, et f_{C_n} la densité associée. On pose :

$$C_s = \bigcup_{n\geq 1} C_n, \quad f = \sum_{n\geq 1} 2^{-n} f_{C_n}.$$

La probabilité Q telle que $\mathrm{d}Q = f\mathrm{d}\mu$, qui est dans $\mathrm{conv}(\mathscr{P})$, admet f pour densité par rapport à μ. Comme $\mu(C_s) > 0$ et $f > 0$ μ-p.p. sur C_s, on a $C_s \in \mathscr{C}$. Par ailleurs, on a aussi $\mu(C_s) = \sup_{C\in\mathscr{C}} \mu(C)$.

Montrons maintenant que Q domine $\mathscr{P}$. Soit $P \in \mathscr{P}$, de densité p par rapport à μ, et A un événement tel que $Q(A) = 0$. Comme $0 = Q(A \cap C_s) = \int_{A\cap C_s} f\mathrm{d}\mu$ et que $f > 0$ μ-p.p. sur C_s, on a $\mu(A \cap C_s) = 0$, d'où $P(A \cap C_s) = 0$ car $P \ll \mu$. Par ailleurs, $P(C_s^c) = 0$. En effet, il est clair que $C_s \subset \{f + p > 0\}$ μ-p.p., et comme $\{f + p > 0\} \in \mathscr{C}$, la propriété de maximalité de C_s montre que $C_s = \{f + p > 0\}$ μ-p.p. Comme $P \ll \mu$, on a aussi $C_s = \{f + p > 0\}$ P-p.p. et donc $P(C_s^c) = P(\{f + p = 0\}) \leq P(\{p = 0\}) = \int_{\{p=0\}} p\mathrm{d}\mu = 0$. En remarquant finalement que $A \subset (A \cap C_s) \cup C_s^c$, on en déduit que $P(A) = 0$, c'est-à-dire que $P \ll Q$. Comme $Q \in \mathrm{conv}(\mathscr{P})$, le théorème est démontré. □

1.5 Estimation

Soit le modèle statistique paramétrique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, avec un espace d'observations individuelles $\mathscr{H} \subset \mathbb{R}^k$ et un espace de paramètres $\Theta \subset \mathbb{R}^d$. Dans ce modèle, le paramètre d'intérêt est θ. Si les n expériences du phénomène sont indépendantes, on a alors $P_\theta = Q_\theta^{\otimes n}$ pour chaque $\theta \in \Theta$.

Dans un soucis de simplicité, on se limitera dans tout le cours au cas où le paramètre d'intérêt est θ, étant entendu que les définitions et la plupart des propriétés qui suivent s'adaptent sans difficulté au cas où le paramètre d'intérêt est une fonction borélienne de θ.

Définition *Un échantillon de loi P_θ est une v.a. canonique sur $(\mathscr{H}^n, P_\theta)$.*

On rappelle qu'une v.a. canonique $(X_1, \cdots, X_n)$ sur $(\mathscr{H}^n, P_\theta)$ est une v.a. qui vérifie pour chaque $i = 1, \cdots, n$:

$$X_i : (x_1, \cdots, x_n) \in \mathscr{H}^n \mapsto x_i \in \mathscr{H}.$$

La taille de l'échantillon est le nombre d'expériences aléatoires. Dans l'exemple de la section 1.1, la taille de l'échantillon est $n = 1000$, et une suite $X_1, \cdots, X_n$ de v.a.i.i.d. issues de la loi $\mathscr{B}(p)$ constitue, après concaténation, un échantillon de la loi $\mathscr{B}(p)^{\otimes n}$. A l'aide de cette modélisation stochastique, l'enjeu est de construire une fonction de l'échantillon, qui fournira l'information sur le paramètre inconnu, noté p_0 dans la section 1.1. Ceci nous amène à la notion d'estimateur, qui est une quantité censé être proche du paramètre. Différentes notions de proximité seront abordées au chapitre 2.

Définitions

1. *Une statistique est une v.a. définie sur $\mathscr{H}^n$ indépendante de θ, i.e. une fonction borélienne définie sur $\mathscr{H}^n$ indépendante de θ.*
2. *Un estimateur (de θ) est une statistique à valeurs dans un sur-ensemble de Θ.*

Remarque Un échantillon de loi P_θ étant une v.a. canonique $(X_1, \cdots, X_n)$, il s'ensuit qu'une statistique s'écrit aussi :

$$g(.) = g(X_1, \cdots, X_n).$$

On utilisera l'une ou l'autre de ces représentations, selon le contexte. Par exemple, pour insister sur le fait que la statistique dépend de l'échantillon, on utilisera la représentation $g(X_1, \cdots, X_n)$. Pour distinguer une statistique d'un estimateur, on notera ce dernier avec un chapeau.

Dans l'exemple de la section 1.1, si $(X_1, \cdots, X_n)$ est un échantillon de la loi $\mathscr{B}(p)^{\otimes n}$, X_1 et $\bar{X}_n$ sont des estimateurs de p. Ces 2 estimateurs n'ont évidemment

pas le même intérêt ; la terminologie du chapitre 2 permettra d'entreprendre une première classification.

1.6 Construction des estimateurs

Soit le modèle statistique paramétrique $(\mathscr{H}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$, avec un espace d'observations individuelles $\mathscr{H} \subset \mathbb{R}^k$ et un espace de paramètres $\Theta \subset \mathbb{R}^d$.

Pour construire un estimateur raisonnable, on utilise en général l'une ou l'autre des 2 procédures suivantes : le principe de la vraisemblance maximale, qui fera l'objet du chapitre 3, ou une méthode *ad hoc* dans laquelle, par le calcul, on observe tout d'abord ce que représente le paramètre θ pour la loi Q_θ, puis on en construit une version empirique.

Examinons en détail la 2ème méthode. Dans un premier temps, on regarde ce que ce paramètre représente pour la loi Q_θ, puis on remplace la mesure Q_θ par sa version empirique. Supposons par exemple que $\theta = \int_{\mathscr{H}} f \mathrm{d}Q_\theta$, pour une certaine fonction connue $f \in L^1(Q_\theta)$. En vertu de la loi des grands nombres, un estimateur raisonnable sera :

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Un tel procédé de construction s'appelle *méthode des moments*, bien qu'il ne concerne pas nécessairement les moments de la loi Q_θ. Bien entendu, ce n'est qu'un procédé de construction, et rien ne nous assure en général qu'un estimateur construit de la sorte ait de bonnes propriétés statistiques. Néanmoins, on retrouve des estimateurs naturels. Par exemple, si θ représente la moyenne de la loi Q_θ, l'estimateur construit par cette méthode sera la *moyenne empirique* :

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

Par ailleurs, si θ représente la variance de la loi Q_θ, l'estimateur sera la *variance empirique* :

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

D'autres procédés de construction d'estimateurs sont envisageables, en fonction

du modèle statistique étudié.

Exemples

1. Dans le modèle $(\mathbb{R}_+^n, \{\mathscr{E}(\lambda)^{\otimes n}\}_{\lambda>0})$, le paramètre λ représente l'inverse de la moyenne de la loi $\mathscr{E}(\lambda)$. Un estimateur naturel de λ, construit avec l'échantillon $(X_1, \cdots, X_n)$ de la loi $\mathscr{E}(\lambda)^{\otimes n}$ est donc

$$\hat{\lambda} = \frac{1}{\bar{X}_n}.$$

2. Dans le modèle $(\mathbb{R}_+^n, \{\mathscr{U}([0,\theta])^{\otimes n}\}_{\theta>0})$, θ représente le maximum des valeurs prises par une réalisation de la loi $\mathscr{U}([0,\theta])$. L'estimateur naturel construit avec l'échantillon $(X_1, \cdots, X_n)$ de la loi $\mathscr{U}([0,\theta])^{\otimes n}$ est donc

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

Un autre estimateur, construit cette fois avec la mesure empirique est, par exemple,

$$\hat{\theta} = \frac{1}{2}\bar{X}_n.$$

Chapitre 2

Principes de l'inférence statistique

On s'intéresse ici à des critères de performance des estimateurs, posant ainsi les bases de l'inférence statistique.

Le modèle statistique considéré est $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, avec $\mathscr{H} \subset \mathbb{R}^k$ et $\Theta \subset \mathbb{R}^d$. Rappelons que, pour simplifier les écritures, on suppose que le paramètre d'intérêt, i.e. le paramètre que l'on souhaite estimer avec les observations, est θ. Dans ce qui suit, toutes les définitions et les résultats généraux s'étendent au cas où le paramètre d'intérêt est une fonction $g(\theta)$ de θ.

On désignera par $\mathbb{E}_\theta$ la moyenne sous la loi P_θ : sous la propriété d'intégrabilité adéquate,

$$\mathbb{E}_\theta g(.) = \mathbb{E}_\theta g(X_1, \cdots, X_n) = \int_{\mathscr{H}^n} g(x) P_\theta(\mathrm{d}x),$$

pour $g : \mathscr{H}^n \to \mathbb{R}$ et $(X_1, \cdots, X_n)$ un échantillon de loi P_θ.

2.1 Critères de performance en moyenne

La première propriété que l'on puisse exiger d'un estimateur est qu'il se comporte en moyenne comme son paramètre cible. C'est le concept de biais, décrit ci-dessous.

Dorénavant, on dira qu'une statistique $\hat{\theta}$ est d'*ordre* p si $\hat{\theta} \in L^p(P_\theta)$ pour chaque $\theta \in \Theta$.

Définitions *Soit $\hat{\theta}$ un estimateur d'ordre 1.*

1. *Le biais de $\hat{\theta}$ en θ est $\mathbb{E}_\theta \hat{\theta} - \theta$;*
2. *$\hat{\theta}$ est sans biais lorsque son biais est nul en chaque $\theta \in \Theta$.*
3. *$\hat{\theta}$ est asymptotiquement sans biais si pour chaque $\theta \in \Theta$, $\lim_{n\to\infty} \mathbb{E}_\theta \hat{\theta} = \theta$.*

Pour revenir à l'exemple de la section 1.1, lorsque $(X_1, \cdots, X_n)$ est un échantillon de la loi $\mathscr{B}(p)^{\otimes n}$, les 2 estimateurs X_1 et $\bar{X}_n$ sont sans biais.

La proximité entre l'estimateur et sa cible peut être évaluée grâce à la distance L^2 entre les 2 quantités. Dans ce qui suit, $\|.\|$ désigne la norme euclidienne de $\mathbb{R}^d$.

Définitions *Soit $\hat{\theta}$ un estimateur d'ordre 2.*

1. *Le risque quadratique de $\hat{\theta}$ sous P_θ est*
$$\mathscr{R}(\theta; \hat{\theta}) = \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2.$$
2. *Soit $\hat{\theta}'$ un autre estimateur d'ordre 2. On dit que $\hat{\theta}$ est préférable à $\hat{\theta}'$ lorsque pour chaque $\theta \in \Theta$, $\mathscr{R}(\theta; \hat{\theta}) \leq \mathscr{R}(\theta; \hat{\theta}')$.*
3. *Supposons que $\hat{\theta}$ est sans biais. On dit que $\hat{\theta}$ est de variance uniformément minimum parmi les estimateurs sans biais (VUMSB) si il est préférable à tout autre estimateur sans biais d'ordre 2.*

L'existence d'un estimateur VUMSB n'est en général pas acquise. Nous reviendrons sur ce problème dans la partie 4.3.

Dans la section 1.1, on a remarqué que lorsque $(X_1, \cdots, X_n)$ est un échantillon de la loi $\mathscr{B}(p)^{\otimes n}$, $\mathscr{R}(p; \bar{X}_n) = p(1-p)/n$. Ainsi, à mesure que l'on acquiert de l'information en multipliant les expériences aléatoires, l'estimateur $\bar{X}_n$ gagne en précision. Ce n'est pas le cas pour l'estimateur X_1, dont le risque quadratique vaut $\mathscr{R}(p; X_1) = p(1-p)$. Comme on pouvait s'y attendre, $\bar{X}_n$ est donc préférable à X_1. En fait, $\bar{X}_n$ est VUMSB. Pour le montrer, considérons un estimateur sans biais quelconque $\hat{\varphi} := \hat{\varphi}(X_1, \cdots, X_n)$, et notons :

$$\begin{aligned} L(p; X_1, \cdots, X_n) &= p^{n\bar{X}_n}(1-p)^{n-n\bar{X}_n}, \text{ et} \\ K(p) &= \ln L(p; X_1, \cdots, X_n). \end{aligned}$$

On remarque dans un premier temps que :

$$\mathbb{E}_p K'(p) = \mathbb{E}_p \left(\frac{1}{p} n\bar{X}_n - \frac{1}{1-p}(n - n\bar{X}_n) \right) = 0.$$

Par suite, si var_p et cov_p désignent la variance et la covariance sous la loi $\mathscr{B}(p)^{\otimes n}$:

$$\begin{aligned}\mathrm{cov}_p(\hat{\varphi}, K'(p)) &= \mathbb{E}_p \hat{\varphi} K'(p) = \sum_{i_1,\cdots,i_n \in \{0,1\}} \hat{\varphi}(i_1,\cdots i_n) L'(p; i_1,\cdots,i_n) \\ &= \frac{\mathrm{d}}{\mathrm{d}p}\mathbb{E}_p \hat{\varphi} = 1,\end{aligned}$$

car $\hat{\varphi}$ est sans biais. Comme, d'après l'inégalité de Cauchy-Schwarz,

$$\left(\mathrm{cov}_p(\hat{\varphi}, K'(p))\right)^2 \leq \mathrm{var}_p(\hat{\varphi})\mathrm{var}_p(K'(p)),$$

on a donc

$$\mathrm{var}_p(\hat{\varphi}) \geq \frac{1}{\mathrm{var}_p(K'(p))}.$$

Or,

$$\begin{aligned}\mathrm{var}_p(K'(p)) &= \mathrm{var}_p\left(\frac{1}{p} n\bar{X}_n + \frac{1}{1-p} n\bar{X}_n\right) = \frac{n^2}{p^2(1-p)^2}\mathrm{var}_p\left(\bar{X}_n\right) \\ &= \frac{n}{p(1-p)} = \left(\mathscr{R}(p;\bar{X}_n)\right)^{-1}. \qquad (2.1.1)\end{aligned}$$

On a donc obtenu

$$\mathscr{R}(p;\hat{\varphi}) = \mathrm{var}_p(\hat{\varphi}) \geq \mathscr{R}(p;\bar{X}_n),$$

c'est-à-dire que $\bar{X}_n$ est VUMSB. Cette preuve, qui peut sembler ici miraculeuse, sera formalisée dans les sections 3.3 et 4.1.

Exercice [CAS OÙ LE PARAMÈTRE D'INTÉRÊT EST UNE FONCTION DE θ] Soit le modèle statistique $(\mathbb{R}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$ tel que pour chaque $\theta \in \Theta$, Q_θ admet un moment d'ordre 2. Pour un échantillon $(X_1,\cdots,X_n)$ de loi $Q_\theta^{\otimes n}$, on note :

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

Montrer que $\bar{X}_n$ et S_n^2 sont des estimateurs sans biais de la moyenne et de la variance de la loi Q_θ, respectivement.

On note dorénavant, pour deux vecteurs aléatoires X, Y de carrés intégrables et à valeurs dans $\mathbb{R}^d$:

$$\begin{aligned}&K_\theta(X,Y) = \mathbb{E}_\theta(X - \mathbb{E}_\theta X)^T(Y - \mathbb{E}_\theta Y) = \mathbb{E}_\theta X^T Y - \mathbb{E}_\theta X^T \mathbb{E}_\theta Y \text{ et} \\ &V_\theta(X) = K_\theta(X,X) = \mathbb{E}_\theta \|X - \mathbb{E}_\theta X\|^2.\end{aligned}$$

Noter que $K_\theta(X,Y) = K_\theta(Y,X)$. Par ailleurs, K_θ et V_θ ne représentent pas la covariance et la variance sous la loi P_θ (respectivement notées cov_θ et var_θ), sauf lorsque $d = 1$.

Proposition [DÉCOMPOSITION BIAIS-VARIANCE] *Soit $\hat{\theta}$ un estimateur d'ordre 2. On a alors la décomposition :*

$$\mathscr{R}(\theta;\hat{\theta}) = \|\mathbb{E}_\theta\hat{\theta} - \theta\|^2 + V_\theta(\hat{\theta}).$$

Pour un risque donné, abaisser le biais revient donc à augmenter la variation, et réciproquement.

Preuve On a :

$$\begin{aligned}\mathscr{R}(\theta;\hat{\theta}) &= \mathbb{E}_\theta\|(\hat{\theta} - \mathbb{E}_\theta\hat{\theta}) + (\mathbb{E}_\theta\hat{\theta} - \theta)\|^2 \\ &= \mathbb{E}_\theta\|\hat{\theta} - \mathbb{E}_\theta\hat{\theta}\|^2 + \|\mathbb{E}_\theta\hat{\theta} - \theta\|^2 + 2\mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^T(\mathbb{E}_\theta\hat{\theta} - \theta).\end{aligned}$$

Le résultat en découle, car $\mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\hat{\theta}) = 0$ et $V_\theta(\hat{\theta}) = \mathbb{E}_\theta\|\hat{\theta} - \mathbb{E}_\theta\hat{\theta}\|^2$. □

Proposition *Soit $\hat{\theta}$ un estimateur d'ordre 2. Alors, $\hat{\theta}$ est VUMSB si, et seulement si, pour tout estimateur $\hat{\varphi}$ d'ordre 2 tel que $\mathbb{E}_\theta\hat{\varphi} = 0$ pour chaque $\theta \in \Theta$, on a :*

$$K_\theta(\hat{\varphi},\hat{\theta}) = 0, \quad \forall\theta \in \Theta.$$

Preuve Pour toute la preuve, fixons $\theta \in \Theta$. Supposons que $\hat{\theta}$ est VUMSB. Soit $\hat{\varphi}$ une statistique d'ordre 2 telle que $\mathbb{E}_\theta\hat{\varphi} = 0$. Pour tout $\alpha \in \mathbb{R}$, l'estimateur $\hat{\varphi}_\alpha = \hat{\theta} + \alpha\hat{\varphi}$ est sans biais. Comme $\hat{\theta}$ est sans biais et VUMSB, on a alors :

$$V_\theta(\hat{\theta}) = \mathscr{R}(\theta;\hat{\theta}) \leq \mathscr{R}(\theta;\hat{\varphi}_\alpha) = V_\theta(\hat{\varphi}_\alpha) = V_\theta(\hat{\theta}) + 2\alpha K_\theta(\hat{\theta},\hat{\varphi}) + \alpha^2 V_\theta(\hat{\varphi}).$$

Par suite, on a pour tout $\alpha \in \mathbb{R}$:

$$2\alpha K_\theta(\hat{\theta},\hat{\varphi}) + \alpha^2 V_\theta(\hat{\varphi}) \geq 0.$$

Ce polynôme en α ne peut garder un signe positif que si $K_\theta(\hat{\theta},\hat{\varphi}) = 0$.

