CHAPITRE 4

INTRODUCTION À LA STATISTIQUE

4.1. Modèle statistique et problèmes statistiques

Dans la théorie des probabilités, on dispose un espace de probabilité. En statistique, on ne connaît pas la vraie probabilité $\mathbb{P}$ derière une expérience aléatoire. On suppose cependant que la mesure de probabilité appartient à une famille connue. Le but est donc de tirer des informations les plus précises possible, de la mesure de probabilité $\mathbb{P}$, via les résultats de l'expérience.

On considère un espace mesurable $(\Omega, \mathcal{F})$ muni d'une famille $\mathcal{P}$ de mesures de probabilité sur $(\Omega, \mathcal{F})$. Le triplet $(\Omega, \mathcal{F}, \mathcal{P})$ est appelé un *modèle statistique*. S'il existe une mesure σ-finie μ sur $(\Omega, \mathcal{F})$ telle que toute mesure de probabilité $\mathbb{P} \in \mathcal{P}$ est absolument continue par rapport à μ, on dit que le modèle statistique $(\Omega, \mathcal{F}, \mathcal{P})$ est *dominé* par μ. Dans ce cours, on considère surtout des *modèles paramétriques*, où la famille $\mathcal{P}$ est paramétrée par un sous-ensemble Θ de $\mathbb{R}^d$ ($d \in \mathbb{N}$, $d \geqslant 1$).

Exemple 4.1.1. — On lance n fois un dé à six faces qui est éventuellement piégé. L'espace Ω de cette expérience est donc $\{1, \ldots, 6\}^n$. On le munit de la tribu de tous les événements possibles. On suppose que les n lances sont indépendantes. Ainsi les mesures de probabilité possibles pour cette expérience sont de la forme $\mu^{\otimes n}$, où μ est une mesure de probabilité sur $\{1, \ldots, 6\}$. Cette famille peut donc être paramétrée par

$$\Theta = \{(a_1, \ldots, a_6) \in [0,1]^6 \,|\, a_1 + \cdots + a_6 = 1\}.$$

Bien que les problèmes statistiques sont très variés, on peut les classifiés en trois catégories.

(1) **Estimation de paramètre.** Étant donnée une fonction borélienne $g(.)$ sur Θ, estimer la valeur de la fonction évaluée en paramètre correspondant à la vrai mesure de probabilité.

(2) **Test d'hypothèse.** On décompose l'espace de paramètres Θ en réunion disjointe de deux sous-ensembles Θ_0 et Θ_1. Déterminer si la vraie valeur du paramètre se trouve dans Θ_0 ou plutôt dans Θ_1.

(3) **Région de confiance.** On se donne aussi une fonction borélienne $g(.)$. Le but est de proposer un sous-ensemble de $g(\Theta)$ où la valeur de g évaluée en vraie paramètre est censée appartienir.

4.2. Statistique

Soient $(\Omega, \mathcal{F}, \mathcal{P})$ un modèle statistique et $(E, \mathcal{E})$ un espace mesurable. On appelle *statistique* à valeurs dans $(E, \mathcal{E})$ toute application mesurable de $(\Omega, \mathcal{F})$ vers $(E, \mathcal{E})$. On dit qu'une statistique $S : \Omega \to E$ est *exhaustive* si les lois conditionnelles de $\mathbb{P}$ sachant S sont identiques. Dans la suite, on désigne par $Y : (\Omega, \mathcal{F}) \to (\Omega, \mathcal{F})$ l'application d'identité, appelée la *statistique des observation.* C'est une statistique exhaustive car la loi conditionnelle de n'import quelle mesure de probabitilié sur $(\Omega, \mathcal{F})$ sachant Y est la mesure de Dirac δ_Y.

Exemple 4.2.1. — Considérons le modèle statistique d'échantillonnnage des lois de Bernoulli $\{0,1\}^n$ paramétré par $\Theta = [0,1]$. On désigne par $Y = (Y_1, \ldots, Y_n)$ la statistique des observations. Alors la statistique de somme $S = Y_1 + \cdots + Y_n$ est exhaustive. En effet, pour tout $\theta \in [0,1]$ on a

$$\mathbb{P}_\theta(Y = y \,|\, S = s) = \frac{\mathbb{P}(Y = y, S = s)}{\mathbb{P}(S = s)}.$$

Si $y_1 + \cdots + y_n \neq s$, on a $\mathbb{P}_\theta(Y = y \,|\, S = s) = 0$; sinon on a

$$\mathbb{P}_\theta(Y = y \,|\, S = s) = \frac{\theta^s (1-\theta)^{n-s}}{\binom{n}{s}\theta^s(1-\theta)^{n-s}} = \binom{n}{s}^{-1}.$$

Le calcul des lois conditionnelles pourrait être assez compliqué. Le théorème suivant donne un critère pour les statistiques exhaustive.

