Introduction à la statistique mathématique

Notes de cours pour le master
Version préliminaire

Laurent Serlet

Juin 2007

Table des matières

Chapitre 1

L'échantillonage

L'objectif de la statistique mathématique est principalement d'aider à établir une décision à partir d'un échantilon d'une population trop vaste pour être étudiée en totalité.

On étudie une population comprenant un nombre fini N d'individus discernables au niveau d'un caractère qui varie selon les individus dans un certain ensemble, typiquement $\mathbb{R}$. Le recensement des valeurs de ce caractère sur tous les N individus étant trop couteux, on procède au tirage d'un échantillon dans cette population. On peut procéder de façon aléatoire ou non aléatoire, voire une combinaison des deux. Par exemple, si on sait partitionner la population en catégories qui sont homogènes vis à vis d'un critère, on peut établir un plan de sondage pour extraire un échantillon représentatif de la population. Ces méthodes qui exploitent des renseignements recueillis sur la population sont étudiées en théorie des sondages qui n'est pas l'objet de ce cours.

On va s'interesser aux tirages aléatoires successifs dans la population toute entière en examinant le cas du tirage avec remise et celui du tirage sans remise. Dans chacun de ces deux cas on supposera que tous les individus dans la population où on effectue un tirage ont la même probabilité d'être tirés. On procède à n tirages successifs.

On s'interessera particulièrement à la moyenne et à la variance du caractère sur l'échantillon tiré appelées moyenne et variance empirique.

1.1 Tirage sans remise dans une population finie

On tire n individus, sans remise, dans une population ayant au total N individus. Notons $(x_i ; 1 \leq i \leq N)$ les valeurs du caractère (supposé réel) sur les N individus formant la population totale. Notons ε_i la variable

aléatoire qui vaut 1 si l'individu numéro i fait partie des n individus tirés. En théorie des sondages, une telle variable est appelée variable Cornfield. La moyenne empirique relative à cet échantillon est :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{N} x_i\ \varepsilon_i$$

Les variables $(\varepsilon_i)_{1\leq i\leq N}$ ne sont pas indépendantes. Leurs lois (marginales) sont données par :

$$\mathbb{P}(\varepsilon_i = 1) = 1 - \mathbb{P}(\varepsilon_i = 0) = \frac{n}{N} = \frac{C_{N-1}^{n-1}}{C_N^n}$$

C'est à dire la loi de Bernoulli de paramètre $\frac{n}{N}$. Quelle est alors l'espérance de la moyenne empirique? Elle vaut

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{N} \mathbb{E}(\varepsilon_i)\ x_i = \frac{1}{N}\sum_{i=1}^{N} x_i$$

que l'on notera m et qui représente la valeur moyenne du caractère sur la population toute entière. Quelle est la variance de $\overline{X}_n$?

$$\begin{aligned}\mathbb{V}(\overline{X}_n) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{N} x_i\varepsilon_i - m\right)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{N}(x_i - m)\varepsilon_i\right)^2 \qquad \text{car } \sum_{i=1}^{N}\varepsilon_i = n \\ &= \frac{1}{n^2}\sum_{i=1}^{N}\mathbb{E}(\varepsilon_i^2)(x_i - m)^2 + \frac{1}{n^2}\sum_{1\leq i\neq j\leq N}(x_i - m)(x_j - m)\mathbb{E}(\varepsilon_i\varepsilon_j)\end{aligned}$$

Or

$$\mathbb{E}(\varepsilon_i\varepsilon_j) = \mathbb{P}(\varepsilon_i\varepsilon_j = 1) = \frac{C_{N-2}^{n-2}}{C_N^n} = \frac{(N-2)!}{(n-2)!(N-n)!}.\frac{n!(N-n)!}{N!} = \frac{n(n-1)}{N(N-1)}$$

si $i \neq j$ et toujours $\mathbb{E}(\varepsilon_i^2) = \mathbb{E}(\varepsilon_i) = \frac{n}{N}$. On obtient donc :

$$\mathbb{V}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{N}\frac{n}{N}(x_i - m)^2 + \frac{1}{n^2}\frac{n(n-1)}{N(N-1)}\sum_{1\leq i\neq j\leq N}(x_i - m)(x_j - m).$$

Pour simplifier cette expression, remarquons que

$$\begin{aligned}0 &= \frac{1}{N}\sum_{i=1}^{N} x_i - m = \frac{1}{N}\sum_{i=1}^{N}(x_i - m) = \left(\sum_{i=1}^{N}(x_i - m)\right)^2 \\ &= \sum_{1\leq i\neq j\leq N}(x_i - m)(x_j - m) + \sum_{i=1}^{N}(x_i - m)^2.\end{aligned}$$

On trouve ainsi :

$$\mathbb{V}(\overline{X}_n) = \sum_{i=1}^{N}(x_i - m)^2 . \left(\frac{1}{nN} - \frac{n-1}{nN(N-1)}\right) = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(x_i - m)^2$$

Notons σ^2 la variance de la famille déterministe $(x_i\,;\, 1 \leq i \leq N)$ définie par

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2.$$

On conclut que :

$$\mathbb{V}(\overline{X}_n) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

Une autre grandeur notable pour l'échantillon est la variance empirique :

$$S'^2_n = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

où $\{X_1,...,X_n\} = \{x_i\,;\, \varepsilon_i = 1\}$ est une énumération de l'échantillon tiré. On a aussi

$$S'^2_n = \frac{1}{n}\sum_{j=1}^{n}(X_j^2 + \overline{X}_n^2 - 2\overline{X}_n X_j) = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}_n^2,$$

en notant que $\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. Cela s'écrit encore ou encore : $S'^2_n = \frac{1}{n}\sum_{i=1}^{N} x_i^2\varepsilon_i - \overline{X}_n^2$. Son espérance est :

$$\begin{aligned}\mathbb{E}(S'^2_n) &= \frac{1}{n}\sum_{i=1}^{N} x_i^2\mathbb{E}(\varepsilon_i) - \mathbb{E}(\overline{X}_n^2)\\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mathbb{E}(\overline{X}_n^2)\end{aligned}$$

Mais

$$\mathbb{E}(\overline{X}_n^2) = \mathbb{V}(\overline{X}_n) + (\mathbb{E}(\overline{X}_n))^2 = \frac{N-n}{N-1}\frac{\sigma^2}{n} + m^2$$

Or

$$\frac{1}{N}\sum_{i=1}^{N} x_i^2 = \sigma^2 + m^2$$

d'où finalement

$$\mathbb{E}(S'^2_n) = \frac{N(n-1)}{(N-1)n}\sigma^2.$$

Nous laissons le calcul de $\mathbb{V}(S'^2_n)$ au lecteur courageux.

1.2 Tirage avec remise dans une population finie

Procédons à n tirages successifs avec remise dans une population totale de N individus. Notons X_1,...,X_n les valeurs du caractère obtenues pour ces n individus. Cette notation sous entend qu'un ordre sur les tirages existe mais cela n'a pas d'importance pour ce qui suit. Avec ce mode de tirage les variables X_1,...,X_n sont indépendantes et de même loi donnée pour $x \in \{x_i\,;\, 1 \leq i \leq N\}$ par :

$$\mathbb{P}(X_j = x) = \frac{1}{N}\#\{i \leq N; x_i = x\}$$

Notons comme précédemment :

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j\,; \quad S'^2_n = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}_n^2.$$

On trouve cette fois, puisque les X_j sont i.i.d.

$$\begin{aligned}\mathbb{E}(\overline{X}_n) &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}(X_j) = \mathbb{E}(X_1) = \sum_x x.\frac{1}{N}\#\{i \leq N; x_i = x\} \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i = m\end{aligned}$$

Par ailleurs,

$$\mathbb{V}(\overline{X}_n) = \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{V}(X_j) = \frac{1}{n}\mathbb{V}(X_1)$$

et, avec les notations du paragraphe précédent,

$$\begin{aligned}\mathbb{V}(X_1) &= \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = \sum_x x^2\frac{1}{N}\#\{i \leq N; x_i = x\} - m^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - m^2 = \sigma^2\end{aligned}$$

et on conclut que $V(\overline{X}_n) = \frac{\sigma^2}{n}$. Passons à la variance empirique :

$$\mathbb{E}(S'^2_n) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}(X_j^2) - \mathbb{E}(\overline{X}_n^2) = \mathbb{E}(X_1^2) - \mathbb{E}(\overline{X}_n^2)$$

Mais

$$\begin{aligned}\mathbb{E}(\overline{X}_n^2) &= \frac{1}{n^2}\sum_{1\leq i,j\leq n}\mathbb{E}(X_iX_j) = \frac{1}{n}\mathbb{E}(X_1^2) + \frac{n^2-n}{n^2}\mathbb{E}(X_1X_2) \\ &= \frac{1}{n}\mathbb{E}(X_1^2) + \frac{n^2-n}{n^2}\mathbb{E}(X_1)\mathbb{E}(X_2)\end{aligned}$$

On obtient

$$\begin{aligned}\mathbb{E}(S_n'^2) &= (1-\frac{1}{n})\mathbb{E}(X_1^2) - (1-\frac{1}{n})m^2 \\ &= (1-\frac{1}{n})(\sigma^2+m^2) - (1-\frac{1}{n})m^2 \\ &= \frac{n-1}{n}\sigma^2\end{aligned} \tag{1.1}$$

En conclusion de ces calculs on observera que dans le cas sans remise ou le cas avec remise $E(\overline{X}_n) = m, \quad V(\overline{X}_n) \sim \frac{\sigma^2}{n}$ en étant légèrement inférieure dans le cas sans remise. Egalement dans les deux cas $\mathbb{E}(S_n'^2) \sim \sigma^2$ sans avoir l'égalité et nous y reviendrons dans la suite. Nous noterons que le cas avec remise donne des calculs plus faciles. Ce sera notre modèle statistique fondamental.

1.3 Modèle statistique fondamental

DORÉNAVANT notre modèle d'échantillonnage sera la donnée de n variables aléatoire i.i.d. X_1,X_2,...,X_n. Supposons les par exemple réélles, on peut les construire sur l'espace canonique $(\mathbb{R}^n, B_{\mathbb{R}^n})$ muni de la mesure produit $\mathbb{P} = \mathbb{Q}^{\otimes n}$ où $\mathbb{Q}$ est la loi commune désirée pour les v.a. X_1,...,X_n qui sont alors simplement les coordonnées, $\omega \in \mathbb{R}^n$ s'écrivant $\omega = (X_1(\omega),...,X_n(\omega))$.

Le problème de la statistique mathématique est d'obtenir des renseignements sur la loi $\mathbb{Q}$ au vue de l'échantillon $(X_1,...,X_n)$. On pourra supposer que $\mathbb{Q}$ appartient à une famille $(\mathbb{Q}_\theta)$ de probabilités indexée par un paramètre θ réel ou vectoriel. Il s'agit alors de statistique paramètrique. Si $\mathbb{Q}$ appartient à une famille non paramètrisable de cette manière, il s'agit d'un problème de statistique non paramètrique.

Notons que dans le problème d'échantillonage d'une population finie les v.a. X_1,...,X_n prennent leurs valeurs dans un ensemble fini, l'ensemble des valeurs prises par le caractère étudié sur la population totale. Le cas d'une population totale très grande, assimilable à l'infini nous amène aussi à considérer pour loi $\mathbb{Q}$ commune aux X_i des lois continues, par exemple des lois à densité.

1.4 Grandeurs empiriques

A l'échantillon $X(\omega) = (X_1(\omega),...,X_n(\omega))$ on associe la mesure empirique

$$\mu(\omega) = \sum_{i=1}^{n} \frac{1}{n}\delta_{X_i(\omega)}$$

qui est simplement une mesure ponctuelle ayant un nombre fini d'atomes. Sont alors définis la moyenne empirique $\overline{X}_n$, la variance empirique S'^2_n, la fonction de répartition empirique, la médiane empirique, les moments empiriques centrés ou non comme étant ces quantités évaluées sur la mesure μ. Comment se comportent ces quantités quand n devient grand?

Le comportement des moments est une conséquence de la loi forte des grands nombres de Kolmogorov.

Proposition 1. *Soit $(X_1,...,X_n)$ un n-échantillon de la loi $\mathbb{Q}$ sur $\mathbb{R}$ supposée admettre un moment d'ordre $k \geq 1$ défini par $\int_{\mathbb{R}} x^k \mathbb{Q}(dx)$.*

Alors le moment empirique d'ordre k vérifie quand $n \longrightarrow +\infty$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{p.s.} \int_{\mathbb{R}} x^k \mathbb{Q}(dx)$$

et le moment centré empirique d'ordre k converge p.s. vers le moment centré de $\mathbb{Q}$ d'ordre k: quand $n \longrightarrow +\infty$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^k \xrightarrow{p.s.} \int_{\mathbb{R}} \left(x - \int y\ \mathbb{Q}(dy)\right)^k \mathbb{Q}(dx)$$

En particulier si $\mathbb{Q}$ admet une espérance, c'est la limite p.s. de $\overline{X_n}$ et si $\mathbb{Q}$ admet une variance c'est la limite p.s. de S'^2_n et de S^2_n.

Passons au comportement de la mesure empirique.

Théorème 2. *(Glivenko-Cantelli)*

Soit $(X_1,...,X_n)$ un n-échantillon de la loi $\mathbb{Q}$ sur $\mathbb{R}$. Alors la fonction de répartition empirique converge uniformément vers la fonction de répartition de $\mathbb{Q}$, presque sûrement. Autrement dit, quand $n \longrightarrow +\infty$:

$$\sup_{r\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \geq r\}} - \mathbb{Q}(]-\infty, r])\right| \xrightarrow{p.s.} 0$$

On remarquera que pour énoncer les deux résultats qui précèdent on doit disposer de n-échantillon pour n arbitrairement grand, c'est à dire d'une suite de v.a. indépendantes et de loi $\mathbb{Q}$. On peut construire toutes ces variables simultanément sur $\mathbb{R}^{\mathbb{N}}$ muni de la tribu produit des tribus boréliennes et de la probabilité produit (infini) $\mathbb{Q}^{\mathbb{N}}$. L'existence de ce dernier objet résulte du théorème de prolongement de Kolmogorov.

On appelle statistique d'ordre du n-échantillon $(X_1,...,X_n)$ le réordonnement $(X_{(1)},...,X_{(n)})$ de ces variables c'est à dire :

$$\{X_{(1)},...,X_{(n)}\} = \{X_1,...,X_n\} \quad \text{et } X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

En notant F_n la fonction de répartition empirique i.e.

$$F_n(r) = \frac{1}{n}\#\{i; X_i \leq r\},$$

on remarque que, si $X_{(k+1)} > X_{(k)}$, on a $F_n(X_{(k)}) = k/n$ et dans le cas général cette égalité est une inégalité. Soit $\alpha \in]0,1[$. Le plus petit entier k tel que $\frac{k}{n} > \alpha$ est $[n\alpha]+1$. On appelle fractile empirique d'ordre α la quantité $f_{\alpha,n} = X_{([n\alpha]+1)}$.

Notons maintenant F la fonction de répartition de $\mathbb{Q}$ i.e. $F(r) = \mathbb{Q}(]-\infty,r])$. On appelle fractile d'ordre α de $\mathbb{Q}$ le réel $\eta_\alpha = \inf\{r; F(r) > \alpha\}$.

Proposition 3. *Si η_α est un point de croissance de F alors $f_{\alpha,n} \xrightarrow{p.s.} \eta_\alpha$.*

On entend par "η_α point de croissance de F" que

$$\forall \delta > 0 \begin{cases} F(\eta_\alpha - \delta) < \alpha \\ F(\eta_\alpha + \delta) > \alpha \end{cases}$$

On notera que la seconde condition est automatique.

Preuve. On a $F_n(f_{\alpha,n}) \geq \frac{[n\alpha]+1}{n} \geq \alpha$, d'où en passant à la limite: $F(\underline{\lim}_n f_{\alpha,n}) \geq \alpha$ d'où $\underline{\lim}_n f_{\alpha,n} \geq \eta_\alpha$.

Soit maintenant $\delta > 0$. On a $F(\eta_\alpha + \delta) > \alpha$, donc il existe $\varepsilon > 0$ tel que $F(\eta_\alpha + \delta) \geq \alpha + \varepsilon$. Mais $\frac{1}{n}\sharp\{i; X_i \leq \eta_\alpha + \delta\} \longrightarrow F(\eta_\alpha + \delta) \geq \alpha + \varepsilon$. Cela entraine que pour n assez grand $X([n\alpha]+1) \leq \eta_\alpha + \delta$, d'où $\overline{\lim}_n f_{\alpha,n} \leq \eta_\alpha + \delta$ et ceci pour tout $\delta > 0$. Cela achève la preuve.

Par exemple la médiane empirique est définie par $X_{[\frac{n}{2}]+1}$.

1.5 Cas d'un échantillon gaussien

On considère un n-échantillon $(X_1,...,X_n)$ de la loi normale $\mathbb{Q} = \mathcal{N}(m,\sigma^2)$.

On définit la variance empirique modifiée par :

$$S_n^2 = \frac{n}{n-1} S_n'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2$$

.

Théorème 4. *(Fisher)*

Pour un n-échantillon de la loi gaussienne $\mathcal{N}(m,\sigma^2)$, la moyenne empirique et la variance empirique (modifiée) sont telles que :

(i) $\sqrt{n}\frac{\overline{X_n}-m}{\sigma} \overset{(loi)}{=} \mathcal{N}(0,1)$

(ii) $(n-1)\frac{S_n^2}{\sigma^2} \overset{(loi)}{=} \mathcal{X}^2(n-1)$ *loi du $\mathcal{X}^2$ à $n-1$ degrés de liberté.*

(iii) $\overline{X_n}$ *et* S_n^2 *sont indépendants.*

(iv) $\frac{\sqrt{n}(\overline{X_n}-m)}{S_n} \overset{(loi)}{=} \mathcal{S}_{n-1}$ *loi de Student à n-1 degrés de liberté.*

Preuve.

(i) Puisque le vecteur $X = \begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_n \end{pmatrix}$ est gaussien, la combinaison linéaire

$\overline{X}_n = \frac{X_1+\ldots+X_n}{n}$ suit une loi normale. De plus $\mathbb{E}(\overline{X}_n) = m$ donc la variable $\sqrt{n}\frac{\overline{X}_n - m}{\sigma}$ est centrée. De plus $V(\overline{X}_n) = \frac{\sigma^2}{n}$ donc $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m)$ est de variance 1.