Réciproquement, tout estimateur sans biais $\hat{\psi}$ tel que $\hat{\psi} \in L^2(P_\theta)$ s'écrit $\hat{\psi} = \hat{\theta} - \hat{\varphi}$, où $\hat{\varphi} = \hat{\theta} - \hat{\psi}$ est une statistique telle que $\mathbb{E}_\theta \hat{\varphi} = 0$ et $\hat{\varphi} \in L^2(P_\theta)$. Par hypothèse, on a alors $K_\theta(\hat{\theta}, \hat{\varphi}) = 0$ et la statistique $\hat{\psi}$ vérifie donc :

$$\begin{aligned} \mathscr{R}(\theta; \hat{\psi}) &= V_\theta(\hat{\theta} - \hat{\varphi}) = V_\theta(\hat{\theta}) + V_\theta(\hat{\varphi}) - 2K_\theta(\hat{\theta}, \hat{\varphi}) \\ &= V_\theta(\hat{\theta}) + V_\theta(\hat{\varphi}) \geq V_\theta(\hat{\theta}) = \mathscr{R}(\theta; \hat{\theta}), \end{aligned}$$

ce qui montre que $\hat{\theta}$ est VUMSB. □

Théorème *Soient $\hat{\theta}$ et $\hat{\theta}'$ des estimateurs VUMSB. Alors, pour chaque $\theta \in \Theta$, $\hat{\theta} = \hat{\theta}'$ P_θ-p.s.*

Preuve Fixons $\theta \in \Theta$. Comme la statistique $\hat{\varphi} = \hat{\theta} - \hat{\theta}'$ vérifie les hypothèses du théorème précédent, on a :

$$\begin{aligned} V_\theta(\hat{\theta} - \hat{\theta}') &= \mathbb{E}_\theta(\hat{\theta} - \hat{\theta}')^T(\hat{\theta} - \hat{\theta}') \\ &= \mathbb{E}_\theta(\hat{\theta} - \hat{\theta}')^T(\hat{\theta} - \theta) - \mathbb{E}_\theta(\hat{\theta} - \hat{\theta}')^T(\hat{\theta}' - \theta) \\ &= K_\theta(\hat{\theta} - \hat{\theta}', \hat{\theta}) - K_\theta(\hat{\theta} - \hat{\theta}', \hat{\theta}') = 0, \end{aligned}$$

ce qui montre que $\hat{\theta} = \hat{\theta}'$ P_θ-p.s., car $\hat{\theta}$ et $\hat{\theta}'$ sont sans biais. □

2.2 Critères de performance asymptotique

A mesure que la taille n de l'échantillon croît, l'échantillon contient de plus en plus d'informations sur la vraie valeur du paramètre. On est alors amené à s'intéresser aux propriétés asymptotiques des estimateurs. Dans la suite, sauf mention explicite du contraire, toute propriété de convergence sera entendue pour une taille d'échantillon n qui tend vers l'infini.

Définition *On dit que l'estimateur $\hat{\theta}$ est consistant lorsque pour chaque $\theta \in \Theta$, $\hat{\theta} \xrightarrow{P_\theta} \theta$.*

Dans l'exemple de la section 1.1, l'estimateur $\bar{X}_n$ construit avec un échantillon $(X_1, \cdots, X_n)$ de loi $\mathscr{B}(p)^{\otimes n}$ est consistant, car $\bar{X}_n \xrightarrow{\mathscr{B}(p)^{\otimes n}} p$ pour chaque $p \in]0,1[$.

Remarque Un estimateur peut être asymptotiquement sans biais sans être consistant. De même, un estimateur peut être consistant sans être asymptotiquement

sans biais. Pour se convaincre du second point par exemple, considérons le modèle statistique $(\mathbb{R}^n, \{N(m,1)^{\otimes n}\}_{m\in]0,1[})$, et l'estimateur $\hat{m}$ issu de l'échantillon $(X_1,\cdots,X_n)$ de la loi $N(m,1)^{\otimes n}$, pour $m \in]0,1[$:

$$\hat{m} = \bar{X}_n + \frac{1}{F(-\sqrt{n})}\mathbf{1}_{\{\bar{X}_n \leq 0\}},$$

où F désigne la fonction de répartition de la loi $N(0,1)$. Comme $m > 0$, la loi faible des grands nombres montre que $\hat{m} \xrightarrow{P_m} m$, si $P_m = N(m,1)^{\otimes n}$. Par ailleurs, comme $\bar{X}_n \sim N(m,1/n)$:

$$P_m(\bar{X}_n \leq 0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-m\sqrt{n}} e^{-t^2/2}\mathrm{d}t \geq F(-\sqrt{n}),$$

car $m \leq 1$. Donc $\mathbb{E}_m\hat{m} \geq m+1$, et $\hat{m}$ n'est pas asymptotiquement sans biais.

Exercice [CAS OÙ LE PARAMÈTRE D'INTÉRÊT EST UNE FONCTION DE θ] Soit le modèle statistique $(\mathbb{R}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$ tel que pour chaque $\theta \in \Theta$, Q_θ admet un moment d'ordre 2. Pour un échantillon $(X_1,\cdots,X_n)$ de loi $Q_\theta^{\otimes n}$, on note :

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

Montrer que $\hat{\sigma}_n^2$ est un estimateur biaisé de la variance de Q_θ, mais qu'il est asymptotiquement sans biais et consistant.

Cette propriété ne doit être vue que comme une propriété minimale que doit satisfaire un estimateur raisonnablement constitué. Cependant, elle ne permet pas de préciser l'erreur commise. C'est précisémment l'objet de la définition qui suit.

Définitions *Soit $(v_n)_n$ une suite de réels positifs telle que $v_n \to \infty$. On dit que l'estimateur $\hat{\theta}$ est :*

1. *de vitesse $(v_n)_n$ si, pour chaque $\theta \in \Theta$, il existe une loi $\ell(\theta)$ telle que $v_n(\hat{\theta}-\theta) \xrightarrow{\mathscr{L}/P_\theta} \ell(\theta)$.*
2. *asymptotiquement normal si, en outre, les lois $\ell(\theta)$ sont gaussiennes.*

La performance d'un estimateur est notamment évaluée sur sa vitesse car, pour une précision donnée, plus la vitesse est rapide, moins la taille de l'échantillon

doit être importante. Néanmoins, il ne faut pas oublier qu'un estimateur performant doit aussi pouvoir être calculé via un algorithme de complexité raisonnable. Comme, en principe, ces 2 contraintes s'opposent, il est important de savoir réaliser un compromis entre ces exigences.

Remarque Un estimateur qui possède la propriété **1.** de la définition ci-dessus est consistant. En effet, fixons $\theta \in \Theta$. On suppose pour simplifier que $(v_n)_n$ est croissante, et que $\ell(\theta)$ est une loi sans atomes (sinon, il suffit de travailler sur l'ensemble des points de continuité de la fonction de répartition de la loi de $\|\ell(\theta)\|$; à toutes fins utiles, rappelons que l'ensemble des points de discontinuité d'une v.a.r. est au plus dénombrable). Pour chaque $\varepsilon > 0$, on a

$$P_\theta(\|\hat{\theta} - \theta\| \geq \varepsilon) \leq P_\theta(v_n\|\hat{\theta} - \theta\| \geq v_p\varepsilon),$$

pour tout $p \leq n$. On en déduit que pour tout p,

$$\limsup_{n\to\infty} P_\theta(\|\hat{\theta} - \theta\| \geq \varepsilon) \leq P_\theta(\|\ell(\theta)\| \geq v_p\varepsilon).$$

En faisant finalement tendre p vers $+\infty$, on peut conclure que $\hat{\theta} \xrightarrow{P_\theta} \theta$.

Dans l'exemple de la section 1.1, on a vu que l'estimateur $\bar{X}_n$ construit avec un échantillon $(X_1, \cdots, X_n)$ de loi $\mathscr{B}(p)^{\otimes n}$ est asymptotiquement normal, de vitesse $\sqrt{n}$, car pour chaque $p \in [0,1]$,

$$\sqrt{n}\,(\bar{X}_n - p) \xrightarrow{\mathscr{L}/\mathscr{B}(p)^{\otimes n}} N(0, p(1-p)).$$

Exercice Soit le modèle statistique $(\mathbb{R}^n, \{\mathscr{U}([\theta, \theta+1])^{\otimes n}\}_{\theta\in[0,1]})$. Construire et étudier des estimateurs du paramètre θ, en utilisant les statistiques $\min_{i\leq n} X_i$, $\max_{i\leq n} X_i$ et $\bar{X}_n$ issues d'un échantillon $(X_1, \cdots, X_n)$ de la loi $\mathscr{U}([\theta, \theta+1])^{\otimes n}$.

2.3 Intervalles de confiance

Nous avons déjà vu, dans la section 1.1, un exemple de construction d'intervalle de confiance. L'objectif de cette section est d'en rappeler le principe, sans toutefois rentrer dans un formalisme excessif, qui pourrait être préjudiciable à la compréhension de la démarche.

Dans cette section, le modèle statistique est $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, avec $\Theta \subset \mathbb{R}$. L'observation $(x_1, \cdots, x_n) \in \mathscr{H}^n$ est issue d'une loi P_{θ_0}, avec $\theta_0 \in \Theta$ inconnu.

2.3.1 Intervalle de confiance pour une taille d'échantillon finie

On fixe $\alpha \in]0,1[$.

Définition *Soit T_n une fonction définie sur $\mathscr{H}^n$ et à valeurs dans les intervalles de $\mathbb{R}$ telle que pour chaque $\theta \in \Theta$:*

$$P_\theta\left(\theta \in T_n(.)\right) = (resp.\ \geq)\, 1-\alpha.$$

$T_n(x_1,\cdots,x_n)$ s'appelle intervalle de confiance (resp. par excès) pour θ_0, au niveau de confiance $1-\alpha$.

Ainsi, $\theta_0 \in T_n(x_1,\cdots,x_n)$ avec une P_{θ_0}-probabilité (*resp. au moins*) $1-\alpha$. On peut remarquer d'emblée qu'un intervalle de confiance est d'autant plus intéressant qu'il est de longueur faible, pour un niveau de confiance élevé. Comme ces 2 exigences s'opposent, il est impératif de réaliser un compromis.

Exemple Considérons le cas d'un modèle statistique $\{P_\theta\}_{\theta\in\Theta} = \{Q_\theta^{\otimes n}\}_{\theta\in\Theta}$ pour lequel $\int_{\mathscr{H}} x^2 Q_\theta(\mathrm{d}x) \leq 1$ et $\theta = \int_{\mathscr{H}} x Q_\theta(\mathrm{d}x)$ pour tout $\theta \in \Theta$. Soit $(X_1,\cdots,X_n)$ un échantillon de la loi $Q_\theta^{\otimes n}$. D'après l'inégalité de Bienaymé-Tchebytchev :

$$P_\theta\left(|\bar{X}_n - \theta| > t\right) \leq \frac{\mathrm{var}_\theta(\bar{X}_n)}{t^2} = \frac{\mathrm{var}_\theta(X_1)}{nt^2} \leq \frac{1}{nt^2},\ \forall t > 0.$$

Si t vérifie $(nt^2)^{-1} \leq \alpha$, on a donc

$$P_\theta\left(|\bar{X}_n - \theta| > t\right) \leq \alpha.$$

Pour un tel t, $[\bar{x}_n - t, \bar{x}_n + t]$ est donc un intervalle de confiance par excès pour θ_0, au niveau de confiance $1-\alpha$. On peut trouver des intervalles de confiance plus précis en utilisant, au lieu de l'inégalité de Bienaymé-Tchebytchev, une inégalité exponentielle (inégalité de Bernstein, inégalité de Hoeffding, ...), forcément plus précise.

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

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

$$q_r = \inf\{x \in \mathbb{R} : F(x) \geq r\}.$$

Si F est continue, $F(q_r) = r$. Si, de plus, F est strictement croissante, alors q_r est l'unique solution de l'équation $F(.) = r$.

Preuve Il suffit de remarquer que, comme F est croissante et continue à droite, $F(q_r^-) \leq r \leq F(q_r)$, si $F(q_r^-)$ est la limite à gauche de F en q_r. □

Exemple Considérons le modèle statistique $\{N(m,1)^{\otimes n}\}_{m\in\mathbb{R}}$. Pour $(X_1,\cdots,X_n)$ un échantillon de la loi $P_m = N(m,1)^{\otimes n}$, on a $\sqrt{n}(\bar{X}_n - m) \sim N(0,1)$. Soit t_0 le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$: si Φ est la fonction de répartition de la loi $N(0,1)$, on sait que $\Phi(t_0) = 1-\alpha/2$. Comme la loi $N(0,1)$ possède une densité paire :

$$P_m\left(\sqrt{n}\,|\bar{X}_n - m| \leq t_0\right) = 2\Phi(t_0) - 1 = 1-\alpha.$$

Si les observations $x_1,\cdots,x_n$ sont régies par la loi $N(m_0,1)$, $[\bar{x}_n - t_0/\sqrt{n}, \bar{x}_n + t_0/\sqrt{n}]$ est un intervalle de confiance pour m_0, au niveau $1-\alpha$.

Si l'obtention d'une telle propriété est hors d'atteinte, ou si T_n est trop complexe pour pouvoir être utilisé, on se retranche sur une propriété asymptotique.

2.3.2 Intervalle de confiance asymptotique

Soit $\alpha \in]0,1[$.

Définition *Soit, pour chaque n, T_n une fonction définie sur $\mathscr{H}^n$ et à valeurs dans les intervalles de $\mathbb{R}$ telle que pour chaque $\theta \in \Theta$:*

$$P_\theta\left(\theta \in T_n(.)\right) \longrightarrow 1-\alpha.$$

$T_n(x_1,\cdots x_n)$ s'appelle intervalle de confiance asymptotique pour θ_0 au niveau de confiance $1-\alpha$.

Exemple Supposons par exemple que $\hat{\theta}$ est un estimateur asymptotiquement normal, de vitesse $(v_n)_n$: pour chaque $\theta \in \Theta$,

$$v_n(\hat{\theta} - \theta) \overset{\mathscr{L}/P_\theta}{\longrightarrow} N(0,1). \tag{2.3.1}$$

Notons $q_{1-\alpha/2}$ et $q_{\alpha/2}$ les quantiles d'ordre $1-\alpha/2$ et $\alpha/2$ de la loi $N(0,1)$. Par symétrie de la loi $N(0,1)$, $q_{1-\alpha/2} = -q_{\alpha/2}$. Si $q = q_{1-\alpha/2} > 0$, alors :

$$P_\theta\left(-q \leq v_n(\hat{\theta} - \theta) \leq q\right) \longrightarrow 1-\alpha.$$

L'intervalle de confiance asymptotique au niveau $1-\alpha$ est donc :

$$\left[\hat{\theta}(x_1,\cdots,x_n)-\frac{q}{v_n};\hat{\theta}(x_1,\cdots,x_n)+\frac{q}{v_n}\right].$$

Pour la construction des intervalles de confiance asymptotiques, le lemme de Slutsky (au programme du L3) est souvent utile.

Lemme [SLUTSKY] *Soient $(X_n)_n$ et $(Y_n)_n$ des suites de v.a.r. sur $(\Omega,\mathscr{F},\mathbb{P})$. Si il existe une v.a.r. Y et un réel a tels que $X_n \xrightarrow{\mathbb{P}} a$ et $Y_n \xrightarrow{\mathscr{L}/\mathbb{P}} Y$, alors $(X_n,Y_n) \xrightarrow{\mathscr{L}/\mathbb{P}} (X,Y)$. En particulier, $X_nY_n \xrightarrow{\mathscr{L}/\mathbb{P}} aY$ et $X_n+Y_n \xrightarrow{\mathscr{L}/\mathbb{P}} a+Y$.*

Exemple Supposons à nouveau que $\hat{\theta}$ est un estimateur asymptotiquement normal, de vitesse $(v_n)_n$: pour chaque $\theta\in\Theta$, il existe $\sigma_\theta^2>0$ tel que

$$v_n(\hat{\theta}-\theta)\xrightarrow{\mathscr{L}/P_\theta} N(0,\sigma_\theta^2). \tag{2.3.2}$$

Soit $\hat{\sigma}$ un estimateur consistant de σ_θ. On a recours au lemme de Slutsky pour en déduire de (2.3.2) que pour chaque $\theta\in\Theta$:

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

Par suite, avec les notations de l'exemple précédent :

$$P_\theta\left(-q\le v_n\frac{\hat{\theta}-\theta}{\hat{\sigma}}\le q\right)\longrightarrow 1-\alpha,$$

ou bien, avec une écriture équivalente :

$$P_\theta\left(\theta\in\left[\hat{\theta}-\frac{\hat{\sigma}q}{v_n};\hat{\theta}+\frac{\hat{\sigma}q}{v_n}\right]\right)\longrightarrow 1-\alpha.$$

Comme les quantités $\hat{\theta}$ et $\hat{\sigma}$ qui interviennent dans cet intervalle peuvent être calculées pour les observations $x_1,\cdots,x_n$, cette propriété nous donne l'intervalle de confiance asymptotique recherché.

La δ-*méthode* est aussi souvent utilisée pour la construction d'intervalle de confiance asymptotiques.

Lemme [δ-MÉTHODE] *Soit $(X_n)_n$ une suite de v.a.r. sur $(\Omega,\mathscr{F},\mathbb{P})$, $g:\mathbb{R}\to\mathbb{R}$ une fonction continûment dérivable en x_0 et $(v_n)_n$ une suite de réels tendant vers $+\infty$. Si $v_n(X_n-x_0)\xrightarrow{\mathscr{L}/\mathbb{P}} X$, alors $v_n(g(X_n)-g(x_0))\xrightarrow{\mathscr{L}/\mathbb{P}} g'(x_0)X$.*

Preuve D'après la formule de Taylor-Lagrange, il existe ξ_n compris entre x_0 et X_n tel que

$$g(X_n)=g(x_0)+(X_n-x_0)g'(\xi_n).$$

Comme g' est continue en x_0 et $(X_n)_n$ converge en probabilité vers x_0, on a donc

$$v_n(g(X_n)-g(x_0))=v_n(X_n-x_0)g'(\xi_n)\xrightarrow{\mathscr{L}/\mathbb{P}} g'(x_0)X,$$

d'après le lemme de Slutsky. □

Exemple Supposons que l'on veuille construire un intervalle de confiance asymptotique au niveau $1-\alpha$ pour le paramètre λ, dans le modèle $\{\mathscr{E}(\lambda)^{\otimes n}\}_{\lambda>0}$. Soit $(X_1,\cdots,X_n)$ un échantillon de la loi $\mathscr{E}(\lambda)^{\otimes n}$. D'après le théorème de la limite centrale :

$$\sqrt{n}\left(\bar{X}_n-\frac{1}{\lambda}\right)\xrightarrow{\mathscr{L}/\mathscr{E}(\lambda)^{\otimes n}} N(0,1/\lambda^2).$$

On a recours à la δ-méthode pour en déduire que

$$\sqrt{n}\left(\frac{1}{\bar{X}_n}-\lambda\right)\xrightarrow{\mathscr{L}/\mathscr{E}(\lambda)^{\otimes n}} \frac{1}{\lambda^2}N(0,1/\lambda^2)=\frac{1}{\lambda^3}N(0,1).$$

Finalement, en utilisant l'estimateur consistant $1/\bar{X}_n$, le lemme de Slutsky nous donne

$$\bar{X}_n^{-3}\sqrt{n}\left(\frac{1}{\bar{X}_n}-\lambda\right)\xrightarrow{\mathscr{L}/\mathscr{E}(\lambda)^{\otimes n}} N(0,1).$$

L'intervalle de confiance asymptotique s'en déduit facilement.