Théorème 4.2.2. — *Soit $(\Omega, \mathcal{F}, \mathcal{P})$ un modèle statistique paramétré par Θ, qui est dominé par une mesure σ-finie μ. Pour qu'une statstique S à valeurs dans un espace mesurable $(E, \mathcal{E})$ soit exhaustive, il faut et il suffit que, pour tout $\theta \in \Theta$, la densité $p_\theta(.)$ de $\mathbb{P}_\theta$ par rapport à μ se décompose sous la forme $\psi_\theta(S(\cdot))f(\cdot)$, où f est une fonction positive $\mathcal{F}$-mesurable sur Ω et ψ_θ est une fonction positive $\mathcal{E}$-mesurable sur E.*

Démonstration. — Montrons seulement la partie de suffisance. Soit $A \in \sigma(S)$ et X une variable aléatoire positive et $\mathcal{F}$-mesurable sur Ω. Quitte à remplacer μ par une mesure de probabilité équivalente, on peut supposer que μ est elle-même une mesure de probabilité. On a

$$\begin{aligned}\mathbb{E}^{\mathbb{P}_\theta}[\mathbb{1}_A X] &= \mathbb{E}^{\mu}[\mathbb{1}_A X p_\theta] = \mathbb{E}^{\mu}[\mathbb{1}_A X \psi_\theta(S) f] = \mathbb{E}^{\mu}[\mathbb{1}_A \psi_\theta(S) \mathbb{E}^{\mu}[fX|S]] \\ &= \mathbb{E}^{\mu}\Big[\mathbb{1}_A \psi_\theta(S) f \frac{\mathbb{E}^{\mu}[fX|S]}{\mathbb{E}^{\mu}[f|S]}\Big] = \mathbb{E}^{\mathbb{P}_\theta}\Big[\mathbb{1}_A \frac{\mathbb{E}^{\mu}[fX|S]}{\mathbb{E}^{\mu}[f|S]}\Big].\end{aligned}$$

Cela montre que

$$\mathbb{E}^{\mathbb{P}_\theta}[X|S] = \frac{\mathbb{E}^{\mu}[fX|S]}{\mathbb{E}^{\mu}[f|S]}$$

ne dépend pas de θ. □

4.3. Information de Fisher

On considère un modèle statistique paramétrique $(\Omega, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, où Θ est un sous-ensemble ouvert de $\mathbb{R}^d$. On suppose que toutes les mesures de probabilité $\mathbb{P}_\theta$ sont équivalentes à une mesure σ-finie μ, et on désigne par $p(\theta)$ la densité de Radon-Nikodym $\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mu$. C'est une variable aléatoire strictement positive sur l'espace mesurable $(\Omega, \mathcal{F})$. On considère p comme une fonction définie sur $\Theta \times \Omega$. On suppose que

(1) la fonction p est deux fois différentiable par rapport à θ,

(2) pour tout $\theta \in \Theta$, la matrice de covariance du vecteur $\partial_\theta \ln p$ par rapport à la mesure de probabilité $\mathbb{P}_\theta$ existe.

(3) il existe une fonction $\mathcal{F}$-mesurable $\varphi : \Omega \to \mathbb{R}$ intégrable par rapport à μ telle que

$$\forall\, \omega \in \Omega, \quad \sup_{\theta\in\Theta} \max(\|\partial_\theta p(\theta,\omega)\|, \|\partial_\theta^2 p(\theta,\omega)\|) \leqslant \varphi(\omega).$$

On appelle *information de Fisher* du modèle statistique la matrice de covariance du vecteur $\partial_\theta \ln p$ par rapport à $\mathbb{P}_\theta$, notée comme $I(\theta)$.