(iii) Le vecteur $\begin{pmatrix} \overline{X}_n \\ X_1 - \overline{X}_n \\ . \\ . \\ . \\ X_n - \overline{X}_n \end{pmatrix}$ est gaussien comme image du vecteur gaussien $\begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_n \end{pmatrix}$ par une application linéaire.

Pour voir que $\overline{X_n}$ est indépendante de S_n^2, il suffit de voir que $\overline{X}_n$ est indépendante de $\begin{pmatrix} X_1 - \overline{X}_n \\ . \\ . \\ . \\ X_n - \overline{X}_n \end{pmatrix}$ et par le résultat sur les vecteurs gaussiens cela résultera de $\mathrm{Cov}(\overline{X}_n, X_i - \overline{X}_n) = 0$ pour $1 \leq i \leq n$. Pour un tel i on a en effet :

$$\begin{aligned} \mathrm{Cov}(\overline{X}_n, X_i - \overline{X}_n) &= \frac{1}{n^2}\mathrm{Cov}(\sum_j X_j, nX_i - \sum_k X_k) \\ &= \frac{1}{n^2}(n\sum_j \mathrm{Cov}(X_i, X_j) - \sum_{j,k}\mathrm{Cov}(X_j, X_k)) \\ &= \frac{1}{n^2}(n\underbrace{\mathrm{Cov}(X_i, X_i)}_{\sigma^2} - \sum_j \underbrace{\mathrm{Cov}(X_j, X_j)}_{\sigma^2}) \\ &\quad \text{car seuls ces termes sont non nuls} \\ &= 0 \text{ comme souhaité} \end{aligned} \tag{1.2}$$

(ii) On rappelle que la loi du $\mathcal{X}^2$ à $n-1$ degrés de liberté est la loi de

$$\sum_{i=1}^{n-1} Z_i^2 \quad \text{où} \quad \begin{pmatrix} Z_1 \\ . \\ . \\ . \\ Z_{n-1} \end{pmatrix} \overset{(loi)}{=} \mathcal{N}(O, I_{n-1})$$

c'est à dire que $Z_1, \ldots, Z_{n-1}$ sont i.i.d. de loi normale centrée réduite.

On note que

$$(n-1)\frac{S_n^2}{\sigma^2} = \sum_{i=1}^{n}(\frac{X_i - m}{\sigma} - \frac{1}{n}\sum_{j=1}^{n}\frac{X_j - m}{\sigma})^2.$$

Quitte à remplacer X_i par $\frac{X_i - m}{\sigma}$, on peut se contenter de prouver le résultat quand $m = 0$, $\sigma = 1$ i.e. $X = \begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_n \end{pmatrix}$ suit la loi $\mathcal{N}(0,I_n)$.

Notons que $\begin{pmatrix} X_1 - \overline{X_n} \\ . \\ . \\ . \\ X_n - \overline{X_n} \end{pmatrix} = X - \frac{1}{\sqrt{n}}AX$ où $A = \frac{1}{\sqrt{n}}\begin{pmatrix} 1 & . & . & . & 1 \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ 1 & . & . & . & 1 \end{pmatrix}$

et que

$$(n-1)S_n^2 = \|X - \frac{1}{\sqrt{n}}AX\|^2 = \|U(X - \frac{1}{\sqrt{n}}AX)\|^2$$

pour toute matrice orthogonale U car une telle matrice préserve la norme euclidienne. Le vecteur $U(I - \frac{1}{\sqrt{n}}A)X$ suit la loi normale centrée de matrice de variance covariance :

$$\begin{aligned} \Lambda &= U\left(I - \frac{1}{\sqrt{n}}A\right) I \left(U(I - \frac{1}{\sqrt{n}}A)\right)' \\ &= UU' - \frac{1}{\sqrt{n}}UA'U - \frac{1}{\sqrt{n}}UAU' + \frac{1}{n}UAA'U' \end{aligned}$$

Or on peut trouver une matrice orthogonale U telle que

$U\begin{pmatrix} 1/\sqrt{n} \\ . \\ . \\ . \\ 1/\sqrt{n} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ . \\ . \\ 0 \end{pmatrix}$ car les deux vecteurs $\begin{pmatrix} 1/\sqrt{n} \\ . \\ . \\ . \\ 1/\sqrt{n} \end{pmatrix}$ et $\begin{pmatrix} 1 \\ 0 \\ . \\ . \\ 0 \end{pmatrix}$

sont de norme 1.

On a alors $UA = \begin{pmatrix} 1 & . & . & . & 1 \\ 0 & . & . & . & 0 \\ . & & & & . \\ . & & & & . \\ 0 & . & . & . & 0 \end{pmatrix}$

et $UA'U = U\begin{pmatrix} 1 & 0 & . & . & 0 \\ . & & & & \\ . & & & & . \\ . & & & & . \\ 1 & 0 & . & . & 0 \end{pmatrix} = \begin{pmatrix} \sqrt{n} & 0 & . & . & 0 \\ 0 & & & & . \\ . & & & & . \\ . & & & & . \\ 0 & . & . & . & 0 \end{pmatrix}$

et également $UAA'U' = UA(UA)' = \begin{pmatrix} n & 0 & . & . & 0 \\ 0 & & & & . \\ . & & & & . \\ . & & & & . \\ 0 & . & . & . & 0 \end{pmatrix}$ de sorte qu'au total

$$\Lambda = I - \begin{pmatrix} 1 & 0 & . & . & 0 \\ 0 & & & & . \\ . & & & & . \\ . & & & & . \\ 0 & . & . & . & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & . & . & 0 \\ 0 & 1 & & & \\ & & . & & \\ & & & . & \\ 0 & & & & 1 \end{pmatrix}$$

Ainsi $(n-1)S_n^2 = \|Z\|^2$ où $Z \overset{(loi)}{=} \mathcal{N}(0, \begin{pmatrix} 0 & 0 & . & . & 0 \\ 0 & 1 & & & \\ & & . & & \\ & & & . & \\ 0 & & & & 1 \end{pmatrix})$ et cela montre bien que $(n-1)S_n^2$ suit la loi $\mathcal{X}^2(n-1)$.

(iv)

$$\frac{\sqrt{n}(\overline{X_n} - m)}{S_n} = \frac{\sqrt{n}\frac{(X_n - m)}{\sigma}}{\sqrt{\frac{(n-1)S_n^2/\sigma^2}{n-1}}} \overset{(loi)}{=} \frac{N}{\sqrt{\frac{\mathcal{X}_{n-1}^2}{n-1}}}$$

où N et $\mathcal{X}_{n-1}^2$ sont deux variables indépendantes de lois respectives $\mathcal{N}(0,1)$ et $\mathcal{X}^2(n-1)$. C'est la définition même d'une loi de student à $n-1$ degrés de liberté.

Proposition 5. *Réciproquement au résultat précédent, si $(X_1,...,X_n)$ est un n-échantillon d'une loi $\mathbb{Q}$ et si $\overline{X}_n$, S_n^2désignent comme d'habitude la moyenne et la variance empirique, l'indépendance de $\overline{X}_n$ et S_n^2 implique que $\mathbb{Q}$ est une loi normale.*

La preuve est en exercice ci-dessous.

Exercice 1 (Tirage avec remise dans une population finie). On effectue un tirage avec remise d'un échantillon de taille n dans une population de taille N. On note $Y_1,Y_2,\ldots,Y_N$ les nombres respectifs de fois où les individus $1,\ldots,N$ ont été tirés.
a) Quelle est la loi du vecteur $(Y_1,Y_2,\ldots,Y_N)$?
b) On étudie un caractère dont la valeur sur le i-ième individu est x_i et on note comme d'habitude $\overline{X}_n$ la moyenne empirique de la valeur du caratère sur l'échantillon tiré. En écrivant

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{N} x_i\, Y_i$$

retrouver les valeurs de $\mathbb{E}(\overline{X}_n)$ et $\mathbb{V}(\overline{X}_n)$ obtenues en cours.

Exercice 2 (Variance de la variance empirique). Calculer la variance de la variance empirique S'^2_n relative à un échantillon de taille n tiré avec remise dans une population.

Exercice 3 (Convergence du moment empirique centré). Prouver la proposition du cours qui affirme le résultat suivant : si une loi $\mathbb{Q}$ admet un moment centré d'ordre k alors c'est la limite presque sûre quand $n \to +\infty$ du moment empirique centré d'ordre k d'un n–échantillon de la loi $\mathbb{Q}$.

Exercice 4 (Utilisation du théorème de Fisher). Calculer $\mathbb{V}(S_n^2)$ pour un échantillon gaussien, en utilisant le théorème de Fisher. Cela confirme-t-il le résultat général valable pour une loi quelconque?

Exercice 5 (Caractérisation d'un échantillon gaussien). Soit $X_1,\ldots,X_n$ un n–échantillon d'une loi $\mathbb{Q}$ centrée et de variance finie σ^2, dont la fonction caractéristique sera notée φ. On suppose que les deux variables

$$S_n = \sum_{i=1}^{n} X_i, \quad \Lambda_n = \sum_{i=1}^{n} \left(X_i - \frac{S_n}{n}\right)^2$$

sont indépendantes, c'est à dire que la variance empirique et la moyenne empirique sont indépendantes. Calculer de deux manières, pour t réel, la quantité $\mathbb{E}\left(\Lambda_n\, e^{it\,S_n}\right)$ et en déduire que φ vérifie l'équation différentielle

$$\frac{\varphi''}{\varphi} - \left(\frac{\varphi'}{\varphi}\right)^2 = -\sigma^2$$

Résoudre et en conclure que $\mathbb{Q}$ est une loi normale. La condition de centrage de $\mathbb{Q}$ peut elle être levée?

Chapitre 2

Modèle statistique et réduction des données

Un modèle statistique est une famille $((\Omega,\mathcal{F},\mathbb{P}_\theta), \theta \in \Theta)$ d'espaces probabilisés où les probabilités $\mathbb{P}_\theta$ sont indexées par un indice θ qui varie dans l'ensemble Θ. Quand Θ est une partie de $\mathbb{R}$ ou de $\mathbb{R}^d$ on dit que ce modèle est *paramétrique*. L'essentiel de ce cours concernera le modèle statistique de l'échantillon de taille n, comme défini au chapitre précédent. Dans la pratique on peut construire ce modèle sur $\Omega = \mathbb{R}^n$ avec $\mathcal{F} = \mathcal{B}_{\mathbb{R}^n}$, $\mathbb{P}_\theta = (\mathbb{Q}_\theta)^{\otimes n}$ et l'échantillon est $X(w) = (X_1(w),...,X_n(w))$ où X est simplement l'identité de $\mathbb{R}^n$ dans lui-même. Cette description est appelée espace canonique.

On appelle *statistique* du modèle $((\Omega,\mathcal{F},\mathbb{P}_\theta), \theta \in \Theta)$ toute variable aléatoire ou vecteur aléatoire T sur $(\Omega,\mathcal{F})$. Dans le modèle canonique on a trivialement $T = T \circ X = T(X)$.

Un modèle statistique $((\Omega,\mathcal{F},\mathbb{P}_\theta), \theta \in \Theta)$ est dit *dominé* si pour tout $\theta \in \Theta$, la probabilité $\mathbb{P}_\theta$ est dominée (au sens absolument continue) par une mesure σ-finie μ sur $(\Omega,\mathcal{F})$, ce qui signifie que $\forall B \in \mathcal{F}, \mu(B) = 0 \Rightarrow \mathbb{P}_\theta(B) = 0$ et cela équivaut à l'existence d'une densité de $\mathbb{P}_\theta$ par rapport à μ.

On démontre que si le modèle est dominé, on peut trouver une mesure dominante μ qui a en plus les propriétés suivantes :

- μ est une probabilité ;
- μ est minimale au sens de la domination : toute mesure qui domine tous les $\mathbb{P}_\theta$ domine μ.
- pour tout $A \in \mathcal{F}$, $\mu(A) = 0$ ssi $(\mathbb{P}_\theta(A) = 0$ pour tout $\theta \in \Theta)$

et on peut même choisir cette "dominante privilégiée" de la forme $\sum_n a_n \mathbb{P}_{\theta_n}$ avec des $a_n \geq 0$ de somme 1.

Pour un modèle $(\Omega,\mathcal{F},\mathbb{P}_\theta)$ dominé par $\tilde{\mathbb{P}}$, la densité $f(X,\theta) = \frac{d\mathbb{P}_\theta}{d\tilde{\mathbb{P}}}$ est appelée vraisemblance du modèle.

2.1 Statistique exhaustive

On dit qu'une statistique T du modèle statistique $((\Omega,\mathcal{F},\mathbb{P}_\theta), \theta \in \Theta)$ est *exhaustive* si la loi conditionnelle de X sachant T sous $\mathbb{P}_\theta$ n'est pas fonction de θ. Cela signifie que "toute l'information sur θ donnée par X est dans la valeur $t = T(X)$ et la position de X sur la surface $\{x; T(x) = t\}$ n'apporte aucune information supplémentaire".

Théorème 6 (de factorisation de Fisher-Neyman). *Soit $((\Omega,\mathcal{F},\mathbb{P}_\theta), \theta \in \Theta)$ un modèle statistique dominé par $\tilde{\mathbb{P}}$ avec des densités $f(X,\theta) = \frac{d\mathbb{P}_\theta}{d\tilde{\mathbb{P}}}$. La statistique T est exhaustive si et seulement si il existe $g : \Omega \longrightarrow \mathbb{R}_+$ et $h : \Omega \times \Theta \longrightarrow \mathbb{R}_+$ mesurables telles que :*

$$f(X,\theta) = g(X)\, h(T(X),\theta).$$

Preuve. Supposons d'abord T exhaustive. Soit ψ fonction test bornée, il existe $q(\cdot)$ mesurable bornée telle que pour tout $\theta \in \Theta$, on a $\mathbb{E}_\theta(\psi(X)|T) = q(T)$. Considérons $\hat{\mathbb{P}}$ la dominante privilégiée des $(\mathbb{P}_\theta, \theta \in \Theta)$. On peut donc écrire $\hat{\mathbb{P}} = \sum_n a_n\, \mathbb{P}_{\theta_n}$. Pour φ fonction test bornée,

$$\begin{aligned}
\hat{\mathbb{E}}(\varphi(T)q(T)) &= \sum_n a_n \mathbb{E}_{\theta_n}(\varphi(T)q(T)) \\
&= \sum_n a_n \mathbb{E}_{\theta_n}(\varphi(T)\mathbb{E}_{\theta_n}(\psi(X)|T)) \\
&= \sum_n a_n \mathbb{E}_{\theta_n}(\varphi(T)\psi(X)) \\
&= \hat{\mathbb{E}}(\varphi(T)\psi(X))
\end{aligned}$$

Cela prouve que $\hat{\mathbb{E}}(\psi(X)|T) = q(T) = \mathbb{E}_\theta(\psi(X)|T)$ pour tout θ, d'où :

$$\begin{aligned}
\mathbb{E}_\theta(\psi(X)) &= \mathbb{E}_\theta(\mathbb{E}_\theta(\psi(X)|T)) = \mathbb{E}_\theta(\hat{\mathbb{E}}(\psi(X)|T)) \\
&= \hat{\mathbb{E}}(\hat{f}(X,\theta)\hat{\mathbb{E}}(\psi(X)|T)) \text{ où } \hat{f}(X,\theta) = \frac{d\mathbb{P}_\theta}{d\hat{\mathbb{P}}} \\
&= \hat{\mathbb{E}}(\hat{\mathbb{E}}(\hat{f}(X,\theta)|T)\hat{\mathbb{E}}(\psi(X)|T)) \\
&= \hat{\mathbb{E}}(\hat{\mathbb{E}}(\hat{f}(X,\theta)|T)\psi(X))
\end{aligned}$$

Cela montre que $\frac{d\mathbb{P}_\theta}{d\hat{\mathbb{P}}} = \hat{\mathbb{E}}(\hat{f}(X,\theta)|T)$ que l'on peut noter $h(T,\theta)$. On sait ensuite que $\hat{\mathbb{P}}$ est dominée par $\tilde{\mathbb{P}}$ et on note $g(X)$ la densité $\frac{d\hat{\mathbb{P}}}{d\tilde{\mathbb{P}}}$. On obtient au total que :

$$\frac{d\mathbb{P}_\theta}{d\tilde{\mathbb{P}}} = h(T,\theta)g(X)$$

Réciproquement supposons que

$$f(X,\theta) = \frac{d\mathbb{P}_\theta}{d\tilde{\mathbb{P}}} = h(T,\theta)g(X).$$

Soit φ et ψ deux fonctions test.

$$\begin{aligned}
\mathbb{E}_\theta(\varphi(T)\psi(X)) &= \tilde{\mathbb{E}}(f(X,\theta)\varphi(T)\psi(X)) = \tilde{\mathbb{E}}(h(T,\theta)\varphi(T)g(X)\psi(X)) \\
&= \tilde{\mathbb{E}}(h(T,\theta)\varphi(T)\tilde{\mathbb{E}}(g(X)\psi(X)|T)) \\
&= \mathbb{E}_\theta\left(\varphi(T)\frac{\tilde{\mathbb{E}}(g(X)\psi(X)|T)}{g(X)}\right) \\
&= \mathbb{E}_\theta\left(\varphi(T)\tilde{\mathbb{E}}(g(X)\psi(X)|T)\mathbb{E}_\theta\left(\frac{1}{g(X)}|T\right)\right)
\end{aligned}$$

d'où

$$\mathbb{E}_\theta(\psi(X)|T) = \tilde{\mathbb{E}}(g(X)\psi(X)|T)\mathbb{E}_\theta\left(\frac{1}{g(X)}|T\right)$$

En prenant $\psi \equiv 1$ on obtient $\mathbb{E}_\theta\left(\frac{1}{g(X)}|T\right) = \frac{1}{\tilde{\mathbb{E}}(g(X)|T)}$. D'où finalement :

$$\mathbb{E}_\theta(\psi(X)|T) = \frac{\tilde{\mathbb{E}}(g(X)\psi(X)|T)}{\tilde{\mathbb{E}}(g(X)|T)}.$$

ce qui montre que la loi conditionnelle de X sachant T n'est pas fonction de θ. Pour le problème de la division par 0 on notera que

$$\mathbb{P}_\theta(g(X)=0) = \tilde{\mathbb{E}}(1_{\{g(X)=0\}}\, g(X)\, h(T,\theta)) = 0.$$

Exemple 1 Soit le modèle statistique $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \mathbb{P}_{a,b}^{\otimes n}; a < b \text{ réels})$ où $\mathbb{P}_{a,b}$ est la loi uniforme sur $[a,b]$. La densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ s'écrit:

$$\begin{aligned}
f(X,\theta) &= \prod_{i=1}^n \frac{1}{b-a}\mathbf{1}_{[a,b]}(X_i) \\
&= \frac{1}{(b-a)^n}\mathbf{1}_{\{\forall i,\, a\le X_i\le b\}} \\
&= \frac{1}{(b-a)^n}\mathbf{1}_{\{a\le \min_{1\le i\le n} X_i \le \max_{1\le i\le n} X_i\le b\}} \\
&= \frac{1}{(b-a)^n}\mathbf{1}_{[a,b]^2}(T(X))
\end{aligned}$$

où $T(X) = (\min_i X_i, \max_i X_i)$. Cette statistique T est donc exhaustive.