Chapitre 3

Vraisemblance

La méthode de construction des estimateurs par maximisation de la vraisemblance est sans doute la plus répandue. Le principe de la construction est intuitivement évident : il s'agit de choisir comme estimateur le paramètre pour lequel l'observation est la plus probable, ou la plus vraisemblable ...

Dans tout le chapitre, l'espace des observations individuelles est $\mathscr{H} \subset \mathbb{R}^k$, et l'espace des paramètres est $\Theta \subset \mathbb{R}^d$.

3.1 Le concept de vraisemblance

Définition *On appelle vraisemblance du modèle statistique* $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ *dominé par* μ *toute application* $L : \mathscr{H}^n \times \Theta \to \mathbb{R}_+$ *telle que, pour chaque* $\theta \in \Theta$*, l'application partielle* $L(.;\theta) : \mathscr{H}^n \to \mathbb{R}_+$ *soit un élément de la classe d'équivalence de la densité de* P_θ *par rapport à* μ.

Remarque La vraisemblance, dont l'existence est acquise grâce au théorème de Radon-Nikodym, dépend donc du choix de la mesure dominante du modèle, qui n'est pas unique. De plus, en raison du fait que que chaque densité $\mathrm{d}P_\theta/\mathrm{d}\mu$ n'est unique qu'à une équivalence près, une vraisemblance elle-même n'est pas unique. Malgré cela, nous parlerons de "la" vraisemblance, sachant que, dans la pratique, le choix d'une vraisemblance s'impose souvent par ses propriétés analytiques.

Exemples

1. Dans le modèle statistique $(\{0,1\}^n, \{\mathscr{B}(p)^{\otimes n}\}_{p\in]0,1[})$ de la section 1.1, qui

est dominé par la mesure $(\delta_0+\delta_1)^{\otimes n}$, la vraisemblance L s'exprime par :

$$L(x_1,\cdots,x_n;p)=\mathscr{B}(p)^{\otimes n}(\{x_1,\cdots,x_n\})=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

pour $p\in]0,1[$ et $x_1,\cdots,x_n\in\{0,1\}$.

2. Dans le modèle $(\mathbb{R}^n,\{N(m,\sigma^2)^{\otimes n}\}_{m\in\mathbb{R},\sigma\in\mathbb{R}_+^\star})$, qui est dominé par la mesure de Lebesgue sur $\mathbb{R}^n$, la vraisemblance est :

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

pour $x_i\in\mathbb{R}$, $m\in\mathbb{R}$ et $\sigma\in\mathbb{R}_+^\star$.

Dans le cadre de modèles statistiques issus d'observations indépendantes, l'expression naturelle de la vraisemblance est simple, comme le montre la proposition ci-dessous.

Proposition *Soit $(\mathscr{H},\{Q_\theta\}_{\theta\in\Theta})$ un modèle statistique dominé par la mesure μ, et de vraisemblance L. Alors, la fonction*

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

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

Preuve Il suffit de remarquer que, pour chaque $\theta\in\Theta$,

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

est une version de la densité de $Q_\theta^{\otimes n}$ par rapport à $\mu^{\otimes n}$. □

Reprenons l'exemple de la section 1.1. Les lancers de la pièce ont fourni une suite d'observations $x_1,\cdots,x_n\in\{0,1\}$. Il est naturel de considérer que la loi $\mathscr{B}(p_0)$ qui régit ces observations est la loi qui apporte la plus forte probabilité à cette réalisation. C'est ainsi que, pour donner une valeur approchée de la vraie valeur du paramètre, on est amené à maximiser en p la vraisemblance $L(x_1,\cdots,x_n;p)$: l'idée sous-jacente est que la valeur de p obtenue est celle qui

s'ajuste le mieux aux observations. C'est cette observation qui motive le concept de maximum de vraisemblance.

Définition *Soit* $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ *un modèle statistique dominé, et L la vraisemblance associée. Un estimateur du maximum de vraisemblance (EMV) est une statistique g à valeurs dans* Θ *qui vérifie :*

$$L(x; g(x)) = \sup_{\theta\in\Theta} L(x;\theta),\ \forall x \in \mathscr{H}^n.$$

Ainsi, si $(X_1, \cdots, X_n)$ *est un échantillon de la loi* P_θ*, l'EMV (de* θ*) est* $g(X_1, \cdots, X_n)$.

Bien entendu, ni l'existence, ni l'unicité des EMV ne sont en général pas acquises.

Dans le modèle statistique issu d'observations indépendantes de la proposition précédente, on préfère calculer l'EMV en maximisant la "log-vraisemblance" - c'est-à-dire le logarithme de la vraisemblance- plutôt que la vraisemblance, car celle-ci s'exprime comme :

$$\ln L_n(x_1, \cdots, x_n; \theta) = \sum_{i=1}^{n} \ln L(x_i;\theta).$$

L'intérêt pratique est clair, l'étape de maximisation étant en principe plus facile à mener.

Exemple L'EMV du modèle statistique $(\mathbb{R}^n, \{N(m,1)^{\otimes n}\}_{m\in\mathbb{R}})$ est la moyenne empirique.

3.2 Consistance de l'EMV

L'un des outils de base pour l'étude des EMV est décrit ci-dessous :

Définition-Proposition *Soit* $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ *un modèle statistique identifiable et dominé par* μ*, de vraisemblance L. Pour chaque* $\alpha, \theta \in \Theta$*, on suppose que* $\ln L(.;\alpha) \in L^1(P_\theta)$. *On note :*

$$K(\alpha,\theta) = -\mathbb{E}_\theta \ln \frac{L(.;\alpha)}{L(.;\theta)}$$

l'information de Kullback entre les lois P_α et P_θ. Alors, $K(\alpha,\theta) \geq 0$ et de plus $K(\alpha,\theta) = 0 \Leftrightarrow \alpha = \theta$.

Preuve Tout d'abord, il est clair que $K(\theta,\theta) = 0$. Soient donc $\alpha \neq \theta$. Comme la fonction $t \mapsto -\ln t$ définie sur $\mathbb{R}_+^\star$ est convexe, on a avec l'inégalité de Jensen :

$$\begin{aligned} K(\alpha,\theta) &= -\int_{\mathscr{H}^n} \ln \frac{L(.;\alpha)}{L(.;\theta)} \mathrm{d}P_\theta \\ &\geq -\ln \int_{\mathscr{H}^n} \frac{L(.;\alpha)}{L(.;\theta)} \mathrm{d}P_\theta = -\ln \int_{\mathscr{H}^n} L(.;\alpha)\mathrm{d}\mu = 0. \end{aligned}$$

Supposons que $K(\alpha,\theta) = 0$. On est alors dans un cas d'égalité dans l'inégalité de Jensen. Comme $t \mapsto -\ln t$ définie sur $\mathbb{R}_+^\star$ est strictement convexe, on en déduit qu'il existe $C \in \mathbb{R}_+$ tel que $L(.;\alpha) = CL(.;\theta)$ P_θ-p.s. Or, P_α est absolument continue par rapport à P_θ, de densité $L(.;\alpha)/L(.;\theta)$. Par suite, pour tout borélien $A \subset \mathscr{H}^n$,

$$P_\alpha(A) = \int_A L(.;\alpha)\mathrm{d}\mu = \int_A \frac{L(.;\alpha)}{L(.;\theta)} \mathrm{d}P_\theta = CP_\theta(A).$$

On en déduit tout d'abord que $C = 1$ (prendre $A = \mathscr{H}^n$), puis que $P_\theta = P_\alpha$, ce qui contredit l'identifiabilité du modèle. □

Cette propriété de l'information de Kullback permet d'identifier le paramètre inconnu θ en tant que seule solution de l'équation $K(.,\theta) = 0$. C'est en ce sens que l'information de Kullback donne des informations sur le modèle.

A priori, il n'y a pas de raison pour qu'un EMV soit consistant, comme en atteste l'exemple suivant :

Exemple Soit $(\mathbb{R}^n, \{\mathscr{C}(\theta)^{\otimes n}\}_{\theta>0})$ un modèle statistique, où $\mathscr{C}(\theta)$ désigne la loi sur $\mathbb{R}$, de densité

$$\frac{\theta}{\pi}\frac{1}{\theta^2 + x^2},\ x \in \mathbb{R}.$$

Notons $(X_1,\cdots,X_n)$ un échantillon de la loi $\mathscr{C}(\theta)^{\otimes n}$, avec $\theta > 0$. Un simple calcul nous montre que l'EMV $\hat{\theta}$ est la seule solution de l'équation $\varphi_n(.) = 1/2$, où l'on a noté

$$\varphi_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + (X_i/\alpha)^2}, \quad \alpha > 0.$$

Par ailleurs, on vérifie facilement que pour tous $\alpha_1, \alpha_2 > 0$:

$$|\varphi_n(\alpha_1) - \varphi_n(\alpha_2)| \leq |\alpha_1^2 - \alpha_2^2| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1^2 + X_i^2}.$$

Par l'absurde, supposons que $\hat{\theta}$ est consistant. La loi des grands nombres et cette inégalité nous montrent que

$$\varphi_n(\hat{\theta}) \overset{\mathscr{C}(\theta)^{\otimes n}}{\longrightarrow} \mathbb{E}_\theta \frac{1}{1 + (X/\theta)^2}$$

pour chaque $\theta > 0$. Par suite,

$$\mathbb{E}_\theta \frac{1}{1 + (X/\theta)^2} = \frac{1}{2}, \ \forall \theta > 0,$$

ce qui est impossible car le terme de gauche tend vers 1 lorsque $\theta \to \infty$.

Il est donc nécessaire de donner des conditions suffisantes de consistance des EMV.

Théorème *Soit $(\mathscr{H}, \{Q_\theta\}_{\theta \in \Theta})$ un modèle statistique identifiable et dominé, de vraisemblance L. On suppose que Θ est compact, et que :*

(i) *$\forall x \in \mathscr{H}$, $\ln L(x;.)$ est continu sur Θ ;*

(ii) *$\forall \theta \in \Theta$, il existe $H \in L^1(Q_\theta)$ telle que $\sup_{\alpha \in \Theta} |\ln L(.;\alpha)| \leq H$.*

On note $\hat{\theta}$ l'EMV de θ associé à la vraisemblance

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

du modèle $(\mathscr{H}^n, \{Q_\theta^{\otimes n}\}_{\theta \in \Theta})$. Alors, $\hat{\theta}$ est consistant.

Preuve On fixe $\theta \in \Theta$ et on note $P_\theta = Q_\theta^{\otimes n}$. Soit $(X_1, \cdots, X_n)$ un échantillon de la loi P_θ et, pour chaque $\alpha \in \Theta$:

$$\begin{aligned} U_n(\alpha) &= -\frac{1}{n} \ln L_n(X_1, \cdots, X_n; \alpha) = -\frac{1}{n} \sum_{i=1}^{n} \ln L(X_i; \alpha) \\ U(\alpha) &= -\mathbb{E}_\theta \ln L(.;\alpha). \end{aligned}$$

Remarquons que $U_n(\hat{\theta}) = \inf_\Theta U_n$ et, par hypothèse, que U est continue. D'après la loi des grands nombres, $U_n \overset{P_\theta}{\longrightarrow} U$ ponctuellement ; nous allons tout d'abord

montrer que cette convergence est en fait uniforme. Pour tout $\eta > 0$, on désigne par $g(.,\eta)$ la fonction définie pour chaque $x \in \mathscr{H}^n$ par

$$g(x,\eta) = \sup_{\|\alpha-\beta\|\leq\eta} |\ln L(x;\alpha) - \ln L(x;\beta)|.$$

On fixe maintenant $\varepsilon > 0$. Comme $g(.,\eta) \leq 2H$ avec $H \in L^1(P_\theta)$ et $g(x,\eta) \to 0$ si $\eta \to 0$ pour tout $x \in \mathscr{H}^n$, on a $\mathbb{E}_\theta g(.,\eta) < \varepsilon/3$ d'après le théorème de Lebesgue, pour une certaine valeur de η que nous fixons dorénavant. On recouvre le compact Θ par N boules fermées de Θ de rayon η :

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

On a dans un premier temps :

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

On en déduit dans un second temps que, puisque $\mathbb{E}_\theta g(.,\eta) < \varepsilon/3$:

$$\begin{aligned}
P_\theta\left(\sup_\Theta |U_n - U| \geq \varepsilon\right) &\leq P_\theta\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i,\eta) + \max_{j=1,\cdots,N} |U_n(\theta_j) - U(\theta_j)| \geq 2\varepsilon/3\right) \\
&\leq P_\theta\left(\max_{j=1,\cdots,N} |U_n(\theta_j) - U(\theta_j)| \geq \varepsilon/3\right) \\
&\quad + P_\theta\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i,\eta) \geq \varepsilon/3\right).
\end{aligned}$$

Or, d'après la la loi des grands nombres, on a à la fois :

$$\max_{j=1,\cdots,N} |U_n(\theta_j) - U(\theta_j)| \xrightarrow{P_\theta} 0 \text{ et } \frac{1}{n}\sum_{i=1}^{n} g(X_i,\eta) \xrightarrow{P_\theta} \mathbb{E}_\theta g(.,\eta) < \varepsilon/3.$$

Ces observations nous permettent de déduire que $\sup_\Theta |U_n - U| \xrightarrow{P_\theta} 0$. En particulier,

$$U_n(\hat{\theta}) = \inf_\Theta U_n \xrightarrow{P_\theta} \inf_\Theta U. \tag{3.2.1}$$

Comme Θ est compact et U est continue, il existe $t \in \Theta$ tel que $U(t) = \inf_\Theta U$. Par suite :

$$U_n(\hat{\theta}) - U_n(\theta) \xrightarrow{P_\theta} U(t) - U(\theta) = K(t,\theta).$$

De plus,

$$U_n(\hat{\theta}) - U_n(\theta) = \inf_\Theta U_n - U_n(\theta) \leq 0.$$

On a donc $K(t,\theta) \leq 0$, ce qui montre que $K(t,\theta) = 0$ d'où $t = \theta$. D'après (3.2.1), $U_n(\hat{\theta}) \xrightarrow{P_\theta} U(\theta)$ et, puisque U_n converge uniformément vers U en probabilité, on en déduit que

$$K(\hat{\theta},\theta) = U(\hat{\theta}) - U(\theta) \xrightarrow{P_\theta} 0.$$

Soit $\varepsilon > 0$. Il existe $\gamma > 0$ tel que si $\alpha \in \Theta$ vérifie $\|\alpha - \theta\| \geq \varepsilon$, alors $K(\alpha,\theta) \geq \gamma$. Par conséquent,

$$P_\theta\left(\|\hat{\theta} - \theta\| \geq \varepsilon\right) \leq P_\theta\left(K(\hat{\theta},\theta) \geq \gamma\right) \longrightarrow 0,$$

donc $\hat{\theta}$ tend vers θ en probabilité. □

3.3 Information de Fisher

Dans le cadre d'un modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ de vraisemblance L telle que pour chaque $x \in \mathscr{H}^n$, $\ln L(x;.) \in \mathscr{C}^1$, la fonction *score* au point θ définie par

$$x \mapsto \nabla \ln L(x;\theta),$$

et dans laquelle ∇ désigne le gradient par rapport à θ, évalue la variabilité du modèle. C'est une notion intrinsèque au modèle, en ce sens qu'elle ne dépend ni de la mesure dominante, ni de la vraisemblance. C'est ce qui justifie la définition qui suit.

Par convention, dès que l'on parle de gradient (*resp.* hessienne), il est sous-entendu que la fonction est de classe $\mathscr{C}^1$ (*resp.* $\mathscr{C}^2$).

Définition *Soit $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ un modèle statistique dominé de vraisemblance L. On suppose que Θ est ouvert, et que pour chaque $\theta \in \Theta$: $\nabla \ln L(.;\theta) \in L^2(P_\theta)$.*

On appelle information de Fisher la fonction

$$I : \theta \mapsto \mathrm{var}_\theta\left(\nabla \ln L(.;\theta)\right) = \left(\mathrm{cov}_\theta\left(\frac{\partial}{\partial\theta_i}\ln L(.;\theta), \frac{\partial}{\partial\theta_j}\ln L(.;\theta)\right)\right)_{i,j=1,\cdots,d}.$$

Lorsque nous parlerons d'information de Fisher, il sera sous-entendu que les hypothèses imposées dans cette définition seront satisfaites.

L'information de Fisher est donc une fonction à valeurs dans l'ensemble des matrices semi-définies positives qui évalue le pouvoir de discrimination du modèle entre 2 valeurs proches du paramètre d'intérêt. En effet, on voit directement dans le cas $d = 1$ que $I(\theta)$ grand traduit une grande variation de la nature des probabilités du modèle au voisinage de P_θ, d'où une discrimination de la vraie valeur du paramètre inconnu facilitée. A l'inverse, si $I(\theta)$ est petit, la loi est très piquée : c'est mauvais, car on est amené à rechercher le maximum de la vraisemblance dans une région très vaste. Ce sont ces propriétés de $I(\theta)$ qui fournissent une information sur le modèle.

Pour illustrer ces affirmations, reprenons le modèle de la section 1.1, pour lequel la vraisemblance vaut, si $p \in]0,1[$ et $x_1,\cdots,x_n \in \{0,1\}$:

$$L(x_1,\cdots,x_n;p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

On a déjà vu dans la relation (2.1.1) que :

$$I(p) = \mathrm{var}_p\left(\nabla \ln L(.;p)\right) = \frac{n}{p(1-p)}.$$

Dans ce modèle, l'incertitude est faible pour p proche de 0 et 1 alors qu'elle est grande pour $p = 1/2$. Ceci se traduit bien par une information $I(p)$ maximale pour p proche de 0 et 1, et minimale pour $p = 1/2$.

Dans une situation d'échantillonage i.i.d., l'information de Fisher est proportionnelle à la taille de l'échantillon. Cette propriété, que nous montrons ci-dessous, légitime encore plus ce concept en tant que mesure d'une quantité d'information.

Proposition *Soit $(\mathscr{H}, \{Q_\theta\}_{\theta\in\Theta})$ un modèle statistique dominé d'information de Fisher I. Alors, l'information de Fisher I_n du modèle $(\mathscr{H}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$ vaut $I_n(\theta) =$*

$nI(\theta)$ pour chaque $\theta \in \Theta$.

Preuve Si L désigne la vraisemblance du modèle $(\mathscr{H}, \{Q_\theta\}_{\theta\in\Theta})$, la vraisemblance L_n du modèle $(\mathscr{H}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$ est :

$$L_n(x_1,\cdots,x_n;\theta) = \prod_{i=1}^n L(x_i;\theta).$$

Le score de ce dernier modèle est donc :

$$\nabla \ln L_n(x_1,\cdots,x_n;\theta) = \sum_{i=1}^n \nabla \ln L(x_i;\theta).$$

Si $(X_1,\cdots,X_n)$ est un échantillon de la loi $P_\theta = Q_\theta^{\otimes n}$, on a alors par indépendance :

$$I_n(\theta) = \mathrm{var}_\theta\left(\sum_{i=1}^n \nabla \ln L(X_i;\theta)\right) = \sum_{i=1}^n \mathrm{var}_\theta\left(\nabla \ln L(X_i;\theta)\right) = nI(\theta).$$

□

Du point de vue des calculs, on se réfèrera souvent à la proposition qui suit, dont l'objectif principal est de donner une forme simplifiée pour la matrice d'information de Fisher. Dans la suite, $\nabla^2 g(\theta)$ désigne la matrice Hessienne de $g : \Theta \to \mathbb{R}$ évaluée en $\theta \in \Theta$.