Lemme 4.3.1. — *Pour tout $\theta \in \Theta$, on a*

$$\mathbb{E}^{\mathbb{P}_\theta}[\partial_\theta \ln p(\theta)] = 0.$$

Démonstration. — Comme $p(\theta)$ est la densité d'une mesure de probabilité par rapport à μ, on a

$$\forall\, \theta \in \Theta, \quad \int_\Omega p(\theta)\, \mathrm{d}\mu = 1,$$

d'où (la condition (3) comme ci-dessus nous permet de dériver sous signe somme)

$$\forall\, \theta \in \Theta, \quad \int_\Omega \partial_\theta p(\theta)\,\mathrm{d}\mu = \mathbb{E}^{\mathbb{P}_\theta}\Big[\frac{\partial_\theta p(\theta)}{p(\theta)}\Big] = \mathbb{E}^{\mathbb{P}_\theta}[\partial_\theta \ln p(\theta)] = 0.$$

□

Théorème 4.3.2. — *La matrice d'information de Fisher est égale à*

$$-\mathbb{E}^{\mathbb{P}_\theta}[\partial_\theta^2 \ln p(\theta)].$$

Démonstration. — On a vu que

$$\int_\Omega \partial_\theta p(\theta)\,\mathrm{d}\mu = 0.$$

On dérive encore une fois cette formule pour obtenir

$$\int_\Omega \partial_\theta^2 p(\theta)\,\mathrm{d}\mu = \mathbb{E}^{\mathbb{P}_\theta}[p(\theta)^{-1}\partial_\theta^2 p(\theta)] = 0.$$

En outre, on a

$$\partial_\theta^2 \ln p(\theta) = \partial_\theta(p(\theta)^{-1}\partial_\theta p(\theta)) = -\frac{\partial_\theta p(\theta)^{\otimes 2}}{p(\theta)^2} + \frac{\partial_\theta^2 p(\theta)}{p(\theta)}.$$

D'où

$$\mathbb{E}^{\mathbb{P}_\theta}[\partial_\theta^2 \ln p(\theta)] = -\mathbb{E}^{\mathbb{P}_\theta}\Big[\frac{\partial_\theta p(\theta)^{\otimes 2}}{p(\theta)^2}\Big] = \mathrm{cov}^{\mathbb{P}_\theta}(\partial_\theta \ln p(\theta)),$$

où la dernière égalité provient du lemme 4.3.1. Le résultat est donc démontré.

□

4.4. Estimateur sans biais

On fixe dans ce paragraphe un modèle statistique $(\Omega, \mathcal{F}, \mathcal{P})$ paramétré par $\Theta \subset \mathbb{R}^k$. Étant donnée une fonction borélienne g sur Θ à valeurs dans $\mathbb{R}^m$, on appelle *estimateur* de g toute statistique sur le modèle à valeurs dans $\mathbb{R}^m$. On dit qu'un statistique $\widehat{g} : \Omega \to \mathbb{R}^m$ est un *estimateur sans biais* si $\mathbb{E}^{\mathbb{P}_\theta}[\widehat{g}] = g(\theta)$ quel que soit $\theta \in \Theta$.

Pour tout estimateur S, on définit $R_M(S)$ comme l'application de Θ vers l'ensemble des matrices symétriques qui envoie $\theta \in \Theta$ en

$$\mathbb{E}^{\mathbb{P}_\theta}[(\widehat{g} - g(\theta))(\widehat{g} - g(\theta))^\intercal],$$

appelée la matrice de risque quadratique. Si $\widehat{g}_1$ et $\widehat{g}_2$ sont deux estimateurs de g tels que $R_M(\widehat{g}_2) - R_M(\widehat{g}_1)$ soit une matrice semi-positive, on dit que $\widehat{g}_1$ est *quadratiquement préférable* à $\widehat{g}_2$. On voit aussitôt que, si $\widehat{g}$ est un estimateur

sans biais, alors $R_M(\widehat{g})$ n'est rien d'autre que la matrice de variance-covariance de $\widehat{g}$.