Exemple 2 Considérons le modèle statistique $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \mathbb{P}_m^{\otimes n}; m \in \mathbb{R})$ où $\mathbb{P}_m$ est la loi gaussienne $\mathcal{N}(m,\sigma^2)$ où m est inconnu mais σ^2 est fixé. La vraisemblance est :

$$\begin{aligned}
f(X,m) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-m)^2}{2\sigma^2}} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2}\; e^{\frac{m}{\sigma^2}\sum_{i=1}^n X_i - \frac{nm^2}{2\sigma^2}} \\
&= g(X)h(T(X),m)
\end{aligned}$$

$$\text{où} \quad \begin{cases} g(X) = (2\pi\sigma)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2} \\ h(T(X),m) = e^{\frac{m}{2\sigma^2}(2\,T(X)-nm)} \\ T(X) = \sum_{i=1}^n X_i \end{cases}$$

On en déduit que la statistique $T(X) = \sum_i X_i$ est exhaustive donc aussi la moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

2.2 Statistique exhaustive minimale

Définition. Une statistique exhaustive S est minimale si pour toute autre statistique exhaustive T on a $S = \psi(T)$ où $\psi(.)$ est une certaine application mesurable.

Théorème 7. *(Lehmann-Scheffé)*

Soit $(\Omega,\mathcal{F},\mathbb{P}_\theta;\ \theta \in \Theta)$ un modèle statistique dominé (par $\tilde{\mathbb{P}}$) de vraisemblance $f(\cdot,\theta)$. Pour que la statistique S soit exhaustive minimale il suffit qu'on ait l'équivalence :

$$S(x) = S(y) \Longleftrightarrow \frac{f(x,\theta)}{f(y,\theta)} \text{ est constante par rapport à } \theta$$

Preuve. Pour tout $y \in S(\Omega)$, on choisit de façon mesurable $S^-(y)$ tel que $S(S^-(y)) = y$. On fixe $\theta_0 \in \Theta$. Pour tout $x \in \Omega$ on a $S(S^-(S(x))) = S(x)$, d'où :

$$\frac{f(x,\theta)}{f(S^-(S(x)),\theta)} = \frac{f(x,\theta_0)}{f(S^-(S(x)),\theta_0)}$$

Ainsi $f(x,\theta) = g(x)h(S(x),\theta)$ où $\begin{cases} h(y,\theta) = f(S^-(y),\theta) \\ g(x) = \frac{f(x,\theta_0)}{f(S^-(S(x)),\theta_0)} \end{cases}$

Cela prouve déjà que S est exhaustive. Considèrons T une autre statistique exhaustive. On a donc

$$f(x,\theta) = \tilde{g}(x)\ \tilde{h}(T(x),\theta)$$

Si $T(x) = T(y)$ on a :

$$\frac{f(x,\theta)}{f(y,\theta)} = \frac{\tilde{g}(x)}{\tilde{g}(y)}\frac{\tilde{h}(T(x),\theta)}{\tilde{h}(T(y),\theta)} = \frac{\tilde{g}(x)}{\tilde{g}(y)} \text{ indépendant de } \theta.$$

Donc $S(x) = S(y)$. On a donc $T(x) = T(y) \Longrightarrow S(x) = S(y)$ donc S est une fonction de T.

Exemple. Dans le cas de l'exemple 2 ci-dessus on a :

$$\frac{f(x,m)}{f(y,m)} = \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2}} e^{\frac{m}{\sigma^2}(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i)}$$

est indépendante de m ssi $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Cela prouve que la statistique $S(X) = \sum_{i=1}^n X_i$ ou encore $\overline{X}_n$ est exhaustive minimale.

2.3 Complétude

Définition. On dit que le modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta;\ \theta\in\Theta)$ est complet si pour toute $\psi:\Omega\longrightarrow\mathbb{R}$ mesurable et $\mathbb{P}_\theta$-intégrable pour tout $\theta\in\Theta$, on a :

$$\left(\forall\theta\in\Theta,\int_\Omega\psi(x)d\mathbb{P}_\theta(x)=0\right)\Longrightarrow\psi=0,\quad\mathbb{P}_\theta\text{–pp pour tout }\theta\in\Theta.$$

On dit qu'une statistique S sur Ω est complète si pour toute ψ bornée

$(\forall\theta\in\Theta,\mathbb{E}_\theta(\psi(S))=0)\Longrightarrow\psi=0$ pp par rapport à la loi de S sous $\mathbb{P}_\theta$, pour tout $\theta\in\Theta$.

Théorème 8. *Soit S une statistique exhaustive et complète, à valeurs dans $\mathbb{R}^d$, sur le modèle $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta\in\Theta)$. Alors S est exhaustive minimale.*

Preuve. Soit T une autre statistique exhaustive. Il s'agit de prouver $S=\gamma(T)$. Quitte à prouver le résultat pour chaque composante de S, on peut supposer que $d=1$. Quitte à composer S par une fonction bijective bornée on peut supposer que S est bornée. Posons alors

$$\gamma(t)=\mathbb{E}_\theta(S|T=t),\quad\delta(s)=\mathbb{E}_\theta(\gamma(T)|S=s)$$

où la valeur du paramètre θ dans ces expressions n'a pas d'importance compte tenu de l'exhaustivité de T et S. Remarquons aussi la bonne définition puisque S est bornée. Alors:

$$\begin{aligned}\mathbb{E}_\theta(S) &= \mathbb{E}_\theta(\mathbb{E}_\theta(S|T))=\mathbb{E}_\theta(\gamma(T))\\ &= \mathbb{E}_\theta(\mathbb{E}_\theta(\gamma(T)|S))=\mathbb{E}_\theta(\delta(S))\end{aligned}$$

d'où $\mathbb{E}_\theta(S-\delta(S))=0$ et cela est vrai pour tout θ donc $\delta=id$ p.p. par rapport aux lois de S sous les $\mathbb{P}_\theta$. Ainsi $\mathbb{E}_\theta(\gamma(T)|S)=S$ Alors la variance de S s'écrit

$\mathbb{V}_\theta(S)=\mathbb{E}_\theta(\mathbb{V}_\theta(S|T))+\underbrace{\mathbb{V}_\theta(\mathbb{E}_\theta(S|T))}_{\mathbb{V}_\theta(\gamma(T))}$ par la formule d'analyse de la variance

Par cette même formule on a aussi :

$$\mathbb{V}_\theta(\gamma(T))=\mathbb{E}_\theta(\mathbb{V}_\theta(\gamma(T)|S))+\mathbb{V}_\theta(\underbrace{\mathbb{E}_\theta(\gamma(T)|S)}_{\delta(S)=S})$$

En sommant on trouve:

$$\mathbb{E}_\theta(\mathbb{V}_\theta(S|T))+\mathbb{E}_\theta(\mathbb{V}_\theta(\gamma(T)|S))=0$$

D'où $\mathbb{V}_\theta(S|T)=0=\mathbb{V}_\theta(\gamma(T)|S)$. La première égalité entraine que $S=\mathbb{E}_\theta(S|T)=\gamma(T)$ ce qui achève la preuve.

2.4 Liberté

Définition. Une statistique sur un modèle statistique de paramètre θ est dite *libre* si sa loi ne dépend pas de θ.

Exemple: On a vu que sur le modèle $(\mathbb{R}^n,\mathcal{B}_{\mathbb{R}^n},\mathcal{N}(m,\sigma^2)^{\otimes n})$ paramétré par (m,σ^2), les statistiques

$$\sqrt{n}\frac{\overline{X_n}-m}{\sigma} \quad \text{et} \quad \frac{n-1}{\sigma^2}S_n^2$$

sont libres.

Théorème 9 ("de Basu"). *Soit S une statistique exhaustive et complète sur le modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta\in\Theta)$. Alors S est $\mathbb{P}_\theta$-indépendante de toute statistique libre sur ce modèle.*

Preuve. Voir exercice ci-dessous.

2.5 Exercices

Exercice 6 (Exhaustivité dans le modèle gaussien et poissonien).
a) On considère le modèle $(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n),\mathbb{Q}_\theta^{\otimes n})$ où $\mathbb{Q}_\theta$ est la loi normale (unidimensionnelle) $\mathcal{N}(m,\sigma^2)$ et $\theta=(m,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*$. En utilisant le théorème de Fisher-Neyman, trouver une statistique exhaustive pour ce modèle.
b) Même question si $\mathbb{Q}_\theta$ est la loi de Poisson de paramètre $\theta\in\mathbb{R}_+^*$. La statistique trouvée est elle complète? Le modèle est il complet?

Exercice 7 (Exhaustivité et minimalité d'une statistique). On considère le modèle statistique d'un n–échantillon $X=(X_1,\ldots,X_n)$ de vraisemblance

$$L(X,\theta)=\prod_{i=1}^{n}\frac{1}{\ln(\alpha)}\,\frac{1}{X_i}\,\mathbf{1}_{]\theta,\alpha\,\theta[}(X_i)$$

où $\alpha>1$ est une constante fixée et le paramètre inconnu θ varie dans $\mathbb{R}_+^*$. Montrer que la statistique

$$\left(\min_{1\le i\le n}X_i,\ \max_{1\le i\le n}X_i\right)$$

est exhaustive minimale. Le modèle est il complet?

Exercice 8 (Formule d'analyse de la variance). Pour S et T deux statistiques sur $(\Omega,\mathcal{F},\mathbb{P})$, montrer que

$$\mathbb{V}(S)=\mathbb{E}\left(\mathbb{V}(S|T)\right)+\mathbb{V}\left(\mathbb{E}(S|T)\right)$$

où $\mathbb{V}(S|T)$ est la variance conditionnelle de S sachant T c'est à dire

$$\mathbb{V}(S|T)=\mathbb{E}\left((S-\mathbb{E}(S|T))^2\Big|T\right)$$

Exercice 9 (Modèle à paramètre de position). On considère le modèle statistique $\left(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n),\mathbb{Q}_\theta^{\otimes n}\right)$ à paramètre de position $\theta \in \mathbb{R}$ c'est à dire que la loi $\mathbb{Q}_\theta$ est la translatée de $\mathbb{Q}_0$ par θ, autrement dit

$$\int \mathbb{Q}_\theta(dx)\ \psi(x) = \int \mathbb{Q}_0(dx)\ \psi(x+\theta)$$

Montrer que l'étendue de l'échantillon $X = (X_1,\ldots,X_n)$ définie par

$$R(X) = \max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i$$

est libre. Préciser sa loi si $\mathbb{Q}_\theta$ est la loi uniforme sur $]\theta,\theta+1[$.

Exercice 10 (Théorème de Basu). Soit S une statistique exhaustive complète sur $(\Omega,\mathcal{F},\mathbb{P}_\theta;\ \theta\in\Theta)$ et T une statistique libre.
a) Pour φ mesurable bornée et

$$\psi(s) = \mathbb{E}_\theta\left(\varphi(T)\right) - \mathbb{E}_\theta\left(\varphi(T)|S=s\right)$$

que vaut $\mathbb{E}_\theta\left(\psi(S)\right)$?
b) En déduire que S et T sont indépendantes.
c) Comment appliquer ce résultat pour obtenir une nouvelle justification de l'indépendance dans le théorème de Fisher sur les échantillons gaussiens?

Chapitre 3

Modèle statistique : classification information

Dans ce chapitre nous définissons une classe importante de modèles paramètriques : les modèles exponentiels. Nous introduisons également le concept d'information qui interviendra à nouveau dans le prochain chapitre consacré à l'estimation.

3.1 la famille exponentielle

Définition : Soit $(\Omega,\mathcal{F},\mathbb{P}_\theta\,;\,\theta\in\Theta)$ un modèle statistique dominé par une mesure σ-finie et $L(x,\theta)$ la vraisemblance. On dit que ce modèle est un modèle exponentiel ou que la famille $(\mathbb{P}_\theta)$ est une famille exponentielle si la vraisemblance s'écrit :

$$L(x,\theta)=\beta(\theta)\xi(x)\exp(\alpha(\theta).T(x))$$

où

$T:\Omega\longrightarrow\mathbb{R}^d$ est une statistique dite statistique naturelle.

$\alpha:\Theta\longrightarrow\mathbb{R}^d$ est une application mesurable dite paramètre naturel.

$\beta:\Theta\longrightarrow\mathbb{R}_+$, $\xi:\Omega\longrightarrow\mathbb{R}_+$ sont mesurables. L'entier d est appelé dimension du modèle.

Exemple 1: Soit Λ une matrice symétrique $d\times d$ définie positive et $\mathbb{P}_\theta$ la loi normale $\mathcal{N}(\theta,\Lambda)$ sur $\mathbb{R}^d$ où $\theta\in\mathbb{R}^d$. La vraisemblance (par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$) s'écrit :

$$\begin{aligned}L(x,\theta) &= \frac{1}{(2\pi)^{d/2}(det\Lambda)^{1/2}}\exp\left(-\frac{1}{2}(x-\theta)'\Lambda^{-1}(x-\theta)\right)\\ &= \underbrace{e^{-\frac{1}{2}\theta'\Lambda^{-1}\theta}}_{\beta(\theta)}\underbrace{\frac{1}{\sqrt{(2\pi)^d det\Lambda}}e^{-\frac{1}{2}x'\Lambda^{-1}x}}_{\xi(x)}\underbrace{e^{\theta.\Lambda^{-1}x}}_{e^{\alpha(\theta).T(x)}}\end{aligned}$$

On a la forme d'un modèle exponentiel avec $\alpha(\theta) = \theta$, $T(x) = \Lambda^{-1}x$ ou encore $\alpha(\theta) = \Lambda^{-1}\theta$, $T(x) = x$. Noter que ce modèle n'est pas celui d'un n-échantillon d'une loi donnée.

Proposition 10. *Si* $(\mathbb{R},\mathcal{B}_{\mathbb{R}},\mathbb{Q}_\theta;\theta \in \Theta)$ *est un modèle exponentiel alors le modèle d'un n-échantillon* $(\mathbb{R}^n,\mathcal{B}_{\mathbb{R}^n},\mathbb{Q}_\theta^{\otimes n};\theta \in \Theta)$ *est un modèle exponentiel.*

Preuve. Soit μ la mesure dominant les $(\mathbb{Q}_\theta)$, l'hypothèse s'écrit

$$\frac{d\mathbb{Q}_\theta}{d\mu} = \beta(\theta)\xi(x)\exp(\alpha(\theta).T(x))$$

Alors

$$\frac{d\mathbb{Q}_\theta^{\otimes n}}{d\mu^{\otimes n}} = \beta(\theta)^n \prod_{i=1}^{n} \xi(x_i) \quad \exp\underbrace{(\sum_{i=1}^{n} \alpha(\theta).T(x_i))}_{\tilde{\alpha}(\theta).\tilde{T}(x)}$$

où $\tilde{\alpha}(\theta) = (\alpha(\theta),...,\alpha(\theta))$, $\tilde{T}(x) = (T(x_1),...,T(x_2))$. D'où le résultat.

Reste donc à examiner le cas de quelques lois classiques pour voir si elles appartiennent à la famille exponentielle. C'est le cas pour

1. la loi normale $\mathcal{N}(m,\sigma^2)$

$$\begin{aligned} L(x,(m,\sigma^2)) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \\ &\quad \text{avec pour mesure dominante la mesure de Lebesgue sur } \mathbb{R}_+ \\ &= \frac{e^{-\frac{m^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} e^{(-\frac{1}{2\sigma^2},\frac{m}{\sigma^2}).(x^2,x)} \end{aligned}$$

2. la loi binomiale $\mathcal{B}(n,p)$ de paramètre p inconnu. Par rapport à la mesure de comptage sur $\{0,...,n\}$, la vraisemblance est :

$$\begin{aligned} L(x,p) &= C_n^x p^x (1-p)^{n-x} \\ &= C_n^x (1-p)^n e^{x\ln(\frac{p}{1-p})} \end{aligned}$$

3. la loi de Poisson $P(\lambda)$. La mesure dominante sera la mesure de comptage sur $\mathbb{N}$.

$$\begin{aligned} L(x,\lambda) &= e^{-\lambda}\frac{\lambda^x}{x!} \\ &= e^{-\lambda}\frac{1}{x!}e^{x\ln(\lambda)} \end{aligned}$$

4. la loi exponentielle $\xi(\theta)$

$$L(x,\theta) = \theta\ e^{-\theta x}$$

5. et plus généralement la loi gamma $\gamma(p,\theta)$. Par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$, la vraisemblance est :

$$L(x,(p,\theta)) = \frac{\theta^p}{\Gamma(p)} e^{-\theta x} x^{p-1} = \frac{\theta^p}{\Gamma(p)} e^{(-\theta, p-1).(x, \ln x)}$$

Par contre la loi uniforme sur $[0,\theta]$ i.e. $L(x,\theta) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x)$ ou la loi de Cauchy centrée en θ i.e. $L(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$ ne se mettent pas sous forme exponentielle.

On dira qu'on a affaire au modèle exponentiel standard sur $\mathbb{R}^n$ si la vraisemblance s'écrit, par rapport à la mesure dominante ν,

$$L(x,\theta) = c(\theta)^{-1} e^{x.\theta}$$

où $c(\theta) = \int_{\mathbb{R}^n} e^{y.\theta} d\nu(y)$ étant entendu que cette intégrale est finie, ce qui sera vrai quand θ appartient à un certain ensemble appelé espace naturel des paramètres (dont on montre qu'il s'agit d'une partie convexe de $\mathbb{R}^n$).