Proposition *Soit $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ un modèle statistique dominé par μ, de vraisemblance L et d'information de Fisher I. Soit $\theta \in \Theta$. On suppose qu'il existe un voisinage $V \subset \Theta$ de θ tel que $\sup_{\alpha\in V} \|\nabla L(.;\alpha)\| \in L^1(\mu)$. Alors :*

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

(ii) *si, en outre, $\sup_{\alpha\in V} \|\nabla^2 L(.;\alpha)\| \in L^1(\mu)$, on a $I(\theta) = -\mathbb{E}_\theta \nabla^2 \ln L(.;\theta)$.*

Les conditions de cette proposition ne sont pas aussi restrictives qu'elle peuvent le sembler, car elle sont satisfaites par bon nombre de modèles statistiques. Comme nous allons le voir, il s'agit essentiellement de donner des conditions pour faire passer l'opération de dérivation sous une intégrale.

Preuve On commence par remarquer que, sous la condition $\sup_{\alpha\in V} \|\nabla L(.;\alpha)\| \in L^1(\mu)$, on a d'après le théorème de Lebesgue :

$$\int_{\mathscr{H}^n} \nabla L(x;\theta)\mu(\mathrm{d}x) = \nabla \int_{\mathscr{H}^n} L(x;\theta)\mu(\mathrm{d}x) = 0.$$

Par suite,

$$\mathbb{E}_\theta \nabla \ln L(.;\theta) = \int_{\mathscr{H}^n} (\nabla \ln L(x;\theta))\, L(x;\theta)\mu(\mathrm{d}x) = \int_{\mathscr{H}^n} \nabla L(x;\theta)\mu(\mathrm{d}x) = 0,$$

d'où **(i)**. Pour montrer **(ii)**, on remarque dans un premier temps que d'après **(i)**,

$$\begin{aligned} I(\theta) &= \left(\mathrm{cov}_\theta \left(\frac{\partial}{\partial \theta_i} \ln L(.;\theta), \frac{\partial}{\partial \theta_j} \ln L(.;\theta) \right) \right)_{i,j=1,\cdots,d} \\ &= \left(\mathbb{E}_\theta \frac{\partial}{\partial \theta_i} \ln L(.;\theta) \frac{\partial}{\partial \theta_j} \ln L(.;\theta) \right)_{i,j=1,\cdots,d}. \end{aligned} \tag{3.3.1}$$

Soit alors $i, j = 1, \cdots, d$. Pour $x \in \mathscr{H}^n$, on a

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln L(x;\theta) = \frac{\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} L(x;\theta) \right)}{L(x;\theta)} - \frac{\left(\frac{\partial}{\partial \theta_i} L(x;\theta) \right) \left(\frac{\partial}{\partial \theta_j} L(x;\theta) \right)}{L^2(x;\theta)}.$$

Il est bon de remarquer que chacune des expressions qui interviennent dans le membre de droite est une fonction de x qui est dans $L^1(P_\theta)$: c'est clair pour le 1er terme car $\nabla^2 L(.;\theta) \in L^1(\mu)$; c'est vrai aussi pour le 2nd membre sous la condition d'existence de l'information de Fisher, i.e. $\nabla \ln L(.;\theta) \in L^2(P_\theta)$. Le théorème de Lebesgue montre que sous l'hypothèse $\sup_{\alpha \in V} \|\nabla^2 L(.;\alpha)\| \in L^1(\mu)$, on a :

$$\int_{\mathscr{H}^n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} L(x;\theta)\mu(\mathrm{d}x) = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int_{\mathscr{H}^n} L(x;\theta)\mu(\mathrm{d}x) = 0.$$

Par suite,

$$\begin{aligned} \mathbb{E}_\theta \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln L(.;\theta) &= \int_{\mathscr{H}^n} \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln L(x;\theta) \right) L(x;\theta)\mu(\mathrm{d}x) \\ &= -\int_{\mathscr{H}^n} \left(\frac{\partial}{\partial \theta_i} L(x;\theta) \right) \left(\frac{\partial}{\partial \theta_j} L(x;\theta) \right) \frac{1}{L(x;\theta)} \mu(\mathrm{d}x) \\ &= -\mathbb{E}_\theta \frac{\partial}{\partial \theta_i} \ln L(.;\theta) \frac{\partial}{\partial \theta_j} \ln L(.;\theta). \end{aligned}$$

D'après (3.3.1), cette dernière quantité coincide avec $-I(\theta)_{ij}$, d'où **(ii)**. □

Cette proposition légitime la définition qui suit.

Définition *On dit que le modèle statistique dominé* $(\mathscr{H}^n, \{P_\theta\}_{\theta \in \Theta})$ *dominé et de vraisemblance L est régulier si pour chaque* $\theta \in \Theta$ *:*

(i) *son information de Fisher en θ existe et est inversible ;*
(ii) $\mathbb{E}_\theta \nabla \ln L(.;\theta) = 0$ et $I(\theta) = -\mathbb{E}_\theta \nabla^2 \ln L(.;\theta)$.

La proposition précédente nous donne donc des conditions suffisantes de régularité d'un modèle. A nouveau, il est entendu dans cette définition que les conditions d'existence de l'information de Fisher sont satisfaites. De même, on n'évoque l'espérance d'une v.a. que lorsque celle-ci existe.

3.4 Normalité asymptotique de l'EMV

Théorème *Soit $(\mathscr{H}, \{Q_\theta\}_{\theta\in\Theta})$ un modèle dominé régulier, de vraisemblance L et d'information de Fisher I tel que, pour chaque $\theta \in \Theta$, il existe un voisinage $V \subset \Theta$ de θ avec $\sup_{\alpha\in V} \|\nabla^2 \ln L(.;\alpha)\| \in L^1(P_\theta)$. On note $\hat{\theta}$ l'EMV de θ associé à la vraisemblance*

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

du modèle $(\mathscr{H}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\Theta})$. Si $\hat{\theta}$ est consistant, alors il est asymptotiquement normal, de vitesse $\sqrt{n}$ et de variance asymptotique $I(\theta)^{-1}$:

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \overset{\mathscr{L}/Q_\theta^{\otimes n}}{\longrightarrow} N(0, I(\theta)^{-1}),\ \forall \theta \in \Theta.$$

Remarque Si les conditions de régularité du modèle ne sont certainement pas optimales pour garantir un tel résultat, il n'en reste pas moins qu'il est nécessaire d'imposer une certaine régularité. Considérons en effet le cas du modèle $(\mathbb{R}_+^n, \{\mathscr{U}([0,\theta])^{\otimes n}\}_{\theta>0})$. Sa vraisemblance L_n s'écrit pour $\theta > 0$:

$$L_n(x_1,\cdots,x_n;\theta) = \begin{cases} \theta^{-n} & \text{si } 0 \le x_1,\cdots,x_n \le \theta; \\ 0 & \text{sinon.} \end{cases}$$

L'EMV calculé à partir d'un échantillon $(X_1,\cdots,X_n)$ de loi $\mathscr{U}([0,\theta])^{\otimes n}$ est donc $\hat{\theta} = \max_{1\le i\le n} X_i$. Calculons maintenant sa vitesse de convergence. En adoptant la notation $P_\theta = \mathscr{U}([0,\theta])^{\otimes n}$, on a pour chaque $0 < t < n\theta$:

$$\begin{aligned} P_\theta\left(n\left(\theta-\hat{\theta}\right) \le t\right) &= 1 - P_\theta\left(\max_{1\le i\le n} X_i < \theta - \frac{t}{n}\right) \\ &= 1 - \left(1 - \frac{t}{n\theta}\right)^n. \end{aligned}$$

Comme la limite est $1-\exp(-t/\theta)$ dès que $t>0$, on a donc montré que

$$n\left(\theta-\hat{\theta}\right) \stackrel{\mathscr{L}/P_\theta}{\longrightarrow} \mathscr{E}(1/\theta).$$

Ainsi, dans cet exemple de modèle non régulier, ni la vitesse de l'EMV, ni la loi limite, ne correspondent à celles du théorème.

Preuve On fixe $\theta\in\Theta$ et on pose $P_\theta=Q_\theta^{\otimes n}$. Dans la suite, $(X_1,\cdots,X_n)$ est un échantillon de loi P_θ. Pour chaque $\alpha\in\Theta$, on note :

$$\mathscr{L}_n(\alpha)=\ln L_n(X_1,\cdots,X_n;\alpha)=\sum_{i=1}^n \ln L(X_i;\alpha).$$

Comme $\hat{\theta}$ maximise $\mathscr{L}_n$, un développement de Taylor avec reste intégral nous donne :

$$0=\nabla\mathscr{L}_n(\hat{\theta})=\nabla\mathscr{L}_n(\theta)+\left(\int_0^1 \nabla^2\mathscr{L}_n\left(\theta+t(\hat{\theta}-\theta)\right)\mathrm{d}t\right)(\hat{\theta}-\theta). \qquad (3.4.1)$$

Nous examinons séparément chacun des termes qui interviennent dans cette relation. Rappelons que, puisque le modèle est régulier,

$$\mathbb{E}_\theta\nabla\ln L(.;\theta)=0.$$

Par ailleurs, $\mathrm{var}_\theta(\nabla\ln L(.;\theta))=I(\theta)$. Donc, d'après le théorème de la limite centrale :

$$\frac{1}{\sqrt{n}}\nabla\mathscr{L}_n(\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\nabla\ln L(X_i;\theta)\stackrel{\mathscr{L}/P_\theta}{\longrightarrow}N(0,I(\theta)). \qquad (3.4.2)$$

Montrons maintenant que :

$$\frac{1}{n}\int_0^1\nabla^2\mathscr{L}_n\left(\theta+t(\hat{\theta}-\theta)\right)\mathrm{d}t\stackrel{P_\theta}{\longrightarrow}-I(\theta)$$

Notons, pour chaque $x\in\mathscr{H}^n$ et $r>0$:

$$\sigma(x,r)=\sup_{\|\alpha-\theta\|\le r}\|\nabla^2\ln L(x;\alpha)-\nabla^2\ln L(x;\theta)\|.$$

Or, $\sigma(.,r)\in L^1(P_\theta)$ pour r assez petit et de plus, $\ln L(x;.)\in\mathscr{C}^2$ pour chaque $x\in\mathscr{H}^n$. Fixons $\varepsilon>0$. D'après le théorème de Lebesgue, il existe $r>0$ tel que $\mathbb{E}_\theta\sigma(.,r)<\varepsilon/2$. Par ailleurs, comme

$$\frac{1}{n}\int_0^1\nabla^2\mathscr{L}_n\left(\theta+t(\hat{\theta}-\theta)\right)\mathrm{d}t=\frac{1}{n}\sum_{i=1}^n\int_0^1\nabla^2\ln L\left(X_i;\theta+t(\hat{\theta}-\theta)\right)\mathrm{d}t,$$

on obtient :

$$\begin{aligned}
&P_\theta\left(\left\|\frac{1}{n}\int_0^1 \nabla^2\mathscr{L}_n\left(\theta+t(\hat\theta-\theta)\right)\mathrm{d}t+I(\theta)\right\|\geq\varepsilon\right)\\
&\quad\leq P_\theta\left(\left\|\frac{1}{n}\sum_{i=1}^n\int_0^1\left[\nabla^2\ln L\left(X_i;\theta+t(\hat\theta-\theta)\right)-\nabla^2\ln L(X_i;\theta)\right]\mathrm{d}t\right\|\geq\frac{\varepsilon}{2}\right)\\
&\qquad+P_\theta\left(\left\|\frac{1}{n}\sum_{i=1}^n\nabla^2\ln L(X_i;\theta)+I(\theta)\right\|\geq\frac{\varepsilon}{2}\right)\\
&\quad\leq P_\theta\left(\frac{1}{n}\sum_{i=1}^n\sigma(X_i,r)\geq\frac{\varepsilon}{2}\right)+P_\theta\left(\|\hat\theta-\theta\|\geq r\right)\\
&\qquad+P_\theta\left(\left\|\frac{1}{n}\sum_{i=1}^n\nabla^2\ln L(X_i;\theta)+I(\theta)\right\|\geq\frac{\varepsilon}{2}\right).
\end{aligned}$$

Le passage à la dernière inégalité a été obtenu par une intersection avec l'événement $\{\|\hat\theta-\theta\|<r\}$. Or, $\mathbb{E}_\theta\sigma(.,r)<\varepsilon/2$ et $\mathbb{E}_\theta\nabla^2\ln L(.;\theta)=-I(\theta)$ car le modèle est régulier. Comme $\hat\theta$ est consistant, on a donc, d'après la loi des grands nombres :

$$\frac{1}{n}\int_0^1\nabla^2\mathscr{L}_n\left(\theta+t(\hat\theta-\theta)\right)\mathrm{d}t\xrightarrow{P_\theta}-I(\theta).$$

En particulier, $I(\theta)$ étant inversible,

$$P_\theta\left(\frac{1}{n}\int_0^1\nabla^2\mathscr{L}_n\left(\theta+t(\hat\theta-\theta)\right)\mathrm{d}t\text{ inversible}\right)\longrightarrow 1.$$

Or, sur ce dernier événement, d'après (3.4.1) :

$$\sqrt{n}(\hat\theta-\theta)=-\frac{1}{\sqrt{n}}\left(\frac{1}{n}\int_0^1\nabla^2\mathscr{L}_n\left(\theta+t(\hat\theta-\theta)\right)\mathrm{d}t\right)^{-1}\nabla\mathscr{L}_n(\theta).$$

En réunissant toutes les pièces, on en déduit de (3.4.2) que

$$\sqrt{n}\left(\hat\theta-\theta\right)\xrightarrow{\mathscr{L}/P_\theta}I(\theta)^{-1}N(0,I(\theta))=N(0,I(\theta)^{-1}),$$

d'où le théorème. □

Chapitre 4

Classification des statistiques

Comme dans tout domaine des mathématiques, classer les objets en fonction de propriétés communes est un moyen efficace pour entreprendre leurs études.

4.1 Estimateurs efficaces

On suppose dans cette section que l'espace des paramètres $\Theta \subset \mathbb{R}$ est un ouvert, que $\mathscr{H} \subset \mathbb{R}^k$ et que $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est un modèle statistique régulier dominé par μ, de vraisemblance L et d'information de Fisher I.

Dans la section 2.1, nous nous sommes intéressés à des bornes du risque quadratique, et donc de la variance, dans la famille des estimateurs sans biais. Nous poursuivons ici dans cette étude. Avant tout, nous aurons besoin de la définition suivante qui prendra tout son sens avec l'inégalité de Cramer-Rao.

Définition *On dit que $\hat{\theta}$ est un estimateur régulier si il est d'ordre 2 et*

$$\nabla \int_{\mathscr{H}^n} \hat{\theta}(.)L(.;\theta)\mathrm{d}\mu = \int_{\mathscr{H}^n} \hat{\theta}(.)\nabla L(.;\theta)\mathrm{d}\mu.$$

L'intérêt de cette définition réside dans la remarque suivante : sous les notations de cette définition, si l'estimateur régulier $\hat{\theta}$ est sans biais, alors

$$\int_{\mathscr{H}^n} \hat{\theta}(.)\nabla L(.;\theta)\mathrm{d}\mu = \nabla \mathbb{E}_\theta \hat{\theta}(.) = 1.$$

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

une vitesse seuil qu'il serait illusoire de vouloir améliorer.

Théorème [CRAMER-RAO] *Pour tout estimateur $\hat{\theta}$ régulier et sans biais, on a :*

$$\mathscr{R}(\theta,\hat{\theta}) \geq I(\theta)^{-1},\ \forall \theta \in \Theta.$$

Le terme $I(\theta)^{-1}$ s'appelle borne de Cramer-Rao.

Preuve Soit $\theta \in \Theta$. L'inégalité de Cauchy-Schwarz nous donne :

$$\mathscr{R}(\theta,\hat{\theta}) = \mathrm{var}_\theta(\hat{\theta}) \geq \frac{\left(\mathrm{cov}_\theta\left(\hat{\theta}, \nabla \ln L(.;\theta)\right)\right)^2}{\mathrm{var}_\theta\left(\nabla \ln L(.;\theta)\right)}. \tag{4.1.1}$$

Par définition de $I(\theta)$, il suffit donc de montrer que $\mathrm{cov}_\theta\left(\hat{\theta}, \nabla \ln L(.;\theta)\right) = 1$. Comme $\hat{\theta}$ est régulier et sans biais, on a

$$\int_{\mathscr{H}^n} \hat{\theta}(x) \nabla L(x;\theta) \mu(\mathrm{d}x) = 1.$$

Par ailleurs, $\mathbb{E}_\theta \nabla \ln L(.;\theta) = 0$ car le modèle est régulier. En conséquence :

$$\begin{aligned} \mathrm{cov}_\theta\left(\hat{\theta}, \nabla \ln L(.;\theta)\right) &= \int_{\mathscr{H}^n} \hat{\theta}(x) \frac{\nabla L(x;\theta)}{L(x;\theta)} P_\theta(\mathrm{d}x) \\ &= \int_{\mathscr{H}^n} \hat{\theta}(x) \nabla L(x;\theta) \mu(\mathrm{d}x) \\ &= 1, \end{aligned}$$

d'où le théorème. □

Reprenons l'exemple du modèle statistique $(\{0,1\}^n, \{\mathscr{B}(p)^{\otimes n}\}_{p\in]0,1[})$ de la section 1.1. Nous avons montré dans la section 2.1 que l'estimateur $\bar{X}_n$ construit à partir d'un échantillon $(X_1,\cdots,X_n)$ de la loi $\mathscr{B}(p)^{\otimes n}$ est VUMSB, ce qui s'exprime par la propriété :

$$\mathrm{var}_p(\hat{\theta}) = \mathscr{R}(p;\hat{\theta}) \geq \mathscr{R}(p;\bar{X}_n) = \mathrm{var}_p(\bar{X}_n) = \frac{p(1-p)}{n},$$

pour tout autre estimateur sans biais $\hat{\theta}$. Un simple calcul nous montre aussi que l'information de Fisher de ce modèle est précisément

$$I(p) = \frac{n}{p(1-p)}.$$

Ainsi, la borne de l'inégalité de Cramer-Rao, communément appelée *borne de Cramer-Rao*, est atteinte. Cette remarque donne tout son sens à la définition qui suit :

Définition *Un estimateur sans biais d'ordre 2 est dit uniformément efficace si il atteint la borne de Cramer-Rao du modèle.*

Si tout estimateur uniformément efficace est VUMSB, la réciproque n'est pas vraie, et ces 2 notions ne sont donc pas les mêmes. La proposition suivante nous montre qu'il est possible de décrire les estimateurs uniformément efficaces.

Proposition *Soit $\hat{\theta}$ un estimateur régulier et sans biais. Alors, $\hat{\theta}$ est uniformément efficace si, et seulement si, il existe une fonction $\psi : \Theta \to \mathbb{R}$ telle que*

$$\forall \theta \in \Theta,\ \hat{\theta} = \theta + \psi(\theta)\nabla \ln L(.;\theta)\ P_\theta - p.s.$$

Preuve Soit $\theta \in \Theta$. D'après (4.1.1), $\hat{\theta}$ est uniformément efficace si et seulement si

$$\mathrm{var}_\theta(\hat{\theta})\mathrm{var}_\theta(\nabla \ln L(.;\theta)) = \left(\mathrm{cov}_\theta(\hat{\theta}, \nabla \ln L(.;\theta)\right)^2.$$

On est donc dans un cas d'égalité dans l'inégalité de Cauchy-Schwarz, ce qui signifie qu'il existe $\psi(\theta)$ tel que

$$\hat{\theta} - \mathbb{E}_\theta \hat{\theta} = \psi(\theta)\left(\nabla \ln L(.;\theta) - \mathbb{E}_\theta \nabla \ln L(.;\theta)\right) \quad P_\theta - p.s.$$

Comme $\hat{\theta}$ est sans biais et $\nabla \ln L(.;\theta)$ est P_θ-centrée, la proposition est prouvée. □

Bien sûr, cette proposition est un "miroir aux alouettes", dans la mesure où l'estimateur uniformément efficace est alors décrit via le paramètre inconnu θ. En fait, l'intérêt d'une telle représentation réside dans le fait que l'on peut quelquefois en déduire qu'un estimateur est uniformément efficace. On peut ainsi facilement retrouver le fait que la moyenne empirique est l'estimateur VUMSB dans le modèle statistique $(\{0,1\}^n, \{\mathscr{B}(p)^{\otimes n}\}_{p\in]0,1[})$. Pour changer d'exemple, considérons plutôt le modèle statistique $(\mathbb{R}^n, \{N(m,\sigma^2)^{\otimes n}\}_{\sigma>0})$, avec $m \in \mathbb{R}$ connu. Si $(X_1, \cdots, X_n)$ est un échantillon de la loi $N(m,\sigma^2)^{\otimes n}$, l'estimateur

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$$

est sans biais -car m est connu- et régulier. Par ailleurs, la vraisemblance L s'écrit, pour $\sigma > 0$ et $x_1, \cdots, x_n \in \mathbb{R}$:

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

Par suite, sa log-vraisemblance vérifie :

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \ln L(x_1, \cdots, x_n; \sigma^2) &= \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - m)^2 \\ &= \frac{n}{2\sigma^4} \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2 - \sigma^2 \right). \end{aligned}$$

On en déduit de la proposition précédente que $\hat{\sigma^2}$ est uniformément efficace.

4.2 Statistiques exhaustives

Dans cette partie, le modèle statistique étudié est $(\mathscr{H}^n, \{P_\theta\}_{\theta \in \Theta})$, avec $\mathscr{H} \subset \mathbb{R}^k$ et $\Theta \subset \mathbb{R}^d$.

Le principe d'exhaustivité d'une statistique est un principe de réduction des données, qui est basé sur la notion de loi conditionnelle. Dans la suite, $\mathscr{L}_{P_\theta}(Z_1|Z_2)$ désigne la loi conditionnelle, sous P_θ, de Z_1 sachant Z_2.

Définition *La statistique g est dite exhaustive si, pour chaque $\theta \in \Theta$,*

$$\mathscr{L}_{P_\theta}(X_1, \cdots, X_n | g(X_1, \cdots, X_n))$$

ne dépend pas de θ, où $(X_1, \cdots, X_n)$ est un échantillon de loi P_θ.

En clair, l'échantillon n'apporte pas plus d'information sur la valeur du paramètre inconnu qu'une statistique exhaustive. Autrement dit, une statistique exhaustive élimine toute l'information superflue dans l'échantillon, en ne retenant que la partie informative sur le paramètre inconnu.

Reprenons le cas du modèle $(\{0,1\}^n, \{\mathscr{B}(p)^{\otimes n}\}_{p \in]0,1[})$ introduit dans la section 1.1. L'ordre dans lequel sont observés les tirages de "pile" ou "face" n'apporte aucune information supplémentaire sur le paramètre inconnu. Du coup, on

peut résumer la suite des observations $x_1,\cdots,x_n$ par leur somme $x_1+\cdots+x_n$, ce qui indique que l'estimateur $\bar{X}_n$ issu de l'échantillon $(X_1,\cdots,X_n)$ de la loi $\mathscr{B}(p)^{\otimes n}$ est exhaustif. Faisons le calcul pour étayer cette intuition. Pour chaque $y_1,\cdots,y_n\in\{0,1\}$ et $z\in\{0,\cdots,n\}$ tels que $y_1+\cdots+y_n=z$:

$$\begin{aligned}\mathscr{B}(p)^{\otimes n}\left(X_1=y_1,\cdots,X_n=y_n\middle|n\bar{X}_n=z\right) &= \frac{\mathscr{B}(p)^{\otimes n}\Big(X_1=y_1,\cdots,X_n=y_n\Big)}{\mathscr{B}(p)^{\otimes n}(n\bar{X}_n=z)}\\ &= \frac{p^z(1-p)^{n-z}}{C_n^z p^z(1-p)^{n-z}}=\frac{1}{C_n^z}.\end{aligned}$$

Sous $\mathscr{B}(p)^{\otimes n}$, la loi de $(X_1,\cdots,X_n)$ sachant $n\bar{X}_n$ est donc la loi uniforme sur l'ensemble $\{y\in\{0,1\}^n : y_1+\cdots+y_n=n\bar{X}_n\}$. Cette loi ne dépend pas du paramètre p, donc $\bar{X}_n$ est une statistique exhaustive : toute l'information sur p contenue dans l'échantillon $(X_1,\cdots,X_n)$ est en fait contenue dans $\bar{X}_n$.

Le théorème ci-dessous nous donne une caractérisation simple de l'exhaustivité.

Théorème [NEYMAN-FISHER] *Supposons que le modèle $(\mathscr{H}^n,\{P_\theta\}_{\theta\in\Theta})$ est dominé par μ. Une statistique g à valeurs dans $\mathbb{R}^q$ est exhaustive si, et seulement si, il existe 2 applications boréliennes $\psi:\mathbb{R}^q\times\Theta\to\mathbb{R}_+$ et $\gamma:\mathscr{H}^n\to\mathbb{R}_+$ telles que la vraisemblance L pour μ s'écrit :*

$$L(x;\theta)=\psi(g(x),\theta)\gamma(x),\forall(x,\theta)\in\mathscr{H}^n\times\Theta.$$

Il est alors très facile de montrer avec ce théorème qu'une statistique est exhaustive. Par exemple, la moyenne empirique est une statistique exhaustive dans le modèle $(\mathbb{R}^n,\{N(m,1)^{\otimes n}\}_{m\in\mathbb{R}})$, car la vraisemblance pour la mesure de Lebesgue sur $\mathbb{R}^n$ vaut

$$L(x;m)=\left\{\exp\left(-\frac{1}{2}n(\bar{x}_n-m)^2\right)\right\}\left\{\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x}_n)^2\right)\right\},$$

pour tout $x=(x_1,\cdots,x_n)^T\in\mathbb{R}^n$ et $m\in\mathbb{R}$.

Preuve On a vu en dans la section 1.4 qu'il existe, dans le convexifié de $\{P_\theta\}_{\theta\in\Theta}$, une probabilité qui domine le modèle statistique. Pour simplifier la preuve, on va

supposer que la mesure dominante μ est cette mesure, i.e.

$$\mu = \sum_n a_n P_{\theta_n},$$

avec $(\theta_n)_n \subset \Theta$ et $(a_n)_n \subset [0,1]$ tel que $\sum_n a_n = 1$. Dans ce cadre, nous allons montrer que g est exhaustive si, et seulement si

$$L(x;\theta) = \psi(g(x),\theta) \quad \forall (x,\theta) \in \mathscr{H}^n \times \Theta, \tag{4.2.1}$$

pour une fonction mesurable $\psi : \mathbb{R}^q \times \Theta \to \mathbb{R}_+$. Au préalable, remarquons que la loi $P_\theta \circ g^{-1}$ de g est absolument continue par rapport à $\mu \circ g^{-1}$, et de densité $\varphi(.,\theta) = \mathbb{E}_\mu[L(.;\theta)\,|\,g = .]$, si $\mathbb{E}_\mu$ désigne l'espérance sous μ. En effet, on a pour tout $A \in \mathscr{B}(\mathbb{R}^q)$:

$$\begin{aligned} P_\theta \circ g^{-1}(A) &= P_\theta\,(g \in A) = \int_{g^{-1}(A)} L(.;\theta)\mathrm{d}\mu = \int_{g^{-1}(A)} \mathbb{E}_\mu[L(.;\theta)\,|\,g]\,\mathrm{d}\mu \\ &= \int_A \mathbb{E}_\mu[L(.;\theta)\,|\,g = x]\mu \circ g^{-1}(\mathrm{d}x). \end{aligned}$$

d'après le théorème de transfert et par définition de l'espérance conditionnelle.

On suppose tout d'abord que L se factorise comme dans (4.2.1). Soit $\theta \in \Theta$. Comme $P_\theta \circ g^{-1}$ est la loi de g, il faut montrer que pour tout $A \in \mathscr{B}(\mathbb{R}^q)$ et tout $B \in \mathscr{B}(\mathscr{H}^n)$:

$$P_\theta(\{g \in A\} \cap B) = \int_A K(x,B) P_\theta \circ g^{-1}(\mathrm{d}x),$$

avec K un noyau indépendant de θ. Pour tout $A \in \mathscr{B}(\mathbb{R}^q)$ et $B \in \mathscr{B}(\mathscr{H}^n)$:

$$\begin{aligned} P_\theta\,(\{g \in A\} \cap B) &= \int_{\mathscr{H}^n} \mathbf{1}_B\,\mathbf{1}_A \circ g(.)\,\psi(g(.),\theta)\,\mathrm{d}\mu \\ &= \int_{\mathscr{H}^n} \mathbb{E}_\mu\,[\mathbf{1}_B\,\mathbf{1}_A \circ g(.)\,\psi(g(.),\theta)\,|g]\,\mathrm{d}\mu \\ &= \int_{\mathscr{H}^n} \mu(B\,|g)\,\mathbf{1}_A \circ g(.)\,\psi(g(.),\theta)\,\mathrm{d}\mu \\ &= \int_{\mathbb{R}^q} \mu(B\,|\,g = x)\,\mathbf{1}_A(x)\,\psi(x,\theta)\,\mu \circ g^{-1}(\mathrm{d}x), \end{aligned}$$

où on a noté $\mu(B\,|\,g) = \mathbb{E}_\mu[\mathbf{1}_B\,|\,g]$. Pour la dernière chaîne d'égalités, on a utilisé successivement la définition de l'espérance conditionnelle et l'une de ses propriétés fondamentales ($\mathbb{E}_\mu[XY|\mathscr{G}] = X\mathbb{E}_\mu[Y|\mathscr{G}]$ si X est $\mathscr{G}$-mesurable, dès que XY et

Y sont dans $L^1(\mu)$), puis le théorème de transfert. Comme $\mathbb{E}_\mu[L(.;\theta)\,|\,g=.] = \psi(.,\theta)$ est la densité de $P_\theta \circ g^{-1}$ par rapport à $\mu \circ g^{-1}$, on a donc obtenu :

$$\begin{aligned} P_\theta\left(\{g \in A\} \cap B\right) &= \int_A \mu(B\,|g=x)\,\psi(x,\theta)\mu \circ g^{-1}(\mathrm{d}x) \\ &= \int_A \mu(B\,|g=x)\,P_\theta \circ g^{-1}(\mathrm{d}x) \end{aligned}$$

Le noyau de transition $K(x,B) = \mu(B\,|\,g=x)$ associé à la loi conditionnelle sous P_θ de l'échantillon sachant g est indépendant de θ, c'est-à-dire que g est une statistique exhaustive.

Supposons maintenant que g est exhaustive. Soit $\theta \in \Theta$. Comme g est exhaustive, la loi conditionnelle $P_\theta(.\,|\,g=.)$ est indépendante de θ ; notons-là $P(.\,|\,g=.)$. Alors, pour tout $B \in \mathscr{B}(\mathscr{H}^n)$ et $x \in \mathbb{R}^q$:

$$\mu(B\,|\,g=x) = \sum_n a_n P_{\theta_n}(B\,|\,g=x) = P(B\,|\,g=x),$$

i.e. les lois conditionnelles $P(.\,|\,g=.)$ et $\mu(.\,|\,g=.)$ sont les mêmes $\mu \circ g^{-1}$-p.s. Par suite, pour tous $A \in \mathscr{B}(\mathbb{R}^q)$ et $B \in \mathscr{B}(\mathscr{H}^n)$:

$$\begin{aligned} P_\theta(\{g \in A\} \cap B) &= \int_A P(B\,|\,g=x) P_\theta \circ g^{-1}(\mathrm{d}x) \\ &= \int_A \mu(B\,|\,g=x)\varphi(x,\theta)\mu \circ g^{-1}(\mathrm{d}x), \end{aligned}$$

car $\varphi(.,\theta) = \mathbb{E}_\mu[L(.;\theta)\,|\,g=.]$ est la densité de $P_\theta \circ g^{-1}$ par rapport à $\mu \circ g^{-1}$. Par ailleurs, on a aussi par définition de l'espérance conditionnelle :

$$P_\theta(\{g \in A\} \cap B) = \int_{g^{-1}(A)} \mathbf{1}_B L(.;\theta)\mathrm{d}\mu = \int_A \mathbb{E}_\mu[\mathbf{1}_B L(.;\theta)\,|\,g=x]\mu \circ g^{-1}(\mathrm{d}x).$$

Ces égalités étant vraies pour tout $A \in \mathscr{B}(\mathbb{R}^q)$, on en déduit que $\mu \circ g^{-1}$-p.s. :

$$\mathbb{E}_\mu[\mathbf{1}_B\,\varphi(g(.),\theta)\,|\,g=.] = \mu(B\,|\,g=.)\varphi(.,\theta) = \mathbb{E}_\mu[\mathbf{1}_B L(.;\theta)\,|\,g=.].$$

Par suite, on a μ-p.s. :

$$\mathbb{E}_\mu\left[\mathbf{1}_B\left(\varphi(g(.),\theta) - L(.;\theta)\right)\,\middle|\,g\right] = 0,$$

et donc, en particulier, pour tout $B \in \mathscr{B}(\mathscr{H}^n)$:

$$\mathbb{E}_\mu \left[\mathbf{1}_B \left(\varphi(g(.), \theta) - L(.; \theta) \right) \right] = 0.$$

Ceci étant vrai pour tout $B \in \mathscr{B}(\mathscr{H}^n)$, on a bien $L(.;\theta) = \varphi(g(.),\theta)$ μ-p.s., d'où la factorisation (4.2.1) $\square$

Une fois caractérisé par des moyens simples, on remarque -comme on pouvait s'y attendre- que le concept d'exhaustivité permet d'améliorer un estimateur, en terme de risque. C'est l'objet du théorème ci-dessous.

Théorème [RAO-BLACKWELL] *Soit g une statistique, et $\hat{\theta}$ un estimateur d'ordre 2. Si g est exhaustive, alors la statistique $\mathbb{E}_\theta[\hat{\theta}|g]$ est un estimateur préférable à $\hat{\theta}$, et de même biais que $\hat{\theta}$.*

Preuve On fixe $\theta \in \Theta$. Comme g est exhaustive, $\mathbb{E}_\theta[\hat{\theta}|g]$, qui ne dépend pas de θ, est donc un estimateur. Notons-le $\hat{\eta}$. Comme

$$\mathbb{E}_\theta \hat{\eta} = \mathbb{E}_\theta \mathbb{E}_\theta[\hat{\theta}|g] = \mathbb{E}_\theta \hat{\theta},$$

les 2 estimateurs ont même biais. Par ailleurs,

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

où l'on a utilisé le fait que $\hat{\theta}$ et $\hat{\eta}$ ont même biais. Or,

$$\begin{aligned} \mathbb{E}_\theta \left[\left(\hat{\theta} - \hat{\eta} \right)^T (\hat{\eta} - \mathbb{E}_\theta \hat{\eta}) \middle| g \right] &= \mathbb{E}_\theta \left[\hat{\theta} - \hat{\eta} | g \right]^T (\hat{\eta} - \mathbb{E}_\theta \hat{\eta}) \\ &= (\hat{\eta} - \hat{\eta})^T (\hat{\eta} - \mathbb{E}_\theta \hat{\eta}) \\ &= 0, \end{aligned}$$

ce qui montre que

$$\mathbb{E}_\theta \left(\hat{\theta} - \hat{\eta} \right)^T (\hat{\eta} - \mathbb{E}_\theta \hat{\eta}) = \mathbb{E}_\theta \mathbb{E}_\theta \left[\left(\hat{\theta} - \hat{\eta} \right)^T (\hat{\eta} - \mathbb{E}_\theta \hat{\eta}) \middle| g \right] = 0.$$

Donc, $V_\theta(\hat{\theta}) \geq V_\theta(\hat{\eta})$ d'où, d'après la décomposition Biais-Variance :

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

ce qui nous donne le résultat. □

Reprenons le cas du modèle $(\{0,1\}^n, \{\mathscr{B}(p)^{\otimes n}\}_{p\in]0,1[})$ introduit dans la section 1.1. Lorsque $(X_1,\cdots,X_n)$ est un échantillon de la loi $P_p = \mathscr{B}(p)^{\otimes n}$, on sait que X_1 est un estimateur sans biais, et que $\bar{X}_n$ lui est préférable. Nous allons retrouver ce résultat en utilisant le théorème de Rao-Blackwell. On a déjà montré que $\bar{X}_n$ est une statistique exhaustive. D'après le théorème de Rao-Blackwell, $\mathbb{E}_p[X_1|\bar{X}_n]$ est donc un estimateur préférable à X_1. Or, comme $X_1,\cdots,X_n$ sont i.i.d., on a pour tout $j\in\{1,\cdots,n\}$ et $A\in\mathscr{B}(\mathbb{R})$:

$$\begin{aligned}\int_{\{\bar{X}_n\in A\}} \mathbb{E}_p[X_1|\bar{X}_n]\mathrm{d}P_p &= \int_{\{\bar{X}_n\in A\}} X_1\mathrm{d}P_p = \int_{\{\bar{X}_n\in A\}} X_j\mathrm{d}P_p \\ &= \int_{\{\bar{X}_n\in A\}} \mathbb{E}_p[X_j|\bar{X}_n]\mathrm{d}P_p.\end{aligned}$$

Ceci étant vrai pour chaque $A\in\mathscr{B}(\mathbb{R})$, on en déduit de l'unicité de l'espérance conditionnelle que $\mathbb{E}_p[X_1|\bar{X}_n] = \mathbb{E}_p[X_j|\bar{X}_n]$ P_p-p.s. Par suite :

$$\mathbb{E}_p[X_1|\bar{X}_n] = \frac{1}{n}\sum_{j=1}^n \mathbb{E}_p[X_j|\bar{X}_n] = \mathbb{E}_p[\bar{X}_n|\bar{X}_n] = \bar{X}_n,\ P_p - p.s.$$

L'estimateur préférable construit avec le théorème de Rao-Blackwell n'est autre que l'inévitable moyenne empirique !

4.3 Statistiques complètes

Dans cette partie, le modèle statistique étudié est $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, avec $\mathscr{H}\subset\mathbb{R}^k$ et $\Theta\subset\mathbb{R}^d$. Dans la suite, on note aussi :

$$\mathbb{L} = \left\{f : \mathscr{H}^n\to\mathbb{R} \ :\ f\in L^1(P_\theta)\ \forall\theta\in\Theta\right\}$$

Définition *On dit qu'une statistique g à valeurs dans $\mathbb{R}^q$ est complète si, pour toute fonction $\xi : \mathbb{R}^q\to\mathbb{R}$ telle que $\xi\circ g\in\mathbb{L}$:*

$$\mathbb{E}_\theta\,\xi\circ g(.) = 0,\ \forall\theta\in\Theta \implies \xi\circ g = 0\ P_\theta - p.s.,\ \forall\theta\in\Theta.$$

De plus, lorsque $g = Id$, le modèle statistique est dit complet.