Avec un bon choix de θ on est dans cette situation pour les exemples 2,3,4 ci-dessus.

Proposition 11. *Pour le modèle statistique standard sur $\mathbb{R}^n$, donc de vraisemblance s'écrivant par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$:*

$$L(x,\theta) = c(\theta)^{-1} e^{x.\theta}$$

les moments sont donnés par :

$$\mathbb{E}_\theta\left(\prod_{i=1}^n X_i^{k_i}\right) = c(\theta)^{-1} \frac{\partial^k c(\theta)}{\partial\theta_1^{k_1} ... \partial\theta_n^{k_n}}, \quad k = k_1 + ... + k_n$$

à l'intérieur de l'espace des paramètres

Preuve. C'est une application du théorème de dérivation sous le signe intégrale.

Proposition 12. *Pour un modèle exponentiel la statistique naturelle est exhaustive. Si l'espace image du paramètre naturel est d'intérieur non vide, cette statistique naturelle est complète donc exhaustive minimale.*

Preuve. L'exhaustivité est claire par le théorème de factorisation de Fisher-Neyman. Pour le caractère complet supposons que pour g bornée et pour tout $\theta \in \Theta$,

$$\int_{\mathbb{R}^d} g(T(x))\ \beta(\theta)\xi(x)\exp(\alpha(\theta).T(x))\ \ \mu(dx) = 0$$

On a repris les notations de la définition ci-dessus avec une mesure σ-finie dominante μ. D'où, pour $z \in \alpha(\Theta)(\subset \mathbb{R}^d)$

$$\int_{\mathbb{R}^d} g_+(T(x))\ e^{z.T(x)}\xi(x)\ \mu(dx) = \int_{\mathbb{R}^d} g_-(T(x))\ e^{z.T(x)}\xi(x)\ \mu(dx).$$

Cela signifie que les mesures de densité respective g_+ et g_- par rapport à la mesure image par T de $\xi(\cdot)\mu$ ont même transformée de Laplace pour $z \in \alpha(\Theta)$. Si $\alpha(\Theta)$ est d'intérieur non vide cela implique que ces deux mesures sont égales donc

$$\int_{\mathbb{R}^d} \xi(x)\mu(dx)\mathbf{1}_{\{g_+(T(x))\neq g_-(T(x))\}} = 0$$

Il s'ensuit que $\mu\{x\ ;\ g_+(T(x)) \neq g_-(T(x))$ et $\xi(x) > 0\} = 0$ et donc $\mathbb{P}_\theta(g(T(x)) \neq 0$ et $\xi(x) > 0) = 0$ puisque $g_+(x) \neq g_-(x) \Longleftrightarrow g(x) \neq 0$. De plus

$$\mathbb{P}_\theta(\xi(x) = 0) = \int \mathbf{1}_{\xi(x)=0}\xi(x)\beta(\theta)e^{\alpha(\theta).T(x)}\mu(dx) = 0$$

Ainsi $\mathbb{P}_\theta(g(T(x)) \neq 0) = 0$ et ceci pour tout $\theta \in \Theta$, ce qu'il fallait obtenir.

3.2 L'information au sens de Fisher : cas unidimensionnel

On considère un modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta \in \Theta)$ dominé par la mesure σ-finie μ et Θ est un intervalle ouvert de $\mathbb{R}$. On note la vraisemblance $L(x,\theta) = \frac{d\mathbb{P}_\theta}{d\mu}$.

Définition. Sous réserve de bonne définition, l'information de Fisher au point θ est définie par

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\frac{\partial}{\partial\theta}L(x,\theta)}{L(x,\theta)}\right)^2\right]$$

On appelle score la quantité $S = \frac{\frac{\partial}{\partial\theta}L(x,\theta)}{L(x,\theta)} = \frac{\partial}{\partial\theta}\ln L(x,\theta)$. L'information est l'espérance du carré du score et aussi sa variance car

$$\begin{aligned}\mathbb{E}_\theta(S) &= \int_\Omega \frac{\frac{\partial}{\partial\theta}L(x,\theta)}{L(x,\theta)}L(x,\theta)d\mu(x)\\ &= \int_\Omega \frac{\partial}{\partial\theta}L(x,\theta)d\mu(x)\\ &\overset{(*)}{=} \frac{\partial}{\partial\theta}\int_\Omega L(x,\theta)d\mu(x) = \frac{\partial}{\partial\theta}1 = 0\end{aligned}$$

dans le cas où on peut dériver sous l'intégrale pour écrire l'égalité (*).

Proposition 13. *Sous réserve de bonne définition et sous les hypothèses techniques faites dans la preuve ci-dessous, on a :*

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln L(x,\theta)\right)$$

Preuve.

$$\begin{aligned}
\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln L(x,\theta)\right) &= \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}L}{L}\right)\right)\\
&= \mathbb{E}_\theta\left[\partial_\theta\left(\frac{\partial_\theta L}{L}\right)\right] \text{ en simplifiant la notation}\\
&= \mathbb{E}_\theta\left(\frac{\partial_\theta^2 L.L-(\partial_\theta L)^2}{L^2}\right)\\
&= \mathbb{E}_\theta\left(\frac{\partial_\theta^2 L}{L}\right)-I(\theta)=\int_\Omega\frac{\partial^2}{\partial\theta^2}L(x,\theta)\mu(dx)-I(\theta)\\
&= \frac{\partial^2}{\partial\theta^2}(\underbrace{\int_\Omega L(x,\theta)\mu(dx)}_{1})-I(\theta)=-I(\theta)
\end{aligned}$$

en admettant que $\ln L(x,\theta)$ est deux fois dérivable et qu'on peut dériver deux fois sous le signe somme l'intégrale dépendant d'un paramètre: $\int L(x,\theta)\mu(dx)$.

Proposition 14. *Soient $(\Omega_1,\mathcal{F}_1,\mathbb{P}^1_\theta)$ et $(\Omega_2,\mathcal{F}_2,\mathbb{P}^2_\theta)$ deux modèles statistiques dominés respectivement par μ_1 et μ_2 et paramètrés par $\theta\in\Theta$, avec pour informations respectives $I_1(\theta)$ et $I_2(\theta)$. Alors l'information $I(\theta)$ du modèle $(\Omega_1\times\Omega_2,\mathcal{F}_1\otimes\mathcal{F}_2,\mathbb{P}^1_\theta\otimes\mathbb{P}^2_\theta)$ est $I(\theta)=I_1(\theta)+I_2(\theta)$ (sous réserve que les hypothèses d'ordre 2 de la proposition précédente soient vérifiées).*

Preuve. Avec des notations claires $L((x,y),\theta)=L_1(x,\theta).L_2(y,\theta)$ est la vraisemblance de $\mathbb{P}^1_\theta\otimes\mathbb{P}^2_\theta$ par rapport à $\mu_1\otimes\mu_2$. Mais

$$\frac{\partial^2}{\partial\theta^2}\ln L(x,y,\theta)=\frac{\partial^2}{\partial\theta^2}\ln L_1(x,\theta)+\frac{\partial^2}{\partial\theta^2}\ln L_2(y,\theta)$$

D'où

$$\begin{aligned}
-I(\theta) &= \int\int\frac{\partial^2}{\partial\theta^2}\ln L(x,y,\theta)\ \ L_1(x,\theta)L_2(y,\theta)d\mu_1(x)d\mu_2(y)\\
&= \int\int\frac{\partial^2}{\partial\theta^2}\ln L_1(x,\theta)\ \ L_1(x,\theta)L_2(y,\theta)d\mu_1(x)d\mu_2(y)+\ 2^{nd}\text{ terme similaire}\\
&= (\int\frac{\partial^2}{\partial\theta^2}\ln L_1(x,\theta)L_1(x,\theta)d\mu_1(x))\int L_2(y,\theta)d\mu_2(y)+\ldots\text{ par Fubini}\\
&= -I_1(\theta)*1-I_2(\theta)*1
\end{aligned}$$

3.3 L'information au sens de Fisher : cas multidimensionnel, exemples

On suppose maintenant que $(\Omega,\mathcal{F},\mathbb{P}_\theta)$ est un modèle statistique dominé (par μ) paramètré par θ variant dans Θ ouvert de $\mathbb{R}^d$.On appelle maintenant

score le vecteur de $\mathbb{R}^d$ (identifié ensuite à une matrice colonne, comme d'habitude) :

$$S_\theta(x) = \bigtriangledown \ln L(x,\theta) = (\frac{\partial}{\partial\theta_1}\ln L(x,\theta),...,\frac{\partial}{\partial\theta_d}\ln L(x,\theta))$$

en notant comme d'habitude $L(x,\theta)$ la vraisemblance, toujours sous réserve d'existence des dérivées écrites.

Définition. On appelle information de Fisher la matrice $d \times d$ définie par

$$I(\theta) = \mathbb{E}_\theta(S_\theta S'_\theta)$$

On notera que sous les hypothèses techniques habituelles $\mathbb{E}_\theta(S_\theta) = 0$ et $I(\theta)$ est donc aussi la matrice de variance-covariance de S_θ. Il s'agit d'une matrice symétrique positive de terme général :

$$\begin{aligned} I_{ij}(\theta) &= \mathbb{E}_\theta\left(\frac{1}{L(x,\theta)^2}\frac{\partial L(x,\theta)}{\partial\theta_i}\frac{\partial L(x,\theta)}{\partial\theta_j}\right) \\ &= -\mathbb{E}_\theta\left(\frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}\right) \end{aligned}$$

par un calcul similaire à ceux effectués précédemment, sous réserve de certaines conditions techniques.

Exemple : supposons que $\mathbb{P}_\theta$ est la loi normale $\mathcal{N}(m,\sigma^2)$ paramètrée par $\theta = (m,\sigma^2)$. Dans ce cas :

$$\ln L = -\frac{(x-m)^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi - \ln\sigma$$

$$\frac{\partial^2 \ln L}{\partial m^2} = -\frac{1}{\sigma^2}, \quad \frac{\partial^2 \ln L}{\partial m\partial\sigma^2} = -\frac{x-m}{\sigma^4}, \quad \frac{\partial^2 L}{\partial(\sigma^2)^2} = -\frac{(x-m)^2}{(\sigma^2)^3} + \frac{1}{2}\frac{1}{(\sigma^2)^2}$$

$$\mathbb{E}_\theta\left(\frac{\partial^2 L}{\partial(\sigma^2)^2}\right) = -\frac{\sigma^2}{\sigma^6} + \frac{1}{2}\frac{1}{\sigma^4} = -\frac{1}{2\sigma^4}$$

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

3.4 Information et exhaustivité

Théorème 15. *Soit $(\Omega,\mathcal{F},\mathbb{P}_\theta)$ un modèle statistique dominé d'information $I(\theta)$ et T une statistique à valeurs dans $(Y,\mathcal{H})$ et $\mathbb{P}_\theta^T$ la probabilité image de $\mathbb{P}_\theta$ par T, sur $(Y,\mathcal{H})$. Les probabilités $\mathbb{P}_\theta^T$ sont dominées par μ^T image de μ par T et on suppose que l'on peut définir l'information $I^T(\theta)$ de ce modèle image. alors*

$$I^T(\theta) \leq I(\theta)$$

et l'égalité a lieu ssi T est exhaustive.

3.5 Exercices

Exercice 11 (Loi log–normale). On dit qu'une variable aléatoire réelle suit la loi log–normale de paramètres m et σ^2si son logarithme népérien suit une loi normale $\mathcal{N}(m,\sigma^2)$.
a) Calculer la densité d'une telle loi.
b) Un n–échantillon d'une telle loi, paramétré par $\theta = (m,\sigma^2)$ est-il un modèle exponentiel?

Exercice 12 (Propriétés de la fonction score). Soit $(\Omega,\mathcal{F},\mathbb{Q}_\theta)$ un modèle statistique paramétré par θ variant dans Θ intervalle ouvert de $\mathbb{R}$, qu'on suppose dominé par μ. Soit $(\Omega^n,\mathcal{F}^{\otimes n},\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes})$ le modèle statistique correspondant à un n–échantillon $X = (X_1,\ldots,X_n)$. La vraisemblance est

$$L(X,\theta) = \prod_{i=1}^{n} f(X_i,\theta) \text{ où } f(\cdot,\theta) = \frac{d\mathbb{Q}_\theta}{d\mu}$$

Comme d'habitude on appelle fonction score

$$S_n(X,\theta) = \frac{\partial}{\partial\theta} \ln L(X,\theta)$$

Déterminer le comportement asymptotique p.s. de $\frac{1}{n} S_n(X,\theta)$ quand $n \to +\infty$ et la convergence en loi de $\frac{1}{\sqrt{n}} S_n(X,\theta)$ en faisant les hypothèses techniques nécessaires.

Exercice 13 (Information pour une loi de Poisson). Dans le modèle statistique d'un n–échantillon de la loi de Poisson $\mathcal{P}(\theta)$ de paramètre θ, calculer l'information de Fisher sur θ.

Exercice 14 (Information pour une loi de Paréto). Dans le modèle statistique d'un n–échantillon de la loi de densité

$$f(x,\theta) = \frac{1+\theta}{(x+\theta)^2} \; \mathbf{1}_{\{x\geq 1\}}$$

avec $\theta \in \mathbb{R}$, calculer l'information de Fisher sur θ.

Chapitre 4

Estimateurs

Soit $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta \in \Theta)$ un modèle statistique paramétrique. On cherche une statistique $T(X)$ (c'est à dire une fonction mesurable de X qui est ici l'identité sur Ω) qui estime une certaine fonction de θ disons $g(\theta)$. Une telle statistique sera appelée *estimateur* de $g(\theta)$ bien que, pour le moment, nous n'ayons pas spécifié ce qu'on demande vraiment à cette statistique. Nous supposerons que T et $g(\cdot)$ sont à valeurs dans $\mathbb{R}$ ou $\mathbb{R}^d$ et selon, $|\cdot|$ désignera la valeur absolue ou la norme euclidienne canonique.

4.1 Ordre, biais

En choisissant T comme estimateur de $g(\theta)$ on commet une erreur qui est quantifiée par :

$$R_T(\theta) = \mathbb{E}_\theta\left(|T - g(\theta)|^2\right)$$

dite *fonction de risque quadratique* car on a privilégié la fonction de perte $|\cdot|^2$ ce qui est le choix traditionnel. L'estimateur S sera *préférable* à T si

$$\forall\theta \in \Theta,\ R_S(\theta) \leq R_T(\theta)$$

et même *strictement préférable* si en plus, il existe $\theta_1 \in \Theta$ tel que : $R_S(\theta_1) < R_T(\theta_1)$.

Une qualité voulue pour un estimateur T de $g(\theta)$ est d'être *sans biais* ce qui signifie :

$$\mathbb{E}_\theta(T) = g(\theta).$$

Si ce n'est pas le cas on appelle *biais* la quantité $\mathbb{E}_\theta(T) - g(\theta)$. Dans le cas où T est réel et sans biais, on notera que $R_T(\theta) = V_\theta(T)$.

On dira qu'un estimateur est *optimal* s'il est sans biais et de risque quadratique minimal (de variance minimale dans le cas réel) parmi tous les estimateurs sans biais.

Exemple : Soit $(\mathbb{R}^n,\mathcal{B}_{\mathbb{R}^n},\mathbb{Q}_\theta^{\otimes n})$ le modèle statistique du n-échantillon d'une loi $\mathbb{Q}_\theta$. On avait défini l'espérance empirique et la variance empirique modifiée par les formules respectives :

$$\overline{X}_n = \frac{X_1 + ... + X_n}{n}, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Un calcul fait précédemment montre que

$$\mathbb{E}_\theta(\overline{X}_n) = \int x\ \mathbb{Q}_\theta(dx), \quad \mathbb{E}_\theta(S_n^2) = \int x^2 \mathbb{Q}_\theta(dx) - \left(\int x \mathbb{Q}_\theta(dx)\right)^2$$

ce qui montre que $\overline{X}_n$ et S_n^2 sont des estimateurs non biaisés de la moyenne de $\mathbb{Q}_\theta$ et de sa variance, respectivement. Par contre la variance empirique non modifiée

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

est un estimateur biaisé. Toutefois il est asymptotiquement sans biais au sens de la définition ci-dessous.

4.2 Propriétés asymptotiques

Soit (T_n) une suite d'estimateurs pour des modèles du n-échantillon d'une loi $\mathbb{Q}_\theta$. On peut construire tous ces estimateurs sur le même espace $(\mathbb{R}^{\mathbb{N}},\mathcal{B}_{\mathbb{R}^{\mathbb{N}}},\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes\mathbb{N}})$ avec $T_n = T_n(X_1,...,X_n)$, les X_i étant les projections canoniques.

On dit que les estimateurs (T_n) de $g(\theta)$ sont *asymptotiquement sans biais* si

$$\lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = g(\theta).$$

Par exemple $S_n'^2$ est un estimateur de $\mathrm{Var}(\mathbb{Q}_\theta)$ asymptotiquement sans biais puisque $\mathbb{E}_\theta(S_n'^2) = \frac{n-1}{n}\mathrm{Var}(\mathbb{Q}_\theta)$.

Dans ce cadre une propriété souhaitable pour les estimateurs (T_n) de $g(\theta)$ est la suivante . On dira que (T_n) est une suite d'estimateurs de $g(\theta)$ *convergente* si, quand $n \to +\infty$, on a $T_n \to g(\theta)$ en $\mathbb{P}_\theta$–probabilité c'est à dire

$$\forall \eta > 0, \quad \mathbb{P}_\theta(|T_n - g(\theta)| > \eta) \overset{n\to+\infty}{\longrightarrow} 0.$$

Exemples. On a vu dans le chapitre consacré à l'échantillonnage que les moments empiriques, les fractiles empiriques (par exemple la médiane empirique) sont des estimateurs convergents respectivement des moments de $\mathbb{Q}_\theta$, des fractiles de $\mathbb{Q}_\theta$ (par exemple la médiane de $\mathbb{Q}_\theta$). Dans ce cas la convergence a même lieu p.s..

Le théorème de Glivenko-Cantelli affirme que la fonction de répartition empirique est un estimateur convergent de la fonction de répartition au sens

de la convergence p.s. de la norme uniforme, puisqu'il s'agit dans ce cas d'un estimateur dans un espace fonctionnel et non $\mathbb{R}$ ou $\mathbb{R}^d$. Dans la même veine on peut dire que la mesure empirique μ_n tend vers la loi $\mathbb{Q}_\theta$ au sens où p.s. $d(\mu_n,\mathbb{Q}_\theta) \overset{n\to+\infty}{\longrightarrow} 0$ où d désigne une distance induisant la topologie de la convergence étroite des mesures.