Exemple Le modèle binomial $(\{0,\cdots,\ell\},\{\mathscr{B}(\ell,\theta)\}_{\theta\in]0,1[})$ est complet. En effet, soit ξ une fonction numérique d'intégrale nulle sous $P_\theta=\mathscr{B}(\ell,\theta)$, et ceci pour chaque $\theta\in]0,1[$. Alors,

$$0=\sum_{k=0}^{\ell}\xi(k)C_\ell^k\theta^k(1-\theta)^{\ell-k}=(1-\theta)^\ell\sum_{k=0}^{\ell}\xi(k)C_\ell^k\left(\frac{\theta}{1-\theta}\right)^k.$$

Comme cette égalité est valable pour tout $\theta\in]0,1[$, il en résulte que $\xi=0$ sur $\{0,\cdots,\ell\}$, donc $\xi=0$ P_θ-p.s., i.e. le modèle binomial est complet.

Le concept prend tout son sens grâce au résultat suivant :

Théorème [LEHMANN-SCHEFFÉ] *Soit $\hat\theta$ un estimateur sans biais d'ordre 2. Si g est une statistique exhaustive complète, alors la statistique $\mathbb{E}_\theta[\hat\theta|g]$ est l'unique estimateur VUMSB.*

Preuve Soit $\hat\theta'$ un autre estimateur sans biais et tel que $\hat\theta'\in L^2(P_\theta)$ pour chaque $\theta\in\Theta$. On fixe $\theta\in\Theta$, et on note

$$\eta=\mathbb{E}_\theta[\hat\theta|g]\text{ et }\eta'=\mathbb{E}_\theta[\hat\theta'|g].$$

Par exhaustivité de g, η et η' sont des estimateurs. En outre, ils sont sans biais et dans $L^2(P_\theta)$. D'après le lemme de Doob, il existe une fonction borélienne ξ telle que $\eta-\eta'=\xi\circ g$. Donc, comme η et η' sont sans biais :

$$0=\mathbb{E}_\theta(\eta-\eta')=\mathbb{E}_\theta\xi\circ g,$$

ce qui montre que $\eta-\eta'=\xi\circ g=0$ P_θ-p.s. car g est une statistique complète. Pour finir, on remarque que d'après l'inégalité de Jensen pour les espérances conditionnelles (appliquée à la fonction convexe $x\mapsto\|x\|^2$) :

$$\begin{aligned}\mathscr{R}(\theta;\eta)=\mathscr{R}(\theta;\eta') &= V_\theta(\eta')=\mathbb{E}_\theta\|\mathbb{E}_\theta[\hat\theta'|g]-\theta\|^2\\ &\le \mathbb{E}_\theta\mathbb{E}_\theta[\|\hat\theta'-\theta\|^2|g]=V_\theta(\hat\theta')=\mathscr{R}(\theta;\hat\theta'),\end{aligned}$$

ce qui entraîne que η est VUMSB. □

Ainsi, dès que l'on dispose d'une statistique complète, tout estimateur sans biais, même déraisonnable, suffit pour déterminer l'estimateur VUMSB. Pour illustrer cette affirmation, reprenons le modèle $(\{0,1\}^n,\{\mathscr{B}(p)^{\otimes n}\}_{p\in]0,1[})$ de la

section 1.1. Nous allons à nouveau montrer, cette fois à l'aide du théorème de Lehmann-Scheffé, que l'estimateur $\bar{X}_n$ construit avec l'échantillon $(X_1, \cdots, X_n)$ de la loi $P_p = \mathscr{B}(p)^{\otimes n}$ est VUMSB. Comme X_1 est un estimateur sans biais, que $\bar{X}_n$ est une statistique exhaustive et que $\mathbb{E}_p[X_1|\bar{X}_n] = \bar{X}_n$, il reste à prouver que $\bar{X}_n$ est une statistique complète. Sous P_p, la loi de $n\bar{X}_n$ est $\mathscr{B}(n,p)$. Donc, pour chaque fonction ξ à valeurs réelles,

$$\mathbb{E}_p \xi(\bar{X}_n) = \sum_{k=0}^{n} \xi\left(\frac{k}{n}\right) C_n^k p^k (1-p)^{n-k}.$$

Si $\mathbb{E}_p \xi(\bar{X}_n) = 0$ pour chaque $p \in]0,1[$, on a alors $\xi(k/n) = 0$ pour chaque $k \in \{0, \cdots, n\}$ et donc $\xi(\bar{X}_n) = 0$ P_p-p.s. Par suite, $\bar{X}_n$ est une statistique complète.

Chapitre 5

Test statistique

Reprenons la problématique de la section 1.1. Au niveau de confiance 95%, l'intervalle de confiance obtenu pour la valeur de p_0 (la probabilité que la pièce tombe sur pile) est $[0.45, 0.59]$. On n'est donc pas en mesure de préciser si la pièce est ou non équilibrée : un intervalle de confiance ne fournit pas, en général, une procédure de décision.

L'objet de ce chapitre est de construire une procédure de décision, le *test statistique*. Il faut avoir à l'esprit que, outre le fait que cette procédure doit rendre une décision, elle doit aussi garder un contrôle sur ses propres erreurs.

On considère dans ce chapitre un modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$. Il faut noter que ni $\mathscr{H}$, ni Θ n'est spécifié.

5.1 Problème de test

Pour une raison ou une autre, on est amené à penser que la vraie valeur du paramètre θ, i.e. celle qui est issue de l'observation $x_1, \cdots, x_n$, se trouve dans un sous-ensemble Θ_0 de Θ. On formule alors une hypothèse, appelée *hypothèse nulle*, et notée $H_0 : \theta \in \Theta_0$. Cependant, cette hypothèse peut malgré tout être fausse, et on est amené à introduire l'*hypothèse alternative* $H_1 : \theta \in \Theta_1$, avec $\Theta_1 \subset \Theta_0^c$. Un *problème de test* est la confrontation de l'hypothèse nulle H_0 contre l'hypothèse alternative H_1.

A ce niveau, il convient de formuler 2 observations :

- ▷ Θ_1 n'est pas nécessairement égal à Θ_0^c : ceci illustre le fait que, dans un problème de test, il faut choisir une hypothèse alternative qui, en cas de rejet de H_0, présente plus de pertinence que H_0 ;
- ▷ dissymétrie entre H_0 et H_1, car le test est construit à partir de la présomption que H_0 est vraie. De même que dans un procès aux assises, il y a présomption d'innocence, dans un problème de test, il y a présomption de H_0. Comme dans un procès où il faut alors prouver avec certitude que le détenu est coupable pour le condamner, le corollaire de ce principe est qu'il faut montrer que H_0 est peu probable pour la rejeter. De ce point de vue, la décision la plus convaincante est donc de rejeter H_0 !

A l'instar des estimateurs, toute procédure de décision sur un problème de test est élaborée à partir d'une observation $(x_1, \cdots, x_n) \in \mathscr{H}^n$. Un test peut alors être représenté par une fonction de l'observation, qui vaut 0 lorsque celle-ci conduit à accepter H_0 et qui vaut 1 dans le cas contraire.

Définition *Un test pur est une statistique T à valeurs dans $\{0,1\}$: pour l'observation $x \in \mathscr{H}^n$, si $T(x) = 0$ alors H_0 est acceptée ; si $T(x) = 1$ alors H_0 est rejetée. La zone de rejet (resp. d'acceptation) du test est $T^{-1}(\{1\})$ (resp. $T^{-1}(\{0\})$.*

Un test pur correspond donc à une décision binaire, qui ne correspond généralement pas à la complexité des situations envisagées. Considérons en effet le cas d'un problème de test $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$ (i.e. $\Theta_0 = \{0\}$ et $\Theta_1 = \mathbb{R}^\star$). Pour une observation menant à une valeur estimée de θ non nulle, mais proche de 0, doit-on pour autant considérer que H_1 est vraie ? Pour assouplir la nature du test, on est amené à utiliser une statistique de test T prenant ses valeurs dans l'intervalle $[0,1]$.

Définition *Un test stochastique est une statistique T à valeurs dans $[0,1]$: pour l'observation $x \in \mathscr{H}^n$, $T(x)$ est la probabilité de rejeter H_0. La zone de rejet (resp. d'acceptation) du test est $T^{-1}(\{1\})$ (resp. $T^{-1}(\{0\})$. La zone d'hésitation du test est $T^{-1}(]0,1[)$.*

Par défaut, un test est considéré comme étant stochastique, et sa décision est rendue par un lancer de pièce ! Plus précisémment, examinons de quelle manière rendre une décision dans le cadre d'un tel test :

PROCÉDURE DE DÉCISION D'UN TEST. Soit T un test stochastique. Pour l'observation x, $T(x)$ est la probabilité de rejeter H_0. On réalise alors un tirage aléatoire dans $\{0,1\}$ selon une loi $\mathscr{B}(T(x))$: si le résultat du tirage est 0, on décide que H_0 est acceptée ; sinon, H_0 est rejetée.

5.2 Erreurs d'un test

Un test doit être construit à partir d'une erreur fixée au préalable. Le 1er type d'erreur que l'on peut dégager est la probabilité de rejeter H_0 à tort :

Définition *Soit T un test stochastique. Son risque (ou erreur) de 1ère espèce est l'application qui, à chaque $\theta \in \Theta_0$, donne la probabilité de rendre la mauvaise décision :*

$$\begin{array}{rcl} \Theta_0 & \to & [0,1] \\ \theta & \mapsto & \mathbb{E}_\theta T. \end{array}$$

On dit que le test est de niveau (resp. de seuil) α si la probabilité maximale de rejeter H_0 à tort, i.e. l'erreur de 1ère espèce maximale $\sup_{\theta\in\Theta_0} \mathbb{E}_\theta T$, est égale (resp. inférieure) à α.

Si le niveau du test est suffisamment proche de 0 (en pratique inférieur à 5%), la décision de rejeter H_0 est donc convaincante.

Exemple Considérons le modèle statistique $(\mathbb{R}^n, \{N(\theta,1)^{\otimes n}\}_{\theta\in\mathbb{R}})$. Pour un paramètre $\theta_0 \in \mathbb{R}$ fixé, on veut construire un test pur de niveau α pour le problème de test $H_0 : \theta \le \theta_0$ contre $H_1 : \theta > \theta_0$. Soit $\theta \in \mathbb{R}$ fixé, et $(X_1, \cdots, X_n)$ un échantillon de loi $P_\theta = N(\theta,1)^{\otimes n}$. On utilise la *statistique de test* $\sqrt{n}(\bar{X}_n - \theta)$ dont la loi est $N(0,1)$. Notons $z(\alpha)$ le quantile d'ordre $1-\alpha$ de la loi $N(0,1)$, et

$$R = \left\{(y_1, \cdots, y_n) \in \mathbb{R}^n : \sqrt{n}(\bar{y}_n - \theta_0) \ge z(\alpha)\right\}.$$

Alors, pour chaque $\theta \le \theta_0$:

$$\begin{array}{rcl} P_\theta(R) & = & P_\theta\left(\sqrt{n}(\bar{X}_n - \theta) + \sqrt{n}(\theta - \theta_0) \ge z(\alpha)\right) \\ & \le & P_\theta\left(\sqrt{n}(\bar{X}_n - \theta) \ge z(\alpha)\right) = \alpha, \end{array}$$

avec égalité lorsque $\theta = \theta_0$. Par suite, le test $T = \mathbf{1}_R$ est de niveau α.

Pour un test de niveau suffisamment proche de 0, la décision d'accepter H_0 peut être sujette à caution : le test nul, i.e. $T \equiv 0$, pour lequel H_0 est toujours choisie, possède un niveau nul. Un tel test n'est pas informatif, car il conclut toujours à accepter H_0, ceci même si elle n'est pas vraie. Cette observation nous amène à distinguer un autre type d'erreur, la probabilité d'accepter H_0 à tort :

Définition *Soit T un test stochastique. Son risque (ou erreur) de 2ème espèce est l'application qui, à chaque $\theta \in \Theta_1$, donne la probabilité de rendre la mauvaise décision :*

$$\begin{array}{rcl} \Theta_1 & \to & [0,1] \\ \theta & \mapsto & 1 - \mathbb{E}_\theta T. \end{array}$$

Comme l'erreur de 1ère espèce, l'erreur de 2ème espèce se doit d'être faible. Un autre concept équivalent est fréquemment utilisé, la probabilité d'accepter H_1 à raison.

Définition *Soit T un test stochastique. Sa puissance est l'application qui, à chaque $\theta \in \Theta_1$, donne la probabilité de rendre la bonne décision :*

$$\begin{array}{rcl} \Theta_1 & \to & [0,1] \\ \theta & \mapsto & \mathbb{E}_\theta T. \end{array}$$

Le test nul, qui possède un niveau nul, a en revanche un risque de 2ème espèce maximal (il vaut 1) et une puissance nulle. En général, diminuer l'erreur de 1ère espèce se fait au détriment de l'erreur de 2ème espèce, qui a alors tendance à augmenter. Il est donc important de s'orienter vers un compromis entre ces 2 types d'erreurs. De même que dans un procès aux assises, où le principe de présomption d'innocence du prévenu conduit l'avocat général à devoir étayer ses accusations de manière (quasi) irréfutable, le principe de présomption sur H_0 conduit à minimiser en priorité le niveau du test en imposant qu'il ne dépasse pas une valeur fixée. Puis, le test est construit de telle sorte que son erreur de 2ème espèce soit minimale. Cette démarche en deux temps porte le nom de *principe de Neyman*.

Exemple Reprenons le modèle statistique $(\mathbb{R}^n, \{N(\theta,1)^{\otimes n}\}_{\theta\in\mathbb{R}})$. Pour $\theta_0 \in \mathbb{R}$ fixé, on a construit un test pur de niveau α pour le problème de test $H_0 : \theta \leq \theta_0$

contre $H_1 : \theta > \theta_0$. Celui-ci est associé à la région de rejet

$$R = \left\{(y_1, \cdots, y_n) \in \mathbb{R}^n : \sqrt{n}(\bar{y}_n - \theta_0) \geq z(\alpha)\right\},$$

avec $z(\alpha)$ le quantile d'ordre $1-\alpha$ de la loi $N(0,1)$. Soit $\theta \in \mathbb{R}$ fixé, et $(X_1, \cdots, X_n)$ un échantillon de loi $P_\theta = N(\theta, 1)^{\otimes n}$. Si N est une variable aléatoire sur $(\Omega, \mathscr{F}, \mathbb{P})$ de loi $N(0,1)$, $\bar{X}_n$ et $\theta + N/\sqrt{n}$ ont même loi. Par suite,

$$\begin{aligned} P_\theta(R) &= P_\theta\left(\sqrt{n}(\bar{X}_n - \theta_0) \geq z(\alpha)\right) = \mathbb{P}\left(\sqrt{n}\left(\theta + \frac{1}{\sqrt{n}}N - \theta_0\right) \geq z(\alpha)\right) \\ &= \mathbb{P}\left(\sqrt{n}(\theta - \theta_0) + N \geq z(\alpha)\right). \end{aligned}$$

Si $T = \mathbf{1}_R$ est le test pur, la fonction puissance $\theta \mapsto \mathbb{E}_\theta T = P_\theta(R)$ définie sur $]\theta_0, \infty[$ est donc croissante, minorée par α et tend vers 1 lorsque θ tend vers l'infini.

Exemple Reprenons le modèle statistique $(\{0,1\}^n, \{\mathscr{B}(p)^{\otimes n}\}_{p\in]0,1[})$ de la section 1.1. Supposons que l'on veuille décider si oui ou non la pièce est équilibrée, en s'appuyant sur les observations $x_1, \cdots, x_n$ telles que $\bar{x}_n = 0.52$. Ces observations, qui sont régies par la loi $\mathscr{B}(p_0)$ nous indiquent que, si la pièce n'est pas équilibrée, l'alternative raisonnable est que $p_0 > 1/2$. On envisage donc de construire un test pur de $H_0 : p = 1/2$ contre $H_1 : p > 1/2$ au seuil 5%. Soit $t \in \mathbb{R}$ et une région de rejet du type :

$$R = \{(z_1, \cdots, z_n) \in \{0,1\}^n : \bar{z}_n > t\}.$$

Le test pur qui est associé à cette région de rejet est $T = \mathbf{1}_R$. Pour un échantillon $(X_1, \cdots, X_n)$ de la loi $P_{1/2} = \mathscr{B}(1/2)^{\otimes n}$:

$$\begin{aligned} \mathbb{E}_{1/2}T &= P_{1/2}(\bar{X}_n > t) \\ &= P_{1/2}\left(2\sqrt{n}\,(\bar{X}_n - 1/2) > 2\sqrt{n}\,(t - 1/2)\right) \\ &= 1 - F\left(2\sqrt{n}(t - 1/2)\right) + O(n^{-1/2}), \end{aligned}$$

si F est la fonction de répartition de la loi $N(0,1)$, en vertu de l'inégalité de Berry-Essèen. Les valeurs de la fonction de répartition de la loi $N(0,1)$ sont tabulées : on trouve alors, pour les valeurs de t telles que $2\sqrt{n}(t-1/2) \geq 1.64$ i.e. $t \geq 0.53$ car $n = 1000$, que

$$1 - F\left(2\sqrt{n}(t - 1/2)\right) \leq 5\%.$$

En négligeant le terme en $O(n^{-1/2})$, on obtient $\mathbb{E}_{1/2}T \leq 5\%$. Autrement dit, pour les régions de rejet :

$$R = \{(z_1, \cdots, z_n) \in \{0,1\}^n : \bar{z}_n > t\},$$

avec $t \geq 0.53$, le test $T = \mathbf{1}_R$ est de seuil 5%. Par ailleurs, la valeur $t = 0.53$ donne le test de puissance maximale. En conclusion, le test $T = \mathbf{1}_R$ avec

$$R = \{(z_1, \cdots, z_n) \in \{0,1\}^n : \bar{z}_n > t\},$$

est de seuil 5% et de puissance maximale. Avec la valeur de $\bar{x}_n = 0.52$, l'observation $(x_1, \cdots, x_n) \notin R$ c'est-à-dire qu'on est amené à accepter H_0 au niveau 5% : il est donc envisageable, au vu des observations, de considérer que la pièce est équilibrée.

5.3 Comparaison des tests

Pour un test T, une puissance trop faible signifie que l'on peut trouver dans Θ_1 un point θ pour lequel $\mathbb{E}_\theta T$ est faible. Lorsque cette dernière valeur est plus petite que le niveau du test, on se retrouve dans la situation paradoxale où la probabilité d'accepter H_1 à raison est plus petite que la probabilité d'accepter H_1 à tort ! Dans un tel contexte, le test ne sépare pas bien les hypothèses H_0 et H_1. La notion de *test sans biais* formalise cet écueil qu'il convient d'éviter.

Définition *Un test stochastique T de seuil α est dit sans biais si pour tout $\theta \in \Theta_1$, on a $\alpha \leq \mathbb{E}_\theta T$.*

Rien ne nous certifie, en général, qu'un test sans biais existe. Nous reviendrons sur ce problème crucial de la théorie des tests dans la section suivante.