4.3 Borne de Fréchet-Darmois-Cramer-Rao (F.D.C.R.) : cas unidimensionnel

Soit $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta\in\Theta)$ un modèle statistique dominé par μ et paramétré par $\Theta\subset\mathbb{R}$. On note $L(x,\theta)$ la vraisemblance. On suppose que l'information de Fisher peut être définie, soit $I(\theta)$ supposée >0. On rappelle que

$$I(\theta)=\mathbb{V}_\theta(S_\theta)$$

où S_θ est le "score" : $S_\theta(x)=\frac{\partial}{\partial\theta}\ln L(x,\theta)$.

Soit T un estimateur de la fonction réelle $g(\theta)$ de θ , supposé sans biais, c'est à dire

$$\mathbb{E}_\theta(T)=\int_\Omega T(x)L(x,\theta)\mu(dx)=g(\theta).$$

On a supposé que $\frac{\partial L}{\partial\theta}$ existe pour définir l'information de Fisher. En supposant en plus que la dérivation sous l'intégrale est possible dans la formule ci-dessus, on trouve que $g(\cdot)$ est dérivable et

$$\begin{aligned} g'(\theta) &= \int_\Omega T(x)\frac{\partial}{\partial\theta}L(x,\theta)\mu(dx)\\ &= \int_\Omega T(x)S_\theta(x)L(x,\theta)\mu(dx)\\ &= \mathbb{E}_\theta(T\ S_\theta)\\ &= \mathrm{Cov}_\theta(T,S_\theta)\end{aligned}$$

en se rappelant que $\mathbb{E}_\theta(T)=g(\theta)$ et $\mathbb{E}_\theta(S_\theta)=0$. L'inégalité de Cauchy-Schwarz donne alors

$$g'(\theta)^2\le\mathbb{V}_\theta(T)\ \ \mathbb{V}_\theta(S_\theta)=\mathbb{V}_\theta(T)\ \ I(\theta)$$

On peut résumer ce calcul dans un énoncé.

Proposition 16 (Inégalité de F.D.C.R.). *Sous les hypothèses détaillées ci-dessus , la variance d'un estimateur T sans biais de $g(\theta)$ est minorée par:*

$$\mathbb{V}_\theta(T)\ge\frac{g'(\theta)^2}{I(\theta)}.$$

Cette borne conduit au vocabulaire suivant. Un estimateur sans biais est dit *efficace* si sa variance est égale à la borne de Fréchet donnée ci-dessus. On dit qu'une suite d'estimateurs est *asymptotiquement efficace* si la suite des variances tend vers la borne de Fréchet.
Exemple. Dans le cas d'un n-échantillon de la loi $\mathcal{N}(m,\sigma^2)$, en supposant successivement que m est inconnu puis σ inconnu, on a déjà calculé les informations respectives de ces deux modèles

$$I(m) = \frac{n}{\sigma^2} \quad et \quad I(\sigma^2) = \frac{n}{2\sigma^4}$$

On connait deux estimateurs sans biais de m et σ^2 : respectivement $\overline{X}_n$ et S_n^2 et on sait que :

$$\mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n}, \quad \mathbb{V}(S_n^2) = \frac{2\sigma^4}{n-1}.$$

Cela prouve que $\overline{X}_n$ est efficace et que S_n^2, sans être efficace, est asymptotiquement efficace.

4.4 Optimalité et efficacité d'un estimateur

Rappelons comme défini précédemment qu'un ESB, estimateur sans biais, est optimal s'il est de variance minimale. Cette variance minimale sera supérieure ou égale à la borne de Fréchet, sous réserve que les hypothèses de l'inégalité de FDCR soient satisfaites. Quand il y a égalité l'estimateur optimal est qualifié d'efficace.

Notons qu'on ne sait pas pour le moment s'il existe un estimateur optimal et a fortiori s'il existe un estimateur efficace. Commençons par un résultat d'unicité.

Théorème 17. *S'il existe un estimateur sans biais optimal, il est unique ($\mathbb{P}_\theta$-p.s. pour tout θ).*

Preuve. Soit T et $\tilde{T}$ deux ESB optimaux supposés réels pour simplifier les écritures. On a donc $\mathbb{V}(T) = \mathbb{V}(\tilde{T})$. Pour λ réel, la considération de l'ESB $T + \lambda(\tilde{T} - T)$ donne

$$\begin{aligned} \forall \lambda \in \mathbb{R},\ \mathbb{V}(T) &\leq \mathbb{V}(T + \lambda(\tilde{T} - T)) \\ &= \mathbb{V}(T) + \lambda^2\, \mathbb{V}(\tilde{T} - T) + 2\lambda\, \mathrm{Cov}(T, \tilde{T} - T) \end{aligned}$$

ce qui conduit à $\mathrm{Cov}(T,\tilde{T} - T) = 0$. Alors :

$$\mathbb{V}(\tilde{T}) = \mathbb{V}(\tilde{T} - T + T) = \mathbb{V}(\tilde{T} - T) + \underbrace{\mathbb{V}(T)}_{\mathbb{V}(\tilde{T})} + 2\underbrace{\mathrm{Cov}(\tilde{T} - T, T)}_{0}.$$

D'où $\mathbb{V}(T - \tilde{T}) = 0$. Dans toute cette preuve, $\mathbb{V}$ désigne la variance sous $\mathbb{P}_\theta$, avec $\theta \in \Theta$. On conclut que $T = \tilde{T}$, $\mathbb{P}_\theta$-p.s. pour tout $\theta \in \Theta$.

Voyons maintenant un procédé qui permet à partir d'un ESB de réduire (au sens large) la variance et dans certains cas d'atteindre l'optimalité.

Théorème 18. *(i) (*Rao-Blackwell*) Soit dans un modèle statistique* $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta\in\Theta)$ *un ESB* T *de* $g(\theta)$ *(réel) et* U *une statistique exhaustive. Alors* $\mathbb{E}(T|U)$ *est un ESB de* $g(\theta)$ *préférable à* T*:*

$$\mathbb{V}_\theta(T)\geq\mathbb{V}_\theta(\mathbb{E}(T|U))\ \textit{pour tout}\ \theta\in\Theta.$$

*(ii) (*Lehmann-Scheffé*) Si en plus* U *est complète, alors* $\mathbb{E}(T|U)$ *est un ESB optimal de* $g(\theta)$*.*

Preuve. L'énoncé du (i) peut sembler ambigü : sous quelle probabilité est calculée $\mathbb{E}(T|U)$. En fait n'importe quelle $\mathbb{P}_\theta$, $\theta\in\Theta$ donne le même résultat car la loi de l'observation X sachant $U=u$ sous $\mathbb{P}_\theta$ ne dépend pas de θ. Alors :

$$\mathbb{E}_\theta\left(\mathbb{E}_\theta\left(T|U\right)\right)=\mathbb{E}_\theta\left(T\right)=g(\theta)$$

donc $\mathbb{E}(T|U)$ est ESB. La formule d'analyse de la variance s'écrit :

$$\mathbb{V}_\theta(T)=\mathbb{V}_\theta(\mathbb{E}_\theta\left(T|U\right))+\mathbb{E}_\theta\left(\mathbb{V}_\theta(T|U)\right).$$

Comme la variance conditionnelle $V_\theta(T|U)$ est positive p.s., on obtient bien l'inégalité annoncée.

(ii) Soit $\tilde{T}$ un ESB de $g(\theta)$. Montrons que $\mathbb{V}_\theta(\tilde{T})\geq\mathbb{V}_\theta(\mathbb{E}(T|U))$. Soit

$$\psi(u)=\mathbb{E}_\theta\left(\tilde{T}|U=u\right)-\mathbb{E}_\theta\left(T|U=u\right)$$

qui est en fait indépendante de θ. Alors $\forall\theta\in\Theta$,

$$\mathbb{E}_\theta\left(\psi(U)\right)=\mathbb{E}_\theta\left(\mathbb{E}_\theta\left(\tilde{T}|U\right)-\mathbb{E}_\theta\left(T|U\right)\right)=\mathbb{E}_\theta\left(\tilde{T}\right)-\mathbb{E}_\theta\left(T\right)=g(\theta)-g(\theta)=0.$$

Comme U est complète cela entraine que $\psi=0$, $\mathbb{P}_\theta^U$–ps. Ainsi $\mathbb{E}(\tilde{T}|U)=\mathbb{E}(T|U)$. Par (i) on a $\mathbb{V}_\theta(\tilde{T})\geq\mathbb{V}_\theta(\mathbb{E}(\tilde{T}|U))=V_\theta(\mathbb{E}(T|U))$ ce qui montre bien l'optimalité de $\mathbb{E}(T|U)$.

Corollaire 19. *Si un ESB est fonction d'une statistique exhaustive complète alors c'est l'unique ESB optimal.*

Exemple de la famille exponentielle. Soit un modèle exponentiel $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta\in\Theta)$ par rapport à une mesure dominante μ ce qui, rappelons le, signifie que la vraisemblance s'écrit :

$$L(x,\theta)=e^{\alpha(\theta).T(x)}\beta(\theta)\xi(x)$$

où $T(x)$, dite statistique naturelle, est à valeurs dans $\mathbb{R}^d$, ainsi que $\alpha(\theta)$. Quant à Θ supposons que c'est une partie d'intérieur non vide de $\mathbb{R}^k$. Alors

on sait que la statistique naturelle $T(x)$ est exhaustive et complète. C'est donc un ESB de :

$$g(\theta) = \mathbb{E}_\theta(T) = \int_\Omega T(x)L(x,\theta)\mu(dx).$$

Supposons par exemple que l'on remplace μ par $\xi(\dot{)}\ \mu(\cdot)$ et qu'on reparamètre le modèle pour remplacer $\alpha(\theta)$ par θ. On a alors

$$L(x,\theta) = e^{-\Phi(\theta)}e^{\theta.T(x)} \quad \text{où } e^{\Phi(\theta)} = \int_\Omega e^{\theta.T(x)}\mu(dx).$$

On dit parfois que le modèle exponentiel est sous forme canonique. Dans ce cas $T(x)$ et θ sont supposés varier dans $\mathbb{R}^d$. On a alors

$$g(\theta) = \mathbb{E}_\theta(T(x)) = e^{-\Phi(\theta)}\int_\Omega T(x)e^{-\theta.T(x)}\mu(dx).$$

Alors la i-ième composante de ce vecteur de $\mathbb{R}^d$ est :

$$\begin{aligned} e^{-\Phi(\theta)}\int_\Omega T_i(x)e^{\sum_j \theta_j.T_j(x)}\mu(dx) &= e^{-\Phi(\theta)}\frac{\partial}{\partial\theta_i}\int_\Omega e^{-\sum_j \theta_j.T_j(x)}\mu(dx) \\ &= e^{-\Phi(\theta)}\frac{\partial}{\partial\theta_i}(e^{\Phi(\theta)}) = \frac{\partial}{\partial\theta_i}\Phi(\theta) \end{aligned}$$

Ainsi $g(\theta) = \nabla\Phi(\theta)$. On pourrait se demander si cette statistique naturelle $T(x)$ qui est un ESB optimal de $g(\theta) = \nabla\Phi(\theta)$ est aussi efficace. Mais il faut noter que pour le moment l'inégalité de FDCR et donc la notion d'efficacité n'ont été définies que pour θ réel (i.e. $d = 1$ ci-dessus). Nous allons donc étendre au cadre multidimensionnel.

4.5 Inégalité de FDCR : cas multidimensionnel

Considérons le modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta\in\Theta)$ où Θ est un ouvert de $\mathbb{R}^d$, dominé par μ et on note $L(x,\theta) = \frac{d\mathbb{P}_\theta}{d\mu}(x)$ la vraisemblance. On rappelle que la matrice d'information de Fisher est, quand elle est définie :

$$I(\theta) = \mathbb{E}_\theta(S_\theta S'_\theta) \text{ où } S_\theta(x) = \nabla_\theta \ln L(x,\theta) = \frac{1}{L(x,\theta)}\nabla_\theta L(x,\theta).$$

Soit alors une statistique T à valeurs dans $\mathbb{R}^p$ et on note :

$$g(\theta) = \mathbb{E}_\theta(T) = \int_\Omega T(x)L(x,\theta)d\mu(x) \quad (\in\mathbb{R}^p).$$

Sous réserve que l'on puisse dériver sous le signe somme :

$$\frac{\partial}{\partial\theta_j}g(\theta) = \int_\Omega T(x)\frac{\partial L(x,\theta)}{\partial\theta_j}d\mu(x) = \mathbb{E}_\theta(T\frac{\partial}{\partial\theta_j}\ln L(x,\theta)) \quad (\in\mathbb{R}^p).$$

Supposons que la matrice d'information de Fisher $I(\theta)$ est inversible et posons

$$Z = T - g(\theta) - W_\theta I(\theta)^{-1} S_\theta,$$

où

$$\begin{aligned} W_\theta = \left(\frac{\partial g_i(\theta)}{\partial \theta_j}\right)_{\substack{1 \le j \le d \\ 1 \le i \le p}} &= \mathbb{E}_\theta\left(T S'_\theta\right) \\ &= \mathbb{E}_\theta\left[(T - g(\theta)) S'_\theta\right] \\ &\quad \text{en utilisant } \mathbb{E}_\theta\left(S'_\theta\right) = 0 \end{aligned}$$

Le vecteur colonne Z de $\mathbb{R}^p$ est centré sous $\mathbb{P}_\theta$, comme $T - g(\theta)$ et S_θ et sa matrice de variance-covariance est une matrice définie positive qui s'écrit :

$$\begin{aligned} \mathbb{E}_\theta(ZZ') &= \mathbb{E}_\theta\left[\left((T - g(\theta)) - W_\theta I_\theta^{-1} S_\theta\right)\left((T - g(\theta))' - S'_\theta I(\theta)^{-1} W'_\theta\right)\right] \\ &= \mathrm{Cov}_\theta(T) - \mathbb{E}_\theta\left((T - g(\theta)) S'_\theta I(\theta)^{-1} W'_\theta\right) - \mathbb{E}_\theta\left[W_\theta I(\theta)^{-1} S_\theta (T\_g(\theta))'\right] \\ &\qquad + \mathbb{E}_\theta\left[W_\theta I(\theta)^{-1} S_\theta S'_\theta I(\theta)^{-1} W'_\theta\right] \\ &= \mathrm{Cov}_\theta(T) - \mathbb{E}_\theta((T - g(\theta)) S'_\theta) I(\theta)^{-1} W'_\theta - W_\theta I(\theta)^{-1}\left[\mathbb{E}_\theta\left((T - g(\theta)) S'_\theta\right)\right]' \\ &\qquad + W_\theta I(\theta)^{-1} \mathbb{E}_\theta\left(S_\theta S'_\theta\right) I(\theta)^{-1} W'_\theta \\ &\quad \text{en spécifiant les matrices aléatoires et celles qui sont déterministes} \\ &= \mathrm{Cov}_\theta(T) - W_\theta I(\theta)^{-1} W'_\theta - W_\theta I(\theta)^{-1} W'_\theta + W_\theta I(\theta)^{-1} I(\theta) I(\theta)^{-1} W'_\theta \\ &= \mathrm{Cov}_\theta(T) - W_\theta I(\theta)^{-1} W'_\theta \end{aligned}$$

On obtient donc le résultat suivant.

Proposition 20. *Soit T un ESB de $g(\theta)$ à valeurs dans $\mathbb{R}^p$. En faisant les hypothèses de régularité ci-dessus on a :*

$$\mathrm{Cov}_\theta(T) \ge W_\theta I(\theta)^{-1} W'_\theta$$

au sens où la différence de ces deux matrices $p \times p$ est symétrique positive, où

$$W_\theta = \left(\frac{\partial}{\partial \theta_1} g | \ldots | \frac{\partial}{\partial \theta_d} g\right) \qquad \textit{matrice } p \times d$$

et $I(\theta)$ est la matrice d'information de Fisher, supposée inversible, du modèle statistique paramètrée par θ variant dans Θ ouvert de $\mathbb{R}^d$.

Quand il y a égalité on dit que l'ESB T est *efficace*. En particulier si T (et $g(\theta)$) sont à valeurs réelles on trouve :

$$\mathbb{V}_\theta(T) \ge \nabla g(\theta)' I(\theta)^{-1} \nabla g(\theta).$$

Exemple de la famille exponentielle. Soit un modèle exponentiel canonique de statistique naturelle T à valeurs dans $\mathbb{R}^d$. La vraisemblance s'écrit donc :

$$L(x,\theta) = \exp(\theta.T(x) - \Phi(\theta)).$$

On a déjà vu que la statistique naturelle T est un ESB optimal de $g(\theta) = \nabla\Phi(\theta)$. Montrons maintenant plus : cet estimateur est *efficace*, c'est à dire que sa matrice de covariance atteint la borne de Cramer-Rao. Rappelons que $I(\theta) = \mathbb{E}_\theta(\nabla \ln L (\nabla \ln L)')$. Or $\ln L = \theta.T(x) - \Phi(\theta)$, d'où $\nabla \ln L = T - \nabla\Phi = T - g(\theta) = T - \mathbb{E}_\theta(T)$ et donc $I(\theta) = \mathrm{Cov}_\theta(T)$. Par ailleurs :

$$W_\theta = \mathbb{E}_\theta\left[(T - g(\theta))(\nabla \ln L)'\right] = \mathbb{E}_\theta\left[(T - g(\theta))(T - g(\theta))'\right] = \mathrm{Cov}_\theta(T).$$

On a donc égalité entre $\mathrm{Cov}_\theta(T)$ et $W_\theta I(\theta) W_\theta'$ ce qui conclut.

4.6 comportement asymptotique des estimateurs

Au paragraphe 2, on a dit que (T_n) est une suite convergente d'estimateurs de $g(\theta)$ dans le modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta \in \Theta)$ si $T_n \longrightarrow g(\theta)$ en $\mathbb{P}_\theta$–probabilité.

On remarquera que, pour que ceci ait lieu, il suffit que

(i) $\mathbb{E}_\theta(T_n) \longrightarrow g(\theta)$ i.e. (T_n) asymptotiquement sans biais .

(ii) $\mathbb{V}_\theta(T_n) \longrightarrow 0$.

En effet, pour tout $\eta > 0$,

$$\begin{aligned}\mathbb{P}_\theta\left(|T_n - g(\theta)| > \eta\right) &\leq \mathbb{P}_\theta\left(|T_n - \mathbb{E}_\theta(T_n)| > \frac{\eta}{2}\right) + \mathbb{P}_\theta\left(|\mathbb{E}_\theta(T_n) - g(\theta)| > \frac{\eta}{2}\right)\\ &\leq \frac{1}{(\eta/2)^2}\mathbb{V}_\theta(T_n) + \mathbf{1}_{\{|\mathbb{E}_\theta(T_n) - g(\theta)| > \frac{\eta}{2}\}} \overset{n\to+\infty}{\longrightarrow} 0\end{aligned}$$

La proposition qui suit confirme l'idée intuitive que les ESB optimaux (i.e. de variance minimale) construits sur un nombre de plus en plus grand d'observations vont tendre vers la vraie valeur du paramètre. Dans ce qui suit $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta \in \Theta)$ est un modèle statistique avec θ réel sur lequel on dispose de $X_1,X_2,...$ suite de statistiques indépendantes de loi $\mathbb{Q}_\theta$; par exemple on fait une construction par produit infini $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes\mathbb{N}}$ et les X_i sont les projections canoniques.

Proposition 21. *Soit (T_n) une suite d'ESB optimaux de $g(\theta)$ au sens où, pour tout n, T_n est optimal dans la classe des ESB fonctions de $X_1,...,X_n$ et en particulier $T_n = T_n(X_1,...,X_n)$.*

Alors T_n est une suite convergente d'estimateurs de $g(\theta)$.

Preuve. La statistique $S_n = \frac{1}{n}\sum_{k=1}^n T_1(X_k)$ a pour espérance $g(\theta)$ et pour variance $\frac{1}{n}V_\theta(T_1)$. Mais T_n étant optimal, sa variance est inférieure à celle de l'ESB S_n:

$$\mathbb{V}_\theta(T_n) \leq \mathbb{V}_\theta(S_n).$$

et le résultat découle de la remarque précédant la proposition.

Définition : Une suite (T_n) d'estimateurs réels de $g(\theta)$ est dite *asymptotiquement normale* si

$$\sqrt{n}(T_n - g(\theta)) \longrightarrow \mathcal{N}(0,v(\theta))$$

en loi, sous $\mathbb{P}_\theta$ et *normale asymptotiquement efficace* si $v(\theta) = \frac{g'^2(\theta)}{I(\theta)}$ borne de Fréchet (en supposant ici que θ est réel).

Notons toutefois que si la variance d'un ESB T_n est minorée par la borne de Fréchet $\frac{g'^2(\theta)}{nI(\theta)}$, la variance asymptotique $v(\theta)$ peut être plus faible que $\frac{g'^2(\theta)}{I(\theta)}$ (estimateur super efficace).

Exemple de suite d'estimateurs asymptotiquement normale : $(\overline{X}_n)$ moyennes empiriques pour une loi de variance finie : c'est le théorème central limite.

4.7 Exercices

Exercice 15 (Cas biaisé). Soit T un estimateur de réel $g(\theta)$ dans le modèle statistique dominé $(\Omega,\mathcal{F},\mathbb{P}_\theta;\ \theta \in \Theta)$ où l'on peut définir l'information de Fisher $I(\theta)$ supposée > 0. On suppose que T peut être biaisé, de biais $B(\theta) = \mathbb{E}_\theta(T) - g(\theta)$.
1) Montrer que le risque quadratique de l'estimateur T est

$$R_T(\theta) = \mathbb{V}_\theta(T) + B(\theta)^2$$

2) En faisant les hypothèses techniques nécessaires, montrer la minoration

$$R_T(\theta) \geq B(\theta)^2 + \frac{(g'(\theta) + B'(\theta))^2}{I(\theta)}$$

Exercice 16 (Loi de Poisson). Considèrons un n–échantillon de la loi de Poisson de paramètre θ. Comme on le sait θ est égal à l'espérance et on peut l'estimer par la moyenne empirique $\overline{X}_n$ mais θ est aussi égal à la variance et on peut l'estimer par la variance empirique (modifiée) S_n^2. Lequel de ces deux estimateurs est préférable?
Exercice 17 (Recherche d'un ESB). Soit $(\mathbb{R},\mathcal{B}(\mathbb{R}),\mathbb{P}_\theta;\ \theta \in]0,1[)$ le modèle statistique d'un échantillon de taille 1 de la loi binomiale de paramètres $m \geq 1$ fixé et θ. Pouvez vous trouver un estimateur sans biais de $g(\theta) = 1/\theta$?
Exercice 18 (Estimation optimale ou efficace). On considère un n–échantillon $X = (X_1,\ldots,X_n)$ de la loi de densité

$$L(x,\theta) = \theta\ a\ x^{a-1}\ e^{-\theta\, x^a}\ \ \mathbf{1}_{\mathbb{R}_+}(x)$$

où $a > 0$ est un paramètre connu et θ est inconnu.
1) Montrer que ce modèle est un modèle exponentiel dont on précisera la statistique naturelle T.
2) Trouver une statistique exhaustive et complète.
3) Montrer que $(n-1)/T$ est un ESB de θ de variance minimale que l'on calculera.
4) Montrer que cette variance est $>$ à $1/I_n(\theta)$ où $/I_n(\theta)$ est l'information de Fisher. Commenter.

Exercice 19 (Optimisation d'un estimateur). Un industriel produit des articles dont une proportion p inconnue est défectueuse. Il livre ses produits par caisse de k articles. Le client refuse une caisse quand le nombre d'articles défectueux est > 2.
1) Calculer la probabilité $g(p)$ qu'une caisse soit acceptée par le client.
2) Soit X_1,X_2,... les nombres d'articles défectueux dans les caisses successivement livrées. Quelle est la vraisemblance de $X = (X_1,\ldots,X_n)$? Trouver une statistique exhaustive U.
3) Montrer que $T = \mathbf{1}_{\{0,1,2\}}(X_1)$ est un ESB de $g(p)$. Que pensez vous de sa qualité?
4) Calculer l'amélioré de Rao-Blackwell de T par U. Cet estimateur est il optimal?

Exercice 20 (Validité asymptotique de FDCR). Sur le modèle statistique d'un échantillon de taille infinie de la loi normale $\mathcal{N}(\theta,\sigma^2)$ où θ est inconnu, on considère les moyennes empiriques $\overline{X}_n$. On fixe $\alpha > 0$ et on pose

$$S_n = \overline{X}_n \, \mathbf{1}_{\{|\overline{X}_n| \geq n^{-1/4}\}} + \alpha \, \overline{X}_n \, \mathbf{1}_{\{|\overline{X}_n| < n^{-1/4}\}}$$

Montrer que S_n est une suite convergente d'estimateurs de θ, qui est asymptotiquement normale et calculer la variance asymptotique. Pour $\theta = 0$, le résultat obtenu est il en contradiction avec l'inégalité de FDCR?

Chapitre 5

Estimation par maximum de vraisemblance et autres méthodes

5.1 Définition de l'estimateur du maximum de vraisemblance

Pour saisir le sens de la définition d'un tel estimateur, considérons l'exemple introductif suivant. Une personne va faire un tirage d'une boule dans l'une de ces deux urnes :

- Urne 1 : 10 boules blanches et 1 noire,
- Urne 2 : 10 boules noires et 1 blanche.

La statistique qui donne la couleur de la boule tirée est notée $X \in \{B,N\}$. Sa loi est $\mathbb{P}_\theta, \theta \in \{1,2\}$ avec :

$$\mathbb{P}_1 = \frac{10}{11}\delta_{\{B\}} + \frac{1}{11}\delta_{\{N\}}, \quad \mathbb{P}_2 = \frac{1}{11}\delta_{\{B\}} + \frac{10}{11}\delta_{\{N\}},$$

selon l'urne choisie pour le tirage. Comment estimer $\theta \in \{1,2\}$ au vu du tirage? Le bon sens dicte la règle suivante : $X = B \Longrightarrow \theta = 1, \quad X = N \Longrightarrow \theta = 2$ C'est à dire que l'on choisit θ qui maximise la probabilité que X prenne la valeur observée.

Définition. Soit $(\Omega,\mathcal{F}, \mathbb{P}_\theta, \theta \in \Theta)$ un modèle statistique dominé, $X = \mathrm{id}_\Omega$ est l'observation et $L(x,\theta)$ la vraisemblance. On dit que $\hat{\theta}$ est un *estimateur du maximum de vraisemblance* de θ si c'est une statistique qui vérifie :

$$L(X,\hat{\theta}) = \sup_{\theta\in\Theta} L(X,\theta).$$

On écrira EMV en abrégé.

A ce niveau de généralité rien ne prouve qu'un tel estimateur existe et peut se calculer. En général il ne sera pas unique ni dépourvu de biais etc... Toutefois dans les cas usuels les calculs pourront être faits et on prouvera de bonnes propriétés. L'estimation par maximum de vraisemblance est le procédé le plus utilisé en estimation.

Exemple : Soit $X = (X_1,...,X_n)$ un échantillon de la loi uniforme sur $[\theta - \frac{1}{2},\theta + \frac{1}{2}]$ La vraisemblance s'écrit :

$$L(X,\theta) = \prod_{i=1}^{n} \mathbf{1}_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(X_i) = \mathbf{1}_{\{\theta-\frac{1}{2}\leq X_{(1)}\leq X_{(n)}\leq\theta+\frac{1}{2}\}}$$

où $(X_{(1)},...,X_{(n)})$ est la statistique d'ordre de $(X_1,...,X_n)$. Autrement dit $(x_{(1)},...,x_{(n)})$ est le réordonnement croissant de $(x_1,...,x_n)$ et donc en particulier $x_{(1)} = \min_{1\leq i\leq n} x_i$, $x_{(n)} = \max_{1\leq i\leq n} x_i$. Toute statistique $\hat{\theta}$ telle que $x_{(n)} - \frac{1}{2} \leq \hat{\theta} \leq x_{(1)} + \frac{1}{2}$ sera un EMV.

5.2 Propriétés de l'EMV

Propriété. S'il existe une statistique exhaustive T et si le procédé de recherche d'un EMV $\hat{\theta}$ aboutit, il va conduire à un $\hat{\theta}$ fonction de T.

En effet dans ce cas $L(X,\theta) = g(X)h(T(X),\theta)$ par le théorème de factorisation de Neyman-Fisher. A X fixé, maximiser $L(X,\theta)$ revient à maximiser $h(T(X),\theta)$ qui ne dépend que de $T(X)$.

La recherche de l'EMV est facilitée quand la vraisemblance est dérivable par rapport à θ (dans le cas θ réel) car alors :

$$\left[\frac{\partial}{\partial\theta} L(X,\theta)\right]_{\theta=\hat{\theta}} = 0$$

et pour θ vectoriel cela s'écrit : $\nabla_\theta L(X,\hat{\theta}) = 0$. Quand la vraisemblance est > 0 pour x variant sur un domaine fixé, le maximum de L est obtenu quand $\ln L$ admet un maximum et les équations nécessaires ci-dessus s'écrivent :

$$\frac{\partial}{\partial\theta} \ln L(X,\theta) = 0 \quad \text{et dans le cas vectoriel } \nabla_\theta \ln L(X,\hat{\theta}) = 0.$$

C'est à dire avec la notation du "score" : $S(X,\hat{\theta}) = 0$. L'équation ci-dessus est appelée *équation de la log-vraisemblance.*

Exemple : *famille exponentielle.* Considérons une famille exponentielle sous forme "canonique" c'est à dire que la vraisemblance s'écrit :

$$L(x,\theta) = \exp(\theta.T(x) - \Phi(\theta))$$

par rapport à une mesure dominante μ, avec $\Phi(\theta) = \ln\left(\int e^{\theta.T} d\mu\right)$. La log-vraisemblance $\ln L(X,\theta) = \theta.T(X) - \Phi(\theta)$ est dérivable en θ et

$$\nabla_\theta \ln L(X,\theta) = T(X) - \nabla\Phi(\theta) = T(X) - \mathbb{E}_\theta(T).$$

Donc un EMV sera une fonction de T, notons la $\hat{\theta}(T)$, qui vérifie

$$T(X) = \mathbb{E}_{\hat{\theta}(T(X))}(T),$$

c'est à dire que la fonction $t \mapsto \hat{\theta}(t)$ est solution de

$$t = \mathbb{E}_{\hat{\theta}(t)}(T)$$

(p.p. par rapport à la loi de T sous μ). De plus la log-vraisemblance $\ln L(x,\theta)$ est deux fois dérivable (différentiable) en θ et la différentielle seconde est $(-)$ la différentielle seconde de $\Phi(\theta)$, c'est à dire la matrice de terme général :

$$\begin{aligned}
-\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\Phi(\theta) &= -\frac{\partial}{\partial\theta_i}\frac{\int T_j e^{\theta.T} d\mu}{\int e^{\theta.T} d\mu} \\
&= -\frac{\int T_i T_j e^{\theta.T} d\mu \int e^{\theta.T} d\mu - \int T_i e^{\theta.T} d\mu \int T_j e^{\theta.T} d\mu}{\left(\int e^{\theta.T} d\mu\right)^2} \\
&= -\int T_i T_j L(X,\theta) d\mu + \int T_i L(X,\theta) d\mu \int T_j L(X,\theta) d\mu \\
&= -\mathrm{Cov}_\theta(T_i,T_j)
\end{aligned}$$

Il s'agit donc d'une matrice symétrique négative. Si on suppose qu'elle est en fait définie négative pour tout $\theta \in \Theta$, on est assuré que la solution de l'équation de log-vraisemblance sera un maximum global.

Revenons au cas général pour essayer de formuler un résultat général, attendu bien sûr qu'un certain nombre d'hypothèses de régularité et d'intégrabilité sont satisfaites.

Proposition 22. *Soit un modèle statistique $(\tilde{\Omega},\mathbb{P}_\theta,\ \theta \in \Theta)$ sur lequel on dispose d'un échantillon infini $X_1,X_2,X_3,\ldots$ d'une loi $\mathbb{Q}_\theta$ sur Ω ayant une densité f_θ par rapport à une mesure μ. On note $L_n(X,\theta)$ la vraisemblance du n-échantillon de la loi $\mathbb{Q}_\theta, \theta \in \Theta$ où Θ est un ouvert de $\mathbb{R}$.*

Sous réserve d'hypothèses de régularité et d'intégrabilité détaillées dans la preuve qui suit, on peut construire une suite $(\hat{\theta}_n)$ de solutions (respectives) des équations de (log)-vraisemblance:

$$\frac{\partial}{\partial\theta} \ln L_n(X,\hat{\theta}_n) = 0$$

qui converge p.s. vers θ, sous $\mathbb{P}_\theta$ et $(\hat{\theta}_n)$ est une suite asymptotiquement normale et efficace de θ : sous $\mathbb{P}_\theta$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{(loi)} \mathcal{N}(0,\frac{1}{I(\theta)}).$$

Preuve. Notons déjà que :

$$\ln L_n(X,\theta) = \sum_{i=1}^{n} \ln f_\theta(X_i).$$

Cela entraine par la loi des grands nombres que, quand $n \to +\infty$, $\mathbb{P}_\theta - p.s.$,

$$\frac{1}{n} \ln L_n(X,\theta) \longrightarrow \mathbb{E}_\theta (\ln f_\theta(X_1)) = \int_\Omega \mathbb{Q}_\theta(dx) \ln f_\theta(x)$$

en supposant la finitude du membre de droite.

Soit $\theta_0 \in \Theta$. Si l'application $\theta \mapsto f_\theta(x)$ est constante, $\mathbb{Q}_\theta(dx)$-p.s., pour θ voisin de θ_0 éventuellement uniquement à droite ou à gauche , on peut choisir comme solutions de l'équation de Log-vraisemblance $\hat{\theta}_n(X) = \theta_0$ qui converge bien p.s. vers θ_0. Excluons maintenant ce cas. Pour tout $\varepsilon > 0$, il existe δ_1 et $\delta_2 > 0$ tels que

$$\theta_0 - \varepsilon < \theta_0 - \delta_1 < \theta_0 < \theta_0 + \delta_2 < \theta_0 + \varepsilon \text{ et } \frac{f_{\theta_0-\delta_1}(x)}{f_{\theta_0}(x)} \text{ et } \frac{f_{\theta_0+\delta_1}(x)}{f_{\theta_0}(x)}$$

sont non constantes. On en déduit l'inégalité de Jensen stricte :

$$\mathbb{E}_{\theta_0} \left(\ln \frac{f_{\theta_0-\delta_1}(X_1)}{f_{\theta_0}(X_1)} \right) < \ln \mathbb{E}_{\theta_0} \left(\frac{f_{\theta_0-\delta_1}(X_1)}{f_{\theta_0}(X_1)} \right).$$

Mais le membre de droite vaut :

$$\begin{aligned} \ln \mathbb{E}_{\theta_0} \left(\frac{f_{\theta_0-\delta_1}(X_1)}{f_{\theta_0}(X_1)} \right) &= \ln \left(\int_\Omega \frac{f_{\theta_0-\delta_1}(x)}{f_{\theta_0}(x)} f_{\theta_0}(x) \mu(dx) \right) \\ &= \ln \left(\int_\Omega \mathbb{Q}_{\theta_0-\delta_1}(dx) \right) \\ &= \ln 1 = 0 \end{aligned}$$

On a donc

$$\mathbb{E}_{\theta_0} (\ln f_{\theta_0-\delta_1}(X_1)) - \mathbb{E}_{\theta_0} (\ln f_{\theta_0}(X_1)) < 0.$$

Mais cette quantité est la limite de :

$$\frac{1}{n} \left(\ln L_n(X,\theta_0 - \delta_1) - \ln L_n(X,\theta_0) \right).$$

On en déduit que pour n assez grand, $L_n(X,\theta_0 - \delta_1) < L_n(X,\theta_0)$ et de façon similaire, toujours pour n grand , $L_n(X,\theta_0 + \delta_2) < L_n(X,\theta_0)$.

Si on suppose que $\theta \mapsto f_\theta(x)$ est dérivable pour μ-presque tout x, l'application $\theta \mapsto \ln L_n(X,\theta)$ est aussi dérivable p.s..