Exemple Pour chaque $\theta \in \mathbb{R}$, on note Q_θ la loi de densité

$$\exp(-(x-\theta))\,\mathbf{1}_{[\theta,\infty[}(x).$$

On souhaite tester $H_0 : \theta \leq 0$ contre $H_1 : \theta > 0$ au niveau $\alpha \in]0,1[$, dans le modèle statistique $(\mathbb{R}^n, \{Q_\theta^{\otimes n}\}_{\theta\in\mathbb{R}})$. Le test $T = \mathbf{1}_R$ associé à la région de rejet

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

est un test pur pour H_0 contre H_1, de niveau α et sans biais. Pour $\theta \in \mathbb{R}$, notons

$P_\theta = Q_\theta^{\otimes n}$ et $(X_1, \cdots, X_n)$ un échantillon de loi P_θ. Si $\theta \leq 0$:

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

avec égalité si $\theta = 0$, i.e. le test T est de niveau α. De plus, si $\theta > 0$, on a :

$$\mathbb{E}_\theta T = \left[P_\theta\left(X_1 \geq -\frac{\ln\alpha}{n}\right)\right]^n = \left[\int_{\max(\theta,-\ln\alpha/n)}^{\infty} e^{-(t-\theta)}\mathrm{d}t\right]^n.$$

Selon que θ est plus grand ou plus petit que $-\ln\alpha/n$, $\mathbb{E}_\theta T$ vaut 1 ou $\alpha e^{n\theta}$. Comme $\theta > 0$, $\mathbb{E}_\theta T > \alpha$, et T est donc un test sans biais.

Définition *Soit $\alpha \in [0,1]$. On dit qu'un test T de seuil α est uniformément plus puissant parmis tous les tests de seuil α (UPPα) si, pour tout autre test T' de seuil α, on a $\mathbb{E}_\theta T \geq \mathbb{E}_\theta T'$ pour chaque $\theta \in \Theta_1$.*

La notion d'optimalité envisagée est claire, un test UPP étant de puissance maximale pour un niveau fixé. En revanche, la question plus délicate de la caractérisation des tests UPP fera l'objet de la section suivante. Examinons d'emblée quelques propriétés évidentes des tests UPP.

Proposition *Soit $\alpha \in [0,1]$. Un test T de seuil α et UPPα est sans biais.*

Preuve Soit T' le test tel que $T' \equiv \alpha$. Comme T est UPPα, pour tout $\theta \in \Theta_1$, on a $\mathbb{E}_\theta T \geq \mathbb{E}_\theta T' = \alpha$. Donc T est sans biais. □

Proposition *Soient $\alpha \in [0,1]$, T un test et ζ une statistique exhaustive. Alors $\mathbb{E}_\theta[T|\zeta]$ est un test de même puissance et niveau que T. En particulier, $\mathbb{E}_\theta[T|\zeta]$ est UPPα si T est UPPα.*

Preuve Il suffit de remarquer que, pour chaque $\theta \in \Theta$, $\mathbb{E}_\theta[T|\zeta]$ est une statistique indépendante de θ par exhaustivité de ζ et que $\mathbb{E}_\theta T = \mathbb{E}_\theta \mathbb{E}_\theta[T|\zeta]$. □

5.4 Optimalité dans les tests simples

Dans toute la section, on suppose que le modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est dominé par μ, et de vraisemblance L. On fixe aussi 2 paramètres $\theta_0 \neq \theta_1 \in \Theta$, et on s'intéresse au problème de *test simple* suivant :

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

Nous allons étudier, pour ce problème de test simple, des conditions nécessaires et suffisantes pour qu'un test soit UPP. Du fait de leur caractère fondateur dans toute la théorie des tests, et afin de faire mention de leurs auteurs, ces résultats sont regroupés sous la dénomination de "*lemme fondamental de Neyman-Pearson*".

On considère la famille des tests $\mathscr{T}$ suivante : $T \in \mathscr{T}$ si il existe $k \in \mathbb{R}_+$ et $\gamma : \mathscr{H}^n \to [0,1]$ mesurable tels que pour chaque $x \in \mathscr{H}^n$:

$$T(x) = \begin{cases} 1 & \text{si } L(x;\theta_1) > kL(x;\theta_0); \\ \gamma(x) & \text{si } L(x;\theta_1) = kL(x;\theta_0); \\ 0 & \text{si } L(x;\theta_1) < kL(x;\theta_0), \end{cases}$$

L'ensemble $\mathscr{T}$ s'appelle *famille des tests de Neyman-Pearson*. L'ensemble $\mathscr{T}_c$ est le sous-ensemble de $\mathscr{T}$ constitué des tests pour lesquels la fonctions γ est constante. Il convient de remarquer qu'un test de Neyman-Pearson associé à une fonction $\gamma \equiv 0$ est un test pur.

Il est essentiel de remarquer l'aspect *constructif* des résultats qui suivent, tous les tests considérés faisant partie de la famille $\mathscr{T}$.

Le 1er résultat est relatif à l'existence d'un test UPP. Il nous montre qu'il existe toujours un test de $\mathscr{T}_c$ de niveau donné.

Théorème *Soit* $\alpha \in]0,1[$.

1. *Il existe un test de* $\mathscr{T}_c$ *de niveau* α *;*

2. *Si un test de* $\mathscr{T}_c$ *est de niveau* α*, alors il est UPP*α.

Preuve

1. Un test $T \in \mathscr{T}_c$ associé aux paramètres k et γ est de niveau α si

$$\alpha = \mathbb{E}_{\theta_0} T = P_{\theta_0}\left(L(.;\theta_1) > kL(.;\theta_0)\right) + \gamma P_{\theta_0}\left(L(.;\theta_1) = kL(.;\theta_0)\right).$$

Il suffit donc de trouver $(k,\gamma) \in \mathbb{R}_+ \times [0,1]$ vérifiant l'égalité précédente. Comme $P_{\theta_0}(L(.;\theta_0) \neq 0) = 1$, on peut écrire :

$$P_{\theta_0}\left(\frac{L(.;\theta_1)}{L(.;\theta_0)} > k\right) + \gamma P_{\theta_0}\left(\frac{L(.;\theta_1)}{L(.;\theta_0)} = k\right) = \alpha. \tag{5.4.1}$$

Notons k_0 un réel qui vérifie

$$P_{\theta_0}\left(\frac{L(.;\theta_1)}{L(.;\theta_0)} > k_0\right) \leq \alpha \leq P_{\theta_0}\left(\frac{L(.;\theta_1)}{L(.;\theta_0)} \geq k_0\right).$$

Un tel réel existe car $t \mapsto P_{\theta_0}(L(.;\theta_1)/L(.;\theta_0) > t)$ est décroissante. Dans le cas où $P_{\theta_0}(L(.;\theta_1)/L(.;\theta_0) = k_0) = 0$, tout couple (k_0,γ) vérifie (5.4.1). Dans le cas contraire, le couple (k_0,γ_0) avec

$$\gamma_0 = \frac{\alpha - P_{\theta_0}\left(\frac{L(.;\theta_1)}{L(.;\theta_0)} > k_0\right)}{P_{\theta_0}\left(\frac{L(.;\theta_1)}{L(.;\theta_0)} = k_0\right)},$$

vérifie (5.4.1). Ainsi, il existe $T \in \mathscr{T}_c$ de niveau α.

2. Soit $T^\star \in \mathscr{T}_c$ un test de niveau α. On note (k,γ) les paramètres associés à $T^\star$ et, pour simplifier, on suppose que $\gamma \in]0,1[$. Soit T un test de seuil α. On a alors les inclusions :

$$\begin{aligned}
&\{T^\star - T > 0\} \subset \{T^\star > 0\} \subset \{L(.;\theta_1) \geq kL(.;\theta_0)\} \text{ car } \gamma > 0;\\
&\{T^\star - T < 0\} \subset \{T^\star < 1\} \subset \{L(.;\theta_1) \leq kL(.;\theta_0)\} \text{ car } \gamma < 1.
\end{aligned}$$

Par suite, pour tout $x \in \mathscr{H}^n$, $(T^\star(x) - T(x))(L(x;\theta_1) - kL(x;\theta_0)) \geq 0$, et donc

$$(T^\star(x) - T(x))L(x;\theta_1) \geq k(T^\star(x) - T(x))L(x;\theta_0). \tag{5.4.2}$$

On en déduit alors que

$$\begin{aligned}
\mathbb{E}_{\theta_1}T^\star - \mathbb{E}_{\theta_1}T &= \mathbb{E}_{\theta_1}(T^\star - T) = \int_{\mathscr{H}^n}(T^\star - T)L(.;\theta_1)\mathrm{d}\mu\\
&\geq k\int_{\mathscr{H}^n}(T^\star - T)L(.;\theta_0)\mathrm{d}\mu = k\left(\mathbb{E}_{\theta_0}T^\star - \mathbb{E}_{\theta_0}T\right).
\end{aligned}$$

Or, comme $T^\star$ est de niveau α et T de seuil α, $\mathbb{E}_{\theta_0}T^\star = \alpha \geq \mathbb{E}_{\theta_0}T$ d'où $\mathbb{E}_{\theta_1}T^\star \geq \mathbb{E}_{\theta_1}T$, i.e. $T^\star$ est UPPα. □

Le 2nd résultat, en nous montrant que la famille des tests de Neyman-Pearson est suffisamment riche, nous donne des conditions nécessaires pour qu'un test soit UPP.

Théorème *Soient $\alpha \in]0,1[$ et T un test UPPα. Il existe $T' \in \mathscr{T}$ tel que $T = T'$ μ-p.p.*

Preuve Soit $T^\star \in \mathscr{T}_c$ un test de niveau α et UPPα. On note $(k,\gamma) \in \mathbb{R}_+ \times [0,1]$ les paramètres associés au test $T^\star \in \mathscr{T}_c$. Pour simplifier, on suppose que $\gamma \in]0,1[$; dans ce cas, on a vu dans la preuve du théorème précédent (cf inégalité 5.4.2) que

$$R := (T^\star - T)(L(.;\theta_1) - kL(.;\theta_0)) \geq 0.$$

Par l'absurde, supposons que $\mu(R > 0) > 0$. Alors,

$$\int_{\mathscr{H}^n} R \mathrm{d}\mu = \int_{\{R>0\}} R \mathrm{d}\mu > 0$$

et, par suite :

$$\int_{\mathscr{H}^n} (T^\star - T) L(.;\theta_1) \mathrm{d}\mu > k \int_{\mathscr{H}^n} (T^\star - T) L(.;\theta_0) \mathrm{d}\mu.$$

Comme $T^\star$ est de niveau α et T est de seuil α,

$$\int_{\mathscr{H}^n} (T^\star - T) L(.;\theta_0) \mathrm{d}\mu = \mathbb{E}_{\theta_0} T^\star - \mathbb{E}_{\theta_0} T \geq 0,$$

ce qui montre que

$$\mathbb{E}_{\theta_1} T^\star - \mathbb{E}_{\theta_1} T = \int_{\mathscr{H}^n} (T^\star - T) L(.;\theta_1) \mathrm{d}\mu > 0.$$

Or, puisque T et $T^\star$ sont UPPα, $\mathbb{E}_{\theta_1} T^\star = \mathbb{E}_{\theta_1} T$ d'où la contradiction. Il s'ensuit que $\mu(R > 0) = 0$ soit, comme $R \geq 0$: $R = 0$ μ-p.p. Ainsi,

$$T = T^\star \ \mu\text{−p.p. sur } \{L(.;\theta_1) \neq kL(.;\theta_0)\}.$$

Définissons maintenant le test T' tel que pour $x \in \mathscr{H}^n$:

$$T'(x) = \begin{cases} T^\star(x) & \text{si } L(x;\theta_1) \neq kL(x;\theta_0); \\ T(x) & \text{si } L(x;\theta_1) = kL(x;\theta_0), \end{cases}$$

Alors, $T' \in \mathscr{T}$ et $T = T'$ μ-p.p., d'où le théorème. □

5.5 Optimalité dans les tests composites

Le contexte de la section précédente, en ne traitant que le cas d'un problème de test simple, est très restrictif. Néanmoins, il est possible de l'étendre au cas d'hypothèses dites *composites*. Soient $\Theta_0, \Theta_1 \subset \Theta$ avec $\Theta_0 \cap \Theta_1 = \emptyset$. Le problème de test que nous allons étudier est :

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

Puisque nous allons faire appel à des résultats du type Neyman-Pearson, nous supposons aussi que le modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est dominé par μ, et de vraisemblance L.

Théorème *Soit T un test de niveau $\alpha \in]0,1[$ tel qu'il existe $\theta_0 \in \Theta_0$ vérifiant $\mathbb{E}_{\theta_0} T = \alpha$. Si, pour tout $\theta_1 \in \Theta_1$, il existe un test $T_{\theta_1} \in \mathscr{T}_c$ de $H_0' : \theta = \theta_0$ contre $H_1' : \theta = \theta_1$ vérifiant $T = T_{\theta_1}$, alors T est UPPα.*

Preuve Fixons $\theta_1 \in \Theta_1$. Comme $\mathbb{E}_{\theta_0} T = \alpha$, pour le problème de test simple

$$H_0' : \theta = \theta_0 \text{ contre } H_1' : \theta = \theta_1,$$

le test T est de niveau α. Comme $T = T_{\theta_1} \in \mathscr{T}_c$, T est UPPα dans le problème de test de H_0' contre H_1'.

Soit maintenant $T^\star$ un test de H_0 contre H_1 de seuil α. Alors, $T^\star$ est de seuil α pour le problème de test de H_0' contre H_1' car

$$\mathbb{E}_{\theta_0} T^\star \leq \sup_{\theta\in\Theta_0} \mathbb{E}_\theta T^\star \leq \alpha.$$

Or, T est UPPα dans le problème de test de H_0' contre H_1', donc $\mathbb{E}_{\theta_1} T \geq \mathbb{E}_{\theta_1} T^\star$. Comme θ_1 a été choisi arbitrairement dans Θ_1, on en déduit que T est UPPα dans le problème de test de H_0 contre H_1. □

Exemple Reprenons le modèle statistique $(\mathbb{R}^n, \{N(\theta,1)^{\otimes n}\}_{\theta\in\mathbb{R}})$. On a vu que, dans le problème de test de $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$, le test $T = \mathbf{1}_R$ de région de rejet

$$R = \left\{(x_1, \cdots, x_n) \in \mathbb{R}^n : \sqrt{n}(\bar{x}_n - \theta_0) > z(\alpha)\right\},$$

où $z(\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi $N(0,1)$, est un test de niveau α. Nous allons montrer que ce test est UPPα en utilisant le théorème précédent.

On remarque tout d'abord que $\mathbb{E}_{\theta_0} T = P_{\theta_0}(R) = \alpha$. Fixons maintenant $\theta_1 > \theta_0$. Pour tout $\theta \in \mathbb{R}$ et $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$, on a l'écriture

$$L(x;\theta) = \left\{ \exp\left(-\frac{n}{2}(\bar{x}_n - \theta)^2 \right) \right\} \left\{ \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right) \right\}.$$

On en déduit la forme suivante pour le *rapport des vraisemblances* :

$$\begin{aligned} \frac{L(x;\theta_1)}{L(x;\theta_0)} &= \exp\left[-\frac{n}{2} \left((\bar{x}_n - \theta_1)^2 - (\bar{x}_n - \theta_0)^2 \right) \right] \\ &= \exp\left[\sqrt{n}(\theta_1 - \theta_0) \left(\sqrt{n}(\bar{x}_n - \theta_0) - \frac{\sqrt{n}}{2}(\theta_1 - \theta_0) \right) \right]. \end{aligned}$$

Par suite, pour tout $k > 0$:

$$\frac{L(x;\theta_1)}{L(x;\theta_0)} > k \quad \Longleftrightarrow \sqrt{n}(\bar{x}_n - \theta_0) > \frac{\ln k}{\sqrt{n}(\theta_1 - \theta_0)} + \frac{\sqrt{n}}{2}(\theta_1 - \theta_0).$$

Choisissons maintenant $k_0 > 0$ tel que

$$z(\alpha) = \frac{\ln k_0}{\sqrt{n}(\theta_1 - \theta_0)} + \frac{\sqrt{n}}{2}(\theta_1 - \theta_0),$$

et notons T_{θ_1} le test de $\mathscr{T}_c$ associé aux paramètres $(k_0, 0)$, i.e.

$$T_{\theta_1} = \mathbf{1}_{\{L(.;\theta_1) > k_0 L(.;\theta_0)\}}.$$

On a alors $T = T_{\theta_1}$. D'après le théorème précédent, T est donc UPPα.

5.6 Tests asymptotiques

Comme les lois à distance finie ne sont pas toujours évidentes à obtenir, on est amené, à l'instar des intervalles de confiance asymptotiques, à définir la notion de *test asymptotique*.

On considère le problème de test de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$, avec $\Theta_0, \Theta_1 \subset \Theta$ et $\Theta_0 \cap \Theta_1 = \emptyset$. Le modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta \in \Theta})$ dépend de n :

dans le cadre des tests asymptotiques, on fait donc apparaître la taille n de l'échantillon dans la notation du test.

Définition *Un test asymptotique de seuil* $\alpha \in]0,1[$ *est la donnée d'une suite de tests* $(T_n)_n$ *tels que*

$$\sup_{\theta \in \Theta_0} \limsup_n \mathbb{E}_\theta T_n \leq \alpha.$$

La procédure de décision est alors calquée sur celle des tests à taille d'échantillon finie. La seule différence notable est qu'un test asymptotique est construit pour contrôler l'erreur de 1ère espèce, mais seulement asymptotiquement.

Définition *Un test asymptotique* $(T_n)_n$ *est dit convergent si*

$$\forall \theta \in \Theta_1 \quad : \quad \lim_n \mathbb{E}_\theta T_n = 1.$$

Chapitre 6

Statistique des échantillons gaussiens

L'étude statistique des échantillons gaussiens est basée sur 2 résultats fondamentaux portant sur la nature particulière de la projection vecteurs gaussiens. Dans tout ce chapitre, $N_d(m,\Sigma)$ désigne une loi gaussienne sur $\mathbb{R}^d$, de moyenne $m \in \mathbb{R}^d$ et de matrice de variance $\Sigma \in \mathscr{M}_d(\mathbb{R})$.

6.1 Projection de vecteurs gaussiens

Toutes les variables aléatoires de cette section sont implicitement définies sur un espace probabilisé $(\Omega, \mathscr{F}, \mathbb{P})$.

Le théorème ci-dessous est essentiel dans toute la théorie des modèles gaussiens. On rappelle que la *loi de Chi 2* à d degrés de liberté, notée χ^2_d, est la loi de la somme des carrés de d v.a.r.i.i.d. de lois $N_1(0,1)$. Par ailleurs, $\|.\|$ désigne toujours la norme euclidienne.