Or les inégalités qui précèdent montrent que cette application admet un maximum en $\hat{\theta}_n(X) \in]\theta_0 - \delta_1, \theta_0 + \delta_2[$ et en ce point l'équation de log-vraisemblance est satisfaite :

$$\frac{\partial}{\partial\theta} \ln L_n(X, \hat{\theta}_n(X)) = 0, \quad \text{avec } \hat{\theta}_n(X) \in]\theta_0 - \varepsilon, \theta_0 + \varepsilon[.$$

En choisissant ε de la forme $\frac{1}{n}$ ou toute autre suite tendant vers 0, on a par construction $\hat{\theta}_n(X) \longrightarrow \theta_0$, $\mathbb{P}_{\theta_0} - p.s.$.

Montrons maintenant le caractère asymptotiquement gaussien et efficace de $\hat{\theta}_n$. On suppose maintenant que $\theta \mapsto f_\theta(x)$ est deux fois dérivable pour μ-presque tout x. Dans le reste de cette preuve on notera les dérivées par rapport à θ avec un prime '. On suppose que $\theta \mapsto f''_\theta(x)$ est lipschitzienne en θ, avec une constante de Lipschitz qui ne dépend pas de x.

On réécrit alors l'équation de la log-vraisemblance et on la complète en utilisant l'inégalité des accroissements finis :

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \ln L_n(X, \hat{\theta}_n(X)) = \sum_{i=1}^{n} (\ln f_{\hat{\theta}_n})'(X_i) \\ &= \sum_{i=1}^{n} \left\{ (\ln f_{\theta_0})'(X_i) + (\hat{\theta}_n - \theta_0)(\ln f_{\theta_n^i})''(X_i) \right\} \end{aligned}$$

où θ_n^i est compris entre θ_0 et $\hat{\theta}_n$. On réécrit cela :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\ln f_{\theta_0})'(X_i)}{\frac{1}{n} \sum_{i=1}^{n} (\ln f_{\theta_n^i})''(X_i)}$$

Les variables apparaissant au numérateur sont indépendantes de même loi et d'espérance :

$$\begin{aligned} \mathbb{E}_{\theta_0} \left((\ln f_{\theta_0})'(X_i) \right) &= \int_\Omega \frac{f'_{\theta_0}(x)}{f_{\theta_0}(x)} \mathbb{Q}_{\theta_0}(dx) \\ &= \int f'_{\theta_0}(x) \mu(dx) \\ &= \left[\frac{\partial}{\partial\theta} \int f_\theta(x) \mu(dx) \right]_{\theta=\theta_0} \\ &= 0 \end{aligned}$$

en supposant, comme dans le cours sur l'information, que la dérivation sous l'intégrale est possible. Alors la variance de ces variables qu'on supposera exister, est

$$V_{\theta_0} \left((\ln f_{\theta_0})' \right) = I(\theta_0)$$

qu'on supposera > 0 . Alors le théorème central limite donne :

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\ln f_{\theta_0})'(X_i) \longrightarrow \mathcal{N}(0, I(\theta_0)).$$

Par ailleurs, on a aussi avec les mêmes hypothèses techniques supposées que dans le cours sur l'information :

$$I(\theta_0) = -\mathbb{E}_{\theta_0} \left[(\ln f_{\theta_0})''(X_1) \right].$$

On écrit ensuite :

$$(\ln f_{\theta_n^i})''(X_i) = (\ln f_{\theta_0})''(X_i) + \underbrace{O(|\theta_n^i - \theta_0|)}_{O(|\hat{\theta}_n - \theta_0|)}$$

Alors le dénominateur qui apparaît ci-dessus s'écrit :

$$\frac{1}{n} \sum_{i=1}^{n} (\ln f_{\theta_n^i})''(X_i) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} (\ln f_{\theta_0})''(X_i)}_{\stackrel{p.s.}{\to} \mathbb{E}_{\theta_0}((\ln f_{\theta_0})''(X_1)) = -I(\theta_0)} + \underbrace{O(|\hat{\theta}_n - \theta_0|)}_{\stackrel{p.s.}{\to} 0}$$

En combinant cela avec le comportement du numérateur on trouve, sous $\mathbb{P}_{\theta_0}$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \stackrel{(loi)}{\longrightarrow} \mathcal{N}(0, \frac{1}{I(\theta_0)})$$

ce qu'il fallait prouver.

Il existe des versions multi-dimensionnelles de ce résultat, i.e. quand $\Theta \subset \mathbb{R}^d$ où l'on remplace $\frac{1}{I(\theta)}$ par $I(\theta)^{-1}$.

Montrons maintenant que dans les situations favorables l'EMV est le meilleur. Pour la simplicité, on va se placer dans un modèle paramétré unidimensionnel.

Proposition 23. *Si il existe un ESB efficace, cet estimateur coïncide avec l'EMV (sous réserve que les hypothèses de régularité de l'inégalité de FDCR soient satisfaites).*

Preuve. Soit T un ESB de θ efficace. Rappelons que l'inégalité de FDCR s'obtient de la façon suivante :

$$\theta = \mathbb{E}_\theta(T) = \int T(x) L(x, \theta) \mu(dx)$$

d'où

$$\begin{aligned} 1 &= \int T(x) \frac{\partial L(x,\theta)}{\partial \theta} \mu(dx) \\ &= \int (T(x) - \theta) \frac{\partial L}{\partial \theta} \mu(dx) \quad \text{car } \int \frac{\partial L}{\partial \theta} d\mu = 0 \\ &= \mathbb{E}_\theta \left((T - \theta) \frac{\partial \ln L}{\partial \theta} \right) \end{aligned}$$

et on utilise ensuite l'inégalité de Cauchy-Schwarz:

$$1 \leq V_\theta(T)\ \ \mathbb{E}_\theta\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = V_\theta(T)\ \ I(\theta)$$

Si T est efficace $V_\theta(T) = \frac{1}{I(\theta)}$ et l'inégalité de Cauchy-Schwarz est en fait une égalité ce qui implique l'existence de $c(\theta)$ telle que

$$T(X) - \theta = c(\theta)\frac{\partial}{\partial \theta}\ln L(X), \quad \mathbb{P}_\theta\text{-p.s..}$$

Mais $\hat{\theta}$ l'EMV est solution de l'équation de la log-vraisemblance $\frac{\partial \ln L}{\partial \theta}(\hat{\theta}) = 0$ et on trouve

$$T(X) - \hat{\theta}(X) = 0 \quad \mathbb{P}_\theta\text{-p.s..}$$

comme annoncé.

Si on veut estimer une fonction de θ, disons $g(\theta)$ on peut utiliser l'estimateur défini comme suit.
Définition. Si $\hat{\theta}$ est un EMV de θ, on appelle EMV de $g(\theta)$ la statistique $g(\hat{\theta})$.

Cette définition se base sur la remarque suivante. Notons $\beta = g(\theta)$ et supposons que β réalise un nouveau paramétrage du modèle statistique i.e. g est injective. Notons alors $L_\beta(x)$ la vraisemblance c'est à dire $L_\beta(x) = L(x,\theta)$ quand $\beta = g(\theta)$. On sait que pour tout $\theta \in \Theta$,

$$L(X,\theta) \leq L(X,\hat{\theta}(X)).$$

Notons $\hat{\beta} = g(\hat{\theta}(X))$. L'inégalité précédente s'écrit :

$$L_\beta(X) \leq L_{\hat{\beta}}(X).$$

ce qui prouve bien que $\hat{\beta} = g(\hat{\theta})$ est un EMV pour le modèle paramétré par β.

5.3 Retour sur l'échantillon gaussien

Reprenons le cas d'un n-échantillon de la loi gaussienne $\mathcal{N}(m,\sigma^2)$. Il s'agit d'un modèle exponentiel puisqu'on peut écrire la vraisemblance sous la forme :

$$\begin{aligned} L(X,m,\sigma^2) &= (2\pi\sigma^2)^{-\frac{n}{2}}\ \exp -\frac{1}{2\sigma^2}\sum_i (X_i - m)^2 \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\ \exp\left(-\frac{1}{2\sigma^2}, \frac{m}{\sigma^2}\right).\left(\sum_i X_i^2, \sum_i X_i\right) - \frac{nm^2}{2\sigma^2} \end{aligned}$$

On voit ainsi que le paramétrage naturel est $\left(-\frac{1}{2\sigma^2},\frac{m}{\sigma^2}\right)$ et la statistique naturelle $\left(\sum_i X_i^2,\sum_i X_i\right)$. Cette statistique est exhaustive et complète. Notons que

$$\left(\sum_i X_i^2,\sum_i X_i\right) = (nS_n'^2 + n\overline{X}_n^2,n\overline{X}_n) = ((n-1)S_n^2 + n\overline{X}_n^2,n\overline{X}_n)$$

ce qui montre que la statistique $(\overline{X}_n,S_n^2)$ est aussi exhaustive et complète. Il s'ensuit que c'est un estimateur optimal de son espérance : (m,σ^2).

Passons à l'estimation par maximum de vraisemblance. La log-vraisemblance s'écrit :

$$\ln L(X,m,\sigma^2) = cste - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i - m)^2$$

d'où

$\frac{\partial}{\partial m}\ln L = \frac{1}{\sigma^2}\sum_i (X_i - m)$

$\frac{\partial}{\partial(\sigma^2)}\ln L = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (X_i - m)^2$

qui a pour solution : $\begin{cases} \sum_i (X_i - \hat{m}) = 0 \text{ i.e. } \hat{m} = \overline{X}_n \\ et \\ \frac{1}{n}\sum_i (X_i - \hat{m})^2 = \hat{\sigma}^2 \quad \text{i.e} \quad \hat{\sigma}^2 = S_n'^2 \end{cases}$

Ce qui montre que l'EMV ne peut être que $(\overline{X}_n,S_n'^2)$.

Par ailleurs on sait dans le cas gaussien que :

– $\overline{X}_n$ et S_n^2 (ou $S_n'^2$) sont indépendants,

– $\overline{X}_n \overset{(loi)}{=} \mathcal{N}(m,\frac{\sigma^2}{n})$,

– $n\frac{S_n'^2}{\sigma^2} = \frac{(n-1)S_n^2}{\sigma^2} \overset{(loi)}{=} \mathcal{X}^2(n-1)$.

D'où l'on tire que $\sqrt{n}(\overline{X}_n - m) \overset{(loi)}{\longrightarrow} \mathcal{N}(0,\sigma^2)$ (en fait c'est une égalité), et

$$\frac{1}{\sqrt{n-1}}\left((n-1)\frac{S_n^2}{\sigma^2} - (n-1)\right) \overset{(loi)}{\longrightarrow} \mathcal{N}(0,2)$$

par le théorème central limite i.e. $\sqrt{n-1}(S_n^2 - \sigma^2) \overset{(loi)}{\longrightarrow} \mathcal{N}(0,2\sigma^4)$ ou encore $\sqrt{n}(S_n^2 - \sigma^2) \overset{(loi)}{\longrightarrow} \mathcal{N}(0,2\sigma^4)$ et $\sqrt{n}(S_n'^2 - \sigma^2) \overset{(loi)}{\longrightarrow} \mathcal{N}(0,2\sigma^4)$. Par indépendance on peut déduire le comportement du couple des lois limites des marginales :

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - m \\ S_n'^2 - \sigma^2 \end{pmatrix} \overset{(loi)}{\longrightarrow} \mathcal{N}\left(0,\begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right)$$

Cette matrice de variance covariance est précisément l'inverse de la matrice d'information de la loi $\mathcal{N}(m,\sigma^2)$.

L'estimateur $\begin{pmatrix} \overline{X}_n \\ S'^2_n \end{pmatrix}$ (ou $\begin{pmatrix} \overline{X}_n \\ S^2_n \end{pmatrix}$ qui est sans biais) est un estimateur de $\begin{pmatrix} m \\ \sigma^2 \end{pmatrix}$ asymptotiquement normal et asymptotiquement efficace.

Question : la solution de la log-vraisemblance $\begin{pmatrix} \overline{X}_n \\ S'^2_n \end{pmatrix}$ réalise-t-elle vraiment le maximum de la vraisemblance?

On peut déjà noter que la hessienne en ce point est

$$\begin{pmatrix} -\frac{n}{S'^2_n} & 0 \\ 0 & -\frac{n}{2(S'^2_n)^2} \end{pmatrix}$$

qui est bien définie négative.

L'estimateur considéré ici est un exemple d'une méthode plus générale que nous étudions maintenant.

5.4 Estimation par la méthode des moments

Soit $(X_1,...,X_n)$ un n-échantillon d'une loi $\mathbb{Q}_\theta$, $\theta \in \Theta$ inconnu dont on note $m_1(\theta),...,m_k(\theta)$ les k premiers moments. Notons aussi les moments empiriques $\mu_1(X),..,\mu_k(X)$.

Définition. On appelle estimateur de θ par la méthode des moments une statistique $\hat{\theta}(X)$ telle que $\forall i \leq k, \quad \mu_i(X) = m_i(\hat{\theta}(X))$.

On peut travailler avec les moments centrés ou non centrés ou même panacher. L'ordre k dépendra de la dimension de Θ afin que les équations $\forall i \leq k, \quad \mu_i(X) = m_i(\hat{\theta}(X))$ aient une solution unique.

Par exemple à l'ordre 2 cela revient à prendre pour $\hat{\theta}$ la solution de :

$$\begin{aligned} \overline{X}_n &= \int x \mathbb{Q}_\theta(dx) \\ S'^2_n &= \int \left(x - \int x' \mathbb{Q}_\theta(dx') \right)^2 \mathbb{Q}_\theta(dx) \end{aligned}$$

On peut montrer que, sous certaines conditions, ces estimateurs sont convergents et asymptotiquement normaux mais pas en général asymptotiquement efficace. Dans le cadre de ce cours on se contentera de traiter des exemples en TD.

5.5 Exercices

Exercice 21 (Loi de Paréto). On considère un n–échantillon de la loi de densité :

$$f(x,\theta) = \frac{1}{\theta}\, x^{-\left(1+\frac{1}{\theta}\right)}\, \mathbf{1}_{\{x\geq 1\}}$$

1) De quel type de modèle s'agit-il?
2) Trouver l'estimateur de maximum de vraisemblance $\hat{\theta}$. Est ce une statistique exhaustive?
3) Calculer l'information de Fisher du modèle.
4) $\hat{\theta}$ est il efficace? Est ce conforme aux résultats du cours?

Exercice 22 (Loi exponentielle décalée). Considérons un n–échantillon $X = (X_1,\ldots,X_n)$ de la loi de densité

$$f(x,\theta) = e^{\theta - x}\, \mathbf{1}_{\{x\geq\theta\}}$$

1) Trouver l'estimateur de maximum de vraisemblance $\hat{\theta}$ de θ.
2) Est il sans biais? Sinon, donner un ESB $\tilde{\theta}$ de θ.
3) Calculer la variance de $\tilde{\theta}$.
4) Calculer l'information de Fisher du modèle et comparer avec le résultat obtenu en 3). Que se passe-t-il?
5) Soit T un ESB de θ. Trouver un minorant de la variance de T en partant de $\delta = \mathbb{E}_{\theta+\delta}(T) - \mathbb{E}_\theta(T)$ (pour δ quelconque) et en écrivant une formule intégrale pour le membre de droite puis en appliquant l'inégalité de Cauchy-Schwarz. La borne ainsi trouvée est elle atteinte par $\tilde{\theta}$?

Exercice 23 (Cas délicat: loi de Cauchy). On considère un échantillon de la loi de Cauchy de densité

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

1) Est il aisé de calculer un estimateur du maximum de vraisemblance?
2) La méthode des moments est elle envisageable?
3) Quelle est la médiane de la loi de Cauchy considérée?
4) La médiane empirique est elle un estimateur convergent de θ?
5) Pour un $(2n+1)$–échantillon, calculer la densité de la médiane empirique M_n. Donner une forme générale non spécifique à l'expression de $f(x,\theta)$ donnée ci-dessus.
6) Dans le cas où $f(x,\theta) = g(x-\theta)$ avec g paire, comme dans le cas de la loi de Cauchy étudiée ici, montrer que M_n est un estimateur sans biais de θ.
7) Dans le cas de la loi de Cauchy étudiée ici, donner des expressions intégrales de la variance de M_n et de l'information de Fisher du modèle. Que faudrait il vérifier pour savoir si M_n est asymptotiquement efficace?

Chapitre 6

L'estimation par régions de confiance

6.1 Principe

Soit $(\Omega,\mathcal{F},\mathbb{P}_\theta, \theta \in \Theta)$ un modèle statistique et $g : \Theta \longrightarrow \mathbb{R}^d$. On cherche à estimer $g(\theta)$. On se donne $\alpha \in]0,1[$. On note X l'observation.

On appelle *région de confiance* au niveau $1-\alpha$ toute statistique R à valeurs dans l'ensemble des parties de $\mathbb{R}^d$ telle que, pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(R(X) \ni g(\theta)) \geq 1-\alpha.$$

L'idée est donc d'estimer $g(\theta)$ non pas en donnant une valeur qu'on espère proche de la vraie valeur mais en donnant une région qui contient la vraie valeur avec une probabilité assez grande.

La définition ci-dessus -ne serait-ce qu'à travers le terme statistique- pose le problème de la mesurabilité de R. Ce qu'on demande à l'application $R : \Omega \longrightarrow \mathcal{P}(\mathbb{R}^d)$ est que : $\forall y \in g(\Theta), \{R(X) \ni y\} \in \mathcal{F}$..

On remarquera que plus le niveau augmente plus la définition est exigeante et plus $R(X)$ risque d'être grande. Il s'agira donc dans la pratique d'arbitrer entre la valeur du niveau $1-\alpha$ qu'on souhaite élevé et la taille de $R(X)$ qui ne doit pas être trop grande sinon l'information apportée est nulle. Par exemple $\mathbb{R}^d$ tout entier est une région de confiance à tout niveau de toute fonction de θ mais cela n'a aucun intérêt . Ainsi, à un niveau donné, on pourra chercher $R(X)$ de façon que :

- $R(X)$ a une mesure de Lebesgue minimale
- autre critère : R soit sans biais au sens suivant :

$$\forall \theta \in \Theta, \forall h \neq g(\theta) \quad \mathbb{P}_\theta(R(X) \ni h) \leq 1-\alpha.$$

La détermination de régions de confiance se fait souvent à l'aide d'un *pivot*. On appelle pivot une application $\pi : \Omega \times g(\Theta) \longrightarrow \mathbb{R}^k$, mesurable,

telle que la loi de $\omega \mapsto \pi(\omega,g(\theta))$ ne dépende pas de $\theta \in \Theta$. Notons ν cette loi. Soit $\alpha \in]0,1[$ et $B \subset \mathbb{R}^k$ tel que $\nu(B) \geq 1-\alpha$. Alors :

$$R(X) = \{y \in g(\Theta); \quad \pi(X,y) \in B\}$$

est une région de confiance au niveau $1-\alpha$ car

$$\mathbb{P}_\theta(R(X) \ni g(\theta)) = \mathbb{P}_\theta(\pi(X,g(\theta)) \in B) = \nu(B) \geq 1-\alpha$$

Dans ce cadre on essaiera de prendre B de mesure de Lebesgue minimale ; il y a bien sûr beaucoup de régions de $\mathbb{R}^d$ qui vérifient $\nu(B) \geq 1-\alpha$.

Par exemple dans le cas $d=1$ on veut choisir $[a,b]$ tel que $\nu([a,b]) = 1-\alpha$, donc a et b tels que $\nu(]-\infty,a[) = \alpha_1$ et $\nu(]b,+\infty[) = \alpha_2$ avec $\alpha_1+\alpha_2 = \alpha$. Par exemple si ν admet une densité f continue unimodale (i.e. croissante sur $]-\infty,m[$, décroissante sur $]m,+\infty[$) on voit que $[a,b]$ est de longueur minimale pour $[a,b] = \{f \geq c\}$.

Toutefois le contexte peut amener à préférer d'autres intervalles de confiance notamment unilatéraux $]-\infty,b]$ (i.e. $a = -\infty$, $\alpha_2 = \alpha$) ou $[a,+\infty[$ $b = +\infty$, $\alpha_1 = \alpha$).

6.2 Exemples pour un échantillon gaussien

6.2.1 Intervalle de confiance pour la moyenne avec une variance connue.

Considérons donc un n-échantillon d'une loi $\mathcal{N}(m,\sigma^2)$ où m est inconnu. La variable $\sqrt{n}\frac{(\overline{X}_n - m)}{\sigma}$ est un pivot de loi $\nu = \mathcal{N}(0,1)$.

Un niveau $1-\alpha$ étant donné, on choisit γ tel que $\nu([-\gamma,\gamma]) = 1-\alpha$ et on peut affirmer alors que

$$\mathbb{P}_m\left(-\gamma \leq \sqrt{n}\frac{\overline{X}_n - m}{\sigma} \leq \gamma\right) = 1-\alpha$$

i.e.

$$\mathbb{P}_m\left(\overline{X}_n - \frac{\sigma\gamma}{\sqrt{n}} \leq m \leq \overline{X}_n + \frac{\sigma\gamma}{\sqrt{n}}\right) = 1-\alpha$$

6.2.2 Intervalle de confiance pour la moyenne avec une variance inconnue.

Considérons maintenant un n–échantillon d'une loi $\mathcal{N}(m,\sigma^2)$ où m et σ^2 sont inconnus. On sait que $\sqrt{n}\frac{(\overline{X}_n - m)}{S_n}$ est un pivot de même loi ν qu'une variable T_{n-1} de loi de Student à $n-1$ degrés de liberté. On rappelle que S_n désigne la racine de la variance empirique modifiée. La loi de Student est

la loi de $T_{n-1} = \sqrt{n-1}\frac{U}{\sqrt{V}}$ où $U \overset{(loi)}{=} \mathcal{N}(0,1)$ et $V \overset{(loi)}{=} \mathcal{X}^2(n-1)$. Le niveau $1-\alpha$ étant choisi, on détermine γ tel que

$$\nu[-\gamma,\gamma] = \mathbb{P}(-\gamma \leq T_{n-1} \leq \gamma) = 1-\alpha$$

et on a alors

$$\mathbb{P}_{(m,\sigma^2)}\left(\overline{X}_n - \frac{\gamma S_n}{\sqrt{n}} \leq m \leq \overline{X}_n + \frac{\gamma S_n}{\sqrt{n}}\right) = 1-\alpha$$

Comme, lorsque $n \to +\infty$, $T_{n-1} \overset{(loi)}{\longrightarrow} \mathcal{N}(0,1)$, la valeur γ trouvée ici sera proche si n est grand de celle obtenue au paragraphe précédent dans le cas de variance connue ; on a aussi $S_n \overset{p.s.}{\longrightarrow} \sigma$ et l'intervalle de confiance obtenu est en cohérence avec celui du paragraphe précédent.

6.2.3 Intervalle de confiance pour la variance avec une moyenne connue.

Soit $\Sigma_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2$ La variable $n\frac{\Sigma_n^2}{\sigma^2}$ a même loi que $\mathcal{X}^2(n)$ qui suit une loi du $\mathcal{X}^2$ à n degrés de liberté. Pour un niveau $1-\alpha$ donné, soit $\mathcal{X}_1$ et $\mathcal{X}_2$ tels que

$$\mathbb{P}(\mathcal{X}^2(n) \in [\mathcal{X}_1,\mathcal{X}_2]) = 1-\alpha$$

on a

$$\mathbb{P}\left(\frac{n\Sigma_n^2}{\mathcal{X}_2} \leq \sigma^2 \leq \frac{n\Sigma_n^2}{\mathcal{X}_1}\right) = 1-\alpha.$$

6.2.4 Intervalle de confiance pour la variance avec une moyenne inconnue

On utilise cette fois le pivot $(n-1)\frac{S_n^2}{\sigma^2}$ de loi $\mathcal{X}^2(n-1)$.

6.3 Utilisation d'un estimateur asymptotiquement normal

Dans un modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta)$ paramétré par le réel $\theta \in \Theta$, on suppose disposer d'une suite (T_n) d'estimateurs de θ asymptotiquement normale:

$$\frac{T_n - \theta}{\sigma_n(\theta)} \overset{(loi)}{\longrightarrow} \mathcal{N}(0,1).$$

Pour $G \overset{(loi)}{=} \mathcal{N}(0,1)$ déterminons γ tel que $\mathbb{P}(-\gamma \leq G \leq \gamma) = 1-\alpha$, ce niveau $1-\alpha$ étant fixé. Alors

$$\mathbb{P}_\theta(T_n - \gamma\sigma_n(\theta) \leq \theta \leq T_n + \gamma\sigma_n(\theta)) \overset{n\to+\infty}{\longrightarrow} 1-\alpha.$$

On dispose donc d'un intervalle de confiance asymptotique ; apparemment au moins car il faut voir que les bornes de cet intervalle font intervenir $\sigma_n(\theta)$ qui dépend de θ. Il y a donc un travail supplémentaire pour obtenir quelque chose d'utilisable.

Exemple : Intervalle de confiance pour le paramètre d'une loi de Bernoulli.

Considérons un n-échantillon d'une loi de Bernoulli de paramètre p. Alors par le théorème central limite :

$$\frac{\sqrt{n}}{\sqrt{p(1-p)}}(\overline{X}_n - p) \xrightarrow{(loi)} \mathcal{N}(0,1).$$

Pour un niveau $1-\alpha$ donné et γ comme ci-dessus :

$$\mathbb{P}_p\left(-\gamma \leq \frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \gamma\right) \longrightarrow 1-\alpha.$$

Or

$$\begin{aligned}
-\gamma \leq \frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \gamma \quad &\Longleftrightarrow \quad \frac{(\overline{X}_n - p)^2}{\frac{p(1-p)}{n}} \leq \gamma^2 \\
&\Longleftrightarrow \quad \overline{X}_n^2 - 2p\overline{X}_n + p^2 \leq \frac{\gamma^2}{n}p - \frac{\gamma^2}{n}p^2 \\
&\Longleftrightarrow \quad (1+\frac{\gamma^2}{n})p^2 - 2(\overline{X}_n + \frac{\gamma^2}{n})p + \overline{X}_n^2 \leq 0 \\
&\Longleftrightarrow \quad p \in \left[p_1(\overline{X}_n, \frac{\gamma^2}{n}), p_2(\overline{X}_n, \frac{\gamma^2}{n})\right]
\end{aligned}$$

où p_1,p_2 sont les deux racines du binôme apparaissant ci-dessus. On dispose donc d'un intervalle de confiance asymptotique :

$$\mathbb{P}_p\left(p_1(\overline{X}_n, \frac{\gamma^2}{n}) \leq p \leq p_2(\overline{X}_n, \frac{\gamma^2}{n})\right) \longrightarrow 1-\alpha.$$

On peut être encore plus direct en approximant $p(1-p)$ par $\overline{X}_n(1-\overline{X}_n)$ puisque l'on sait que $\overline{X}_n \xrightarrow{p.s.} p$ et on écrira :

$$\mathbb{P}_p\left(\overline{X}_n - \gamma\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq p \leq \overline{X}_n + \gamma\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right) \longrightarrow 1-\alpha$$

Application numérique : On réalise un sondage pré-électoral sur 1000 personnes:

candidat Nicolas 510

candidat Ségolène 490

Notons p la proportion vraie pour Nicolas. Quel est un intervalle de confiance au niveau 95% de p?

On trouve $0{,}48 \leq p \leq 0{,}54$.

6.4 Exercices

Exercice 24. Soit un échantillon $X_1, \ldots, X_n$ de la loi uniforme sur $[0,\theta]$ où $\theta > 0$ est inconnu.
1) Notons $X_{(n)} = \max_{1 \leq i \leq n} X_i$ comme d'habitude. Montrer que $\frac{\theta}{X_{(n)}}$ est un pivot de densité unimodale.
2) Donner un intervalle de confiance pour θ au niveau $1 - \alpha$.

Exercice 25. Soit un échantillon $X_1, \ldots, X_n$ de la loi exponentielle de moyenne $1/\theta > 0$ inconnue.
1) Quelle est la loi de $2\,\theta \sum_{i=1}^{n} X_i$?
2) Déterminer un intervalle de confiance au niveau $1 - \alpha$ de θ.

Chapitre 7

Les tests : principe

On considère un modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta \in \Theta)$. On suppose que l'espace des paramètres Θ est partitionné en 2 parties disjointes :
$\Theta = \Theta_0 \cup \Theta_1$.

Le problème est de décider au vu de l'échantillon X $(= id_\Omega$ comme d'habitude) s'il est raisonnable d'affirmer que $\theta \in \Theta_0$ qu'on appelle l'hypothèse nulle H_0 ou bien que $\theta \in \Theta_1$ qu'on appelle l'hypothèse alternative H_1.

Une idée naturelle est de prendre cette décision en fonction de l'appartenance de X à une certaine région W de Ω qu'on appelle région critique : on rejettera l'hypothèse nulle H_0 (et donc on admettra H_1) si $X \in W$.

On peut généraliser un peu cette idée en introduisant la notion de fonction de test qui désigna simplement une fonction Φ de Ω dans $[0,1]$ et en adoptant pour règle de décision associée à Φ :

- on rejette H_0 si $\Phi(\omega) = 1$ (et donc on accepte H_1),
- on accepte H_0 si $\Phi(\omega) = 0$ (et donc on rejette H_1),
- si $\Phi(\omega) \in]0,1[$: on hésite ! On fait un tirage au sort qui conduit à rejeter H_0 avec probabilité $\Phi(\omega)$.

Dans beaucoup de cas usuels on pratique un test non aléatoire c'est à dire que $\Phi(\omega) \in \{0,1\}$, il n'y a pas d'hésitation, on accepte ou bien on rejette H_0.

En prenant la décision associée à un test on peut commettre deux types d'erreurs : l'erreur de première espèce qui consiste à rejeter l'hypothèse nulle H_0 alors qu'elle est vraie et l'erreur de deuxième espèce qui consiste à accepter l'hypothèse nulle H_0 alors qu'elle est fausse (i.e. l'hypothèse alternative H_1 est vraie).

On quantifie ces erreurs en appelant :

- fonction de risque de première espèce : $\begin{array}{l}\Theta_0 \longrightarrow [0,1] \\ \theta \mapsto \mathbb{E}_\theta[\Phi(X)]\end{array}$
- fonction de risque de deuxième espèce : $\begin{array}{l}\Theta_1 \longrightarrow [0,1] \\ \theta \mapsto \mathbb{E}_\theta[1-\Phi(X)] = 1 - \mathbb{E}_\theta[\Phi(X)]\end{array}$

Plaçons nous par exemple dans le cas d'un test d'hypothèses simples ce qio signifie que Θ_0 et Θ_1 sont des singletons : $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \overline{\{\theta_1\}}$ alors la connaissance du risque se limite à deux réels : l'erreur de première espèce $\mathbb{P}_{\theta_0}$ (rejeter H_0) que certains livres notent $\mathbb{P}_{H_0}$ (rejeter H_0) et l'erreur de deuxième espèce $\mathbb{P}_{\theta_1}$ (accepter H_0) $= \mathbb{P}_{H_1}$ (accepter H_0)

7.1 Optimisation d'un test

Il est bien clair qu'un test $\tilde{\Phi}$ est préférable à un test Φ s'il y a un plus faible risque de première espèce, c'est à dire quand $\theta \in \Theta_0$, et un plus fai ble risque de deuxième espèce, c'est à dire quand $\theta \in \Theta_1$.

Toutefois cela fait deux minimisations à gérer et il n'y a aucune raison qu'un test qui minimise l'erreur de première espèce, minimise aussi celle de deuxième espèce.

En conséquence Neyman et Pearson ont proposé en 1933 de traiter les deux risques de façon non symétrique. On appelle niveau d'un test Φ l'erreur maximale de première espèce.

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\Phi)$$

Définition : un test Φ^* est UPP, uniformément plus puissant , au niveau $\alpha(\in]0,1[)$ si :

- il est de niveau α : $\sup_{\theta\in\Theta_0} \mathbb{E}_\theta(\Phi^*) = \alpha$,
- il minimise l'erreur de deuxième espèce : pour tout test Φ de niveau $\leq \alpha$ et tout $\theta \in \Theta_1$,$\mathbb{E}_\theta(\Phi) \leq \mathbb{E}_\theta(\Phi^*)$.

En anglais on dit UMP "uniformly most powerful". Le terme puissance désigne l'erreur de deuxi'ème espèce c'est à dire $\mathbb{E}_\theta(\Phi)$.

Ainsi un test UPP est un test pour lequel le risque de rejeter à tort H_0 est controllé par α et qui est le plus puissant pour rejeter H_0 quand elle n'est pas vrai.

Peut-on trouver facilement des test UPP? Nous allons étudier dans ce qui suit des situations où l'on dispose d'une méthode générale de construction.

7.2 Le théorème de Neyman-Pearson pour le test d'hypothèses simples

On considère ici un modèle statistique $(\Omega,\mathcal{F},\mathbb{P}_\theta;\theta \in \{\theta_0,\theta_1\})$ et les notations précédentes se réduisent à $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$. On note $L(x,\theta)$ la vraisemblance par rapport à la mesure dominante $\mathbb{P}_{\theta_0} + \mathbb{P}_{\theta_1}$.

Théorème 24. *Pour tout $\alpha \in]0,1[$ il existe un test Φ^* UPP de niveau α de la forme :*

$$\Phi^*(X) = \begin{cases} 1 \text{ si } L(X,\theta_1) > kL(X,\theta_0) \\ \gamma \text{ si } L(X,\theta_1) = kL(X,\theta_0) \\ 0 \text{ si } L(X,\theta_1) < kL(X,\theta_0) \end{cases}$$

avec $k > 0$ et $\gamma \in [0,1]$. De plus tout test UPP de niveau α est p.s. de la forme ci-dessus (avec éventuellement γ aléatoire).

Remarquons que souvent l'évènement $L(X,\theta_1) = kL(X,\theta_0)$ est de probabilité nulle et on pourra construire un test non aléatoire.

Exemple : test de l'espérance d'une loi normale

Considérons un n-échantillon d'une loi normale de variance σ^2 connue et d'espérance inconnue θ pouvant être soit θ_0 soit θ_1 $(\theta_0 < \theta_1)$.

La vraisemblance d'un tel échantillon $X = (X_1,...,X_n)$ s'écrit :

$$\begin{aligned} L(X,\theta) &= (2\pi)^{-\frac{n}{2}}\sigma^{-n}exp - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2 \\ &= (2\pi)^{-\frac{n}{2}}\sigma^{-n}exp - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}X_i^2 - 2\theta\sum_{i=1}^{n}X_i + n\theta^2\right) \end{aligned}$$

Alors $L(X,\theta_1) > kL(X,\theta_0)$ équivaut à $\overline{X}_n > K$ avec :

$$K = \frac{\sigma^2}{n}\frac{\ln k}{\theta_1-\theta_0} + \frac{\theta_1+\theta_0}{2} \text{ et comme d'habitude } \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i.$$

Il reste à déterminer K (et donc k) de façon que la probabilité de la région critique assure un test de niveau α qu'on se sera donné.

On note que sous $\mathbb{P}_\theta$, $\overline{X}_n$ suit la loi $\mathcal{N}(\theta,\frac{\sigma^2}{n})$. Donc en notant F la fonction de répartition de la loi normale $\mathcal{N}(0,1)$, on a :

$$\begin{aligned} \mathbb{P}_{\theta_0}(\overline{X}_n > K) &= 1 - \mathbb{P}_0\left(\frac{\sqrt{n}}{\sigma}(\overline{X}_n-\theta_0) \leq \frac{\sqrt{n}}{\sigma}(K-\theta_0)\right) \\ &= 1 - F\left(\frac{\sqrt{n}}{\sigma}(K-\theta_0)\right) \end{aligned}$$

Ainsi le test sera de niveau α en choisissant :

$$K = \theta_0 + \frac{\sigma}{\sqrt{n}}F^{-1}(1-\alpha).$$

Quelle est alors la puissance β de ce test que l'on sait UPP par le théorème de Neyman-Pearson ? Elle vaut :

$$\begin{aligned} \beta = \mathbb{P}_{\theta_1}(\overline{X}_n > K) &= \mathbb{P}_{\theta_1}\left(\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \theta_1) > \frac{\sqrt{n}}{\sigma}(K - \theta_1)\right) \\ &= 1 - F\left(\frac{\sqrt{n}}{\sigma}(K - \theta_1)\right) \end{aligned}$$

ou de façon complète : $\beta = 1 - F\left(\frac{\sqrt{n}}{\sigma}(\theta_0 - \theta_1) + F^{-1}(1-\alpha)\right)$.