Théorème [COCHRAN] *Soit $X \sim N_n(0, \sigma^2 \mathrm{Id})$ avec $\sigma > 0$, et $L_1 \oplus \cdots \oplus L_p$ une décomposition de $\mathbb{R}^n$ en sous-espaces orthogonaux de dimensions $r_1, \cdots, r_p$. Les projections orthogonales $\pi_1, \cdots, \pi_p$ de X sur $L_1, \cdots, L_p$ sont des vecteurs gaussiens indépendants, et pour chaque $i = 1, \cdots, p$:*

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

Preuve Soit $(e^i_j)_{i,j}$ une base orthonormée de $\mathbb{R}^n$ telle que pour chaque $i = 1, \cdots, p$, $(e^i_j)_{j=1,\cdots,r_i}$ est une base orthonormée de L_i. Pour chaque $i = 1, \cdots, p$, on a :

$$\pi_i = \sum_{j=1}^{r_i} (X^T e^i_j) e^i_j.$$

Les vecteurs $(e^i_j)_{i,j}$ étant orthogonaux, pour tout $i \neq k$, la matrice de covariance entre π_i et π_k, i.e.

$$\mathrm{cov}(\pi_i, \pi_k) = \mathbb{E}\,(\pi_i - \mathbb{E}\pi_i)\,(\pi_k - \mathbb{E}\pi_k)^T = \mathbb{E}\pi_i \pi_k^T = 0.$$

Comme $(\pi_1 \cdots \pi_p)^T$ est un vecteur gaussien (toute combinaison linéaire des v.a.r. $(X^T e^i_j)_{i,j}$ est gaussienne), $\pi_1, \cdots, \pi_p$ sont donc des vecteurs gaussiens indépendants, d'où le premier point.

Fixons $i = 1, \cdots, p$, et calculons tout d'abord, pour tout $j = 1, \cdots, r_i$, la loi de la v.a.r. $X^T e^i_j$. Il est clair que $X^T e^i_j$ est une v.a.r. gaussienne centrée, comme combinaison linéaire des composantes d'un vecteur gaussien centré. De plus, comme les composantes du vecteur $X = (X_1 \cdots X_n)^T$ sont i.i.d. de loi $N_1(0, \sigma^2)$,

$$\mathrm{var}(X^T e^i_j) = \sum_{k=1}^{n} \mathrm{var}(X_k)(e^i_j(k))^2 = \sigma^2 \|e^i_j\|^2 = \sigma^2,$$

où l'on a noté $e^i_j = (e^i_j(1) \cdots e^i_j(n))^T$. Par suite, $X^T e^i_j \sim N_1(0, \sigma^2)$. D'autre part, comme le vecteur aléatoire $(X^T e^i_1 \cdots X^T e^i_{r_i})^T$ est gaussien (car toute combinaison linéaire de ses composantes est une v.a.r. gaussienne), il suffit de montrer que pour tout $j \neq j'$, $\mathrm{cov}(X^T e^i_j, X^T e^i_{j'}) = 0$ pour en déduire que $X^T e^i_1, \cdots, X^T e^i_{r_i}$ sont indépendantes. Or, si $j \neq j'$:

$$\begin{aligned} \mathrm{cov}(X^T e^i_j, X^T e^i_{j'}) &= \mathbb{E}(X^T e^i_j)(X^T e^i_{j'}) = \sum_{k,k'=1}^{n} \mathbb{E}(X_k X_{k'}) e^i_j(k) e^i_{j'}(k') \\ &= \sum_{k=1}^{n} \mathbb{E}(X_k^2) e^i_j(k) e^i_j(k) = \sigma^2 (e^i_j)^T e^i_{j'} = 0. \end{aligned}$$

Nous avons donc montré que les v.a.r. $(X^T e^i_j / \sigma^2)_j$ sont i.i.d., de même loi $N_1(0,1)$. Par suite,

$$\frac{1}{\sigma^2}\|\pi_i\|^2 = \sum_{j=1}^{r_i} \left(\frac{X^T e^i_j}{\sigma}\right)^2 \sim \chi^2_{r_i},$$

d'où le théorème. □

La *loi de Student* à n degrés de liberté, notée T_n, est la loi du quotient $\sqrt{n}X/\sqrt{Y}$, où $X \perp\!\!\!\perp Y$, $X \sim N_1(0,1)$ et $Y \sim \chi^2_n$.

Théorème [FISHER] *Soient* $X = (X_1, \cdots, X_n)^T \sim N_n(\bar{m}, \sigma^2 \mathrm{Id})$ *et* $\bar{m} = (m, \cdots, m)^T$ *avec* $\sigma > 0$ *et* $m \in \mathbb{R}$. *On note*

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

Alors,

(i) $\bar{X}_n \perp\!\!\!\perp S_n$;
(ii) $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$;
(iii) $\sqrt{n}(\bar{X}_n - m)/S_n \sim T_{n-1}$.

Remarques

(a) Le résultat en (iii) est à comparer au résultat classique : $\sqrt{n}(\bar{X}_n - m)/\sigma \sim N_1(0,1)$.

(b) D'après la loi forte des grands nombres, $S_n \to \sigma$ p.s. Par suite, l'assertion (iii), le théorème de la limite centrale unidimensionnel et le lemme de Slutsky montrent que T_n converge en loi vers la loi $N_1(0,1)$.

Preuve Pour simplifier, on considère le cas $m = 0$ et $\sigma = 1$. Soit L le s.e.v. de $\mathbb{R}^n$ engendré par $e = (1, \cdots, 1)^T$. Le projecteur orthogonal P sur L est la matrice $n \times n$ dont tous les coefficients valent $1/n$. On a alors $PX = \bar{X}_n e$ et $(\mathrm{Id} - P)X = (X_1 - \bar{X}_n, \cdots, X_n - \bar{X}_n)^T$. Comme $(\mathrm{Id} - P)X$ est la projection orthogonale de X sur l'orthogonal de L, on déduit du théorème de Cochran que $PX \perp\!\!\!\perp (\mathrm{Id} - P)X$, et en particulier que $\bar{X}_n \perp\!\!\!\perp S_n^2$, d'où (i). De plus, $(n-1)S_n^2 = \|(\mathrm{Id} - P)X\|^2 \sim \chi^2_{n-1}$ d'après le théorème de Cochran, d'où (ii). Enfin, (iii) est conséquence du fait que $\sqrt{n}(\bar{X}_n - m)/\sigma$ et $(n-1)S_n^2/\sigma^2$ sont indépendantes, et de lois respectives $N_1(0,1)$ et χ^2_{n-1}. □

6.2 Tests sur les paramètres

On se donne dans cette partie un modèle statistique $(\mathbb{R}^n, \{N_1(m, \sigma^2)^{\otimes n}\}_{m \in \mathbb{R}, \sigma > 0})$. Le but est de construire des tests ou des intervalles de confiance sur la valeur des paramètres m_0 et σ_0^2 d'un échantillon $x_1, \cdots, x_n$ issu de la loi $N_1(m_0, \sigma_0^2)$. Comme

on l'a vu dans les chapitres précédents, il faut alors construire une statistique dont la loi ne dépend pas des paramètres inconnus du modèle.

Notons $(X_1, \cdots, X_n)$ un échantillon de loi $P_{m,\sigma} = N_1(m, \sigma^2)^{\otimes n}$. On sait alors que

$$\sqrt{n}\frac{\bar{X}_n - m}{\sigma^2} \sim N_1(0,1).$$

Cependant, cette statistique, en faisant intervenir simultanément les 2 paramètres inconnus m et σ, n'est pas utilisable. On se tourne alors vers le théorème de Fisher, qui nous donne les égalités en loi :

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1} \text{ et } \sqrt{n}\frac{\bar{X}_n - m}{S_n} \sim T_{n-1}.$$

L'utilisation de ces statistiques permet de construire facilement des intervalles de confiance pour les valeurs de m_0 et σ_0, à partir des valeurs observées $x_1, \cdots, x_n$.

Considérons par exemple le problème de test $H_0 : m \geq m_1$ contre $H_1 : m < m_1$ au niveau α, avec m_1 un réel fixé. Si $t_{n-1}(\alpha)$ est le quantile d'ordre α de la loi T_{n-1}, on a sous H_0 :

$$\begin{aligned} P_{m,\sigma}\left(\bar{X}_n < m_1 + t_{n-1}(\alpha)\frac{S_n}{\sqrt{n}}\right) &\geq P_{m,\sigma}\left(\bar{X}_n < m + t_{n-1}(\alpha)\frac{S_n}{\sqrt{n}}\right) \\ &= P_{m,\sigma}\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < t_{n-1}(\alpha)\right) = \alpha. \end{aligned}$$

Notons pour chaque $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$,

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i \quad \text{et} \quad s_n^2(y) = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y}_n)^2.$$

Le *test de Student* est le test pur de région de rejet

$$R_{\text{moy}} = \left\{ y = (y_1, \cdots, y_n) \in \mathbb{R}^n : \bar{y}_n < m_1 + t_{n-1}(\alpha)\frac{s_n(y)}{\sqrt{n}} \right\}.$$

Ce test est de niveau α, et la procédure de décision est définie ainsi : on accepte H_0 au niveau α si $(x_1 \cdots x_n)^T \notin R_{\text{moy}}$.

Etudions maintenant le problème de test de $H_0 : \sigma \geq \sigma_1$ contre $H_1 : \sigma < \sigma_1$ au niveau α, avec $\sigma_1 > 0$ fixé. Si $\chi_{n-1}(\alpha)$ est le quantile d'ordre α de la loi χ^2_{n-1}, on a sous H_0 :

$$\begin{aligned} P_{m,\sigma}\left(S_n^2 < \frac{\chi_{n-1}(\alpha)}{n-1}\sigma_1^2\right) &\geq P_{m,\sigma}\left(S_n^2 < \frac{\chi_{n-1}(\alpha)}{n-1}\sigma^2\right) \\ &= P_{m,\sigma}\left((n-1)\frac{S_n^2}{\sigma^2} < \chi_{n-1}(\alpha)\right) = 1-\alpha. \end{aligned}$$

Le *test de Fisher* est le test pur de région de rejet

$$R_{\text{var}} = \left\{ y = (y_1, \cdots, y_n) \in \mathbb{R}^n : s_n^2(y) < \frac{\chi_{n-1}(\alpha)}{n-1}\sigma_1^2 \right\}.$$

Ce test est de niveau α, et la procédure de décision est définie ainsi : on accepte H_0 au niveau α si $(x_1 \cdots x_n)^T \notin R_{\text{var}}$.

6.3 Comparaison de 2 échantillons

On suppose dans cette partie que l'on a 2 suites indépendantes d'observations indépendantes $x = (x_1, \cdots, x_n)$ et $y = (y_1, \cdots, y_p)$, chacune issue de l'une des lois des modèles statistiques $\{N_1(m, \sigma^2)^{\otimes n}\}_{m\in\mathbb{R},\sigma>0}$ et $\{N_1(m, \sigma^2)^{\otimes p}\}_{m\in\mathbb{R},\sigma>0}$. On suppose que ces suites d'observations ont même variance (c'est l'hypothèse dite d'*homoscédasticité*), et on veut construire un test pur portant sur l'égalité des moyennes des suites x et y.

Si m_1 et m_2 représentent les moyennes de chaqun des 2 échantillons, le problème de test s'exprime donc $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$, dont nous allons construire un test pur au niveau α. Notons X un échantillon $(X_1, \cdots, X_n)$ de la loi $N_1(m_1, \sigma^2)^{\otimes n}$ et Y un échantillon $(Y_1, \cdots, Y_p)$ de la loi $N_1(m_1, \sigma^2)^{\otimes p}$. Compte tenu des hypothèses expérimentales, on peut supposer que X et Y sont indépendantes. De plus, $S_n^2(X)$ et $S_p^2(Y)$ désignent les variances empiriques sans biais de X et Y.

Introduisons la statistique

$$Q = \frac{(\bar{X}_n - \bar{Y}_p) - (m_1 - m_2)}{\sqrt{\frac{1}{n} + \frac{1}{p}}}.$$

Puisque X^T et Y^T sont 2 vecteurs gaussiens indépendants, Q est une v.a.r. gaussienne, comme combinaison linéaire d'un vecteur gaussien. Il est clair que Q est centrée, et on montre facilement que la variance de Q est σ^2. En conséquence, $Q \sim N_1(0,\sigma^2)$. Cependant, σ est en général un paramètre inconnu, donc la statistique Q n'est pas utilisable directement pour construire un test statistique.

Notons alors

$$W^2 = (n-1)S_n^2(X) + (p-1)S_p^2(Y).$$

D'après le théorème de Fisher, $(n-1)S_n^2(X) \sim \sigma^2\chi^2_{n-1}$ et $(p-1)S_p^2(Y) \sim \sigma^2\chi^2_{p-1}$. Comme, par ailleurs, $S_n^2(X) \perp\!\!\!\perp S_p^2(Y)$, on a donc $W^2 \sim \sigma^2\chi^2_{n+p-2}$. De plus, $W \perp\!\!\!\perp Q$ d'après le théorème de Fisher. Par définition de la loi de Student, on a donc

$$M = \sqrt{n+p-2}\,\frac{Q}{W} \sim T_{n+p-2}.$$

Comme la loi de M est *libre*, i.e. elle ne dépend pas de paramètres inconnus, la statistique de test à utiliser est M. Désignons par $t_{n+p-2}(\alpha)$ le quantile d'ordre $1-\alpha/2$ de la loi T_{n+p-2}. En utilisant le fait que la loi de Student est symétrique, on vérifie comme dans la section précédente que, avec des notations évidentes, l'ensemble

$$\left\{ (xy)^T \in \mathbb{R}^{n+p} \,:\, \frac{|\bar{x}_n - \bar{y}_p|}{\sqrt{(n-1)s_n^2(x) + (p-1)s_p^2(y)}} \geq \sqrt{\frac{\frac{1}{n}+\frac{1}{p}}{n+p-2}}\, t_{n+p-2}(\alpha) \right\}$$

est une région de rejet pour tester H_0 contre H_1, au niveau α.

Supposons maintenant que l'on veuille tester l'égalité des moyennes dans 3 échantillons gaussiens indépendants. On peut bien sûr reprendre la méthodologie précédente, et réaliser 2 tests d'égalité de moyenne. Mais alors, le niveau du test global ainsi construit est de l'ordre de la somme des niveaux des 2 tests. Pour éviter cette perte de niveau, il faut adopter une démarche radicalement différente, comme nous allons le constater dans la section qui suit.

6.4 Modèle linéaire gaussien

6.4.1 Le problème et sa formulation vectorielle

On suppose dans cette section que l'on dispose de k jeux indépendants d'observations indépendantes $x^1, \cdots, x^k$. On est encore dans le cadre d'un modèle gaussien,

car pour tout i, x^i est une observation du modèle statistique $\{N_1(m,\sigma^2)^{\otimes n_i}\}_{m\in\mathbb{R},\sigma>0}$. Comme dans la section précédente, on impose l'hypothèse d'homoscédasticité du modèle, i.e. les variances de chacun des jeux d'observations sont les mêmes. L'objectif est de construire un test pur portant sur l'égalité des moyennes de ces k jeux d'observations.

Sous l'hypothèse d'homoscédasticité, on peut introduire les échantillons indépendants $X_1 \sim N_1(m_1,\sigma^2)^{\otimes n_1},\cdots,X_k \sim N_1(m_k,\sigma^2)^{\otimes n_k}$ pour construire la statistique de test. Le problème de test s'exprime donc par

$$H_0 : m_1 = \cdots = m_k \quad \text{contre} \quad H_1 : \text{il existe } i \neq j \text{ tel que } m_i \neq m_j.$$

Dans cette formulation, $m_1,\cdots,m_k$ sont des paramètres réels et $\sigma > 0$.

Soit $n = n_1 + \cdots + n_k$, $n_0 = 0$ et, pour chaque $i = 1,\cdots,k$,

$$I_i = \sum_{j=n_1+\cdots+n_{i-1}+1}^{n_1+\cdots+n_i} e_j,$$

où, pour tout $j = 1,\cdots,n$, e_j est le j-ème vecteur de la base canonique de $\mathbb{R}^n$. Notons alors

$$\mu = \sum_{i=1}^{k} m_i I_i,$$

E l'espace vectoriel engendré par les vecteurs $I_1,\cdots,I_k$, et H le sous-espace vectoriel de $\mathbb{R}^n$ engendré par le vecteur $(1\cdots1)^T$. Avec cette écriture, le problème de test s'énonce ainsi :

$$H_0 : \mu \in H \quad \text{contre} \quad H_1 : \mu \in E \setminus H.$$

6.4.2 Statistique de test

Dans la suite, z_F désigne la projection orthogonale de $z \in \mathbb{R}^n$ sur le sous-espace vectoriel F. Si $X = (X_1\cdots X_k)^T$, on a la décomposition :

$$X = \mu + \varepsilon,$$

où $\varepsilon \sim N_n(0,\mathrm{Id})$. Cette formulation porte le nom de *modèle linéaire gaussien*. Dans ce cadre, on observe que :

- ▷ $X_E = \mu + \varepsilon_E$ car $\mu \in E$. En particulier, $X_E - \mu$ est la projection orthogonale de ε sur E ;
- ▷ $X - X_E = \varepsilon - \varepsilon_E$ est la projection orthogonale de ε sur l'orthogonal de E. Cette quantité ne contient pas d'information sur la valeur de μ, mais elle contient des informations sur la dispersion des observations.

En exploitant ces constatations, on obtient directement avec le théorème de Cochran :

Proposition

(i) X_E *est un estimateur sans biais de* μ *;*
(ii) $X_E \perp\!\!\!\perp X - X_E$ *;*
(iii) $\|X - X_E\|^2 \sim \sigma^2 \chi^2_{n-k}$. *En particulier,* $\|X - X_E\|^2/(n-k)$ *est un estimateur sans biais de* σ^2 *;*
(iv) $\|X_E - \mu\|^2 \sim \sigma^2 \chi^2_k$.

Sous H_0, $X_H = \mu + \varepsilon_H$ et donc $X_E - X_H = \varepsilon_E - \varepsilon_H$. Le théorème de Cochran appliqué au vecteur gaussien ε nous montre alors que

$$\|X_E - X_H\|^2 \sim \sigma^2 \chi^2_{k-1}, \text{ et } \quad X_E = \varepsilon - \varepsilon_E \perp\!\!\!\perp X_E - X_H.$$

La *loi de Fisher* de paramètres (i, j), notée $F(i, j)$, est définie comme suit :

$$F(i,j) \sim \frac{j}{i} \frac{U}{V}, \text{ si } \quad U \perp\!\!\!\perp V, \text{ et } \quad U \sim \chi^2_i, \quad V \sim \chi^2_j.$$

D'après la proposition précédente et les observations ci-dessus, sous H_0, on connaît donc la loi de la statistique

$$F = \frac{n-k}{k-1} \frac{\|X_E - X_H\|^2}{\|X - X_E\|^2} \sim F(k-1, n-k).$$

Pour construire la région de rejet, on observe que, si P désigne la loi de X, on a sous H_0,

$$P(F \geq f(\alpha)) = \alpha,$$

si $f(\alpha)$ désigne le quantile d'ordre $1-\alpha$ de la loi $F(k-1, n-k)$. La région de rejet

$$R = \left\{ z \in \mathbb{R}^n : \frac{n-k}{k-1} \frac{\|z_E - z_H\|^2}{\|z - z_E\|^2} \geq f(\alpha) \right\}$$

défini donc un test pur de H_0 contre H_1, au niveau α.

Concaténons les jeux d'observations $x^1, \cdots, x^k$ pour obtenir un vecteur x de $\mathbb{R}^n$. Plus précisémment, $x = (x_1 \cdots x_n)^T$ est le vecteur de $\mathbb{R}^n$ tel que

$$x = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x^i(j)\, e_{n_1+\cdots+n_{i-1}+j},$$

si, pour chaque $i = 1, \cdots, k$, $x^i = (x^i(1), \cdots, x^i(n_i))^T$. La procédure de décision s'énonce alors ainsi : on accepte H_0 au niveau α si $x \notin R$.