***Théorème 4.4.1* (Rao-Blackwell).** — *Si $\widehat{g}$ est un estimateur sans biais et si S est une statistique exhaustive, alors l'estimateur $\widehat{g}_S = \mathbb{E}[\widehat{g}|S]$ est sans biais et quadratiquement préférable à $\widehat{g}$.*

Démonstration. — Bien que l'on n'a pas précisé dans l'énoncé du théorème la mesure de probabilité que l'on utilise à calculer l'espérance conditionnelle, le fait que S est une statistique exhaustive montre que cette espérance conditionnelle ne dépend pas du choix d'une mesure de probabilité dans $\mathcal{P}$. En outre, comme $\mathbb{E}^{\mathbb{P}_\theta}[\mathbb{E}^{\mathbb{P}_\theta}[\widehat{g}|S]] = \mathbb{E}^{\mathbb{P}_\theta}[\widehat{g}] = g(\theta)$, on obtient que la statistique $\widehat{g}_S$ est sans biais. En outre, pour tout $x \in \mathbb{R}^m$, on a

$$\mathbb{E}^{\mathbb{P}_\theta}[(x^\tau(\widehat{g} - g(\theta)))^2] = \mathbb{E}^{\mathbb{P}_\theta}[\mathbb{E}^{\mathbb{P}_\theta}[(x^\tau(\widehat{g} - g(\theta)))^2|S]] \geqslant \mathbb{E}^{\mathbb{P}_\theta}[(x^\tau(\widehat{g}_S - g(\theta)))^2],$$

d'après l'inégalité de Jensen. Le théorème est donc démontré. □

Dans le cas où l'information de Fisher est bien définie, le théorème suivant du à Cramer et Rao donne une minoration pour la matrice de variance-covariance d'un estimateur sans biais. Dans la suite, on suppose que le modèle statistique est dominé par une mesure σ-finie μ et on désigne par $p(\theta, .)$ la densité de Radon-Nikodym de $\mathbb{P}_\theta$ par rapport à μ. On suppose en outre que Θ est un ouvert de $\mathbb{R}^k$ et que la fonction g est différentiable.

***Théorème 4.4.2* (Cramer-Rao).** — *On suppose que l'information de Fisher pour le modèle statistique est bien définie. Soit $\widehat{g}$ un estimateur sans biais tel que $\mathbb{E}^{\mathbb{P}_\theta}[\|\widehat{g}\|^2] < +\infty$ pour tout $\theta \in \Theta$ et que*

$$\int_\Omega p(\theta)\widehat{g}\,\mathrm{d}\mu$$

est différentiable en θ sous signe intégrale. Alors

$$\mathrm{cov}^{\mathbb{P}_\theta}(\widehat{g}) - Dg(\theta)^\tau I(\theta)^{-1} Dg(\theta)$$

est une matrice symétrique semi-positive.

Démonstration. — Pour tout $\theta \in \Theta$, on a

$$\int_\Omega p(\theta)\widehat{g}\,\mathrm{d}\mu = g(\theta).$$

On en déduit

$$\int_\Omega \partial_\theta p(\theta)^\tau \widehat{g}\,\mathrm{d}\mu = Dg(\theta),$$

ou encore

$$\mathbb{E}^{\mathbb{P}_\theta}[\partial_\theta(\ln p(\theta))^\tau \widehat{g}] = Dg(\theta)$$

Considérons le vecteur de variables aléatoires

$$X = (\widehat{g} - g(\theta)) - Dg(\theta)^\tau I(\theta)^{-1}\partial_\theta(\ln p(\theta)).$$

Il est d'espérance nulle. Sa matrice de variance-covariance est

$$\begin{aligned}
&\mathbb{E}^{\mathbb{P}_\theta}[(\widehat{g} - g(\theta))(\widehat{g} - g(\theta))^\tau] - \mathbb{E}^{\mathbb{P}_\theta}[(\widehat{g} - g(\theta))(\partial_\theta \ln p(\theta))^\tau (I(\theta)^{-1})^\tau Dg(\theta)] \\
&- \mathbb{E}^{\mathbb{P}_\theta}[Dg(\theta)^\tau I(\theta)^{-1}\partial_\theta(\ln p(\theta))(\widehat{g} - g(\theta))^\tau] \\
&+ \mathbb{E}^{\mathbb{P}_\theta}[Dg(\theta)^\tau I(\theta)^{-1}\partial_\theta(\ln p(\theta))(\partial_\theta \ln p(\theta))^\tau (I(\theta)^{-1})^\tau Dg(\theta)],
\end{aligned}$$

qui est égale à

$$\mathrm{cov}^{\mathbb{P}_\theta}(\widehat{g}) - Dg(\theta)^\tau I(\theta)^{-1} Dg(\theta).$$

Le théorème est donc démontré. □

4.5. Le maximum de vraisemblance

On considère un modèle statistique de la forme $((E, \mathcal{E})^{\otimes n}, (\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n})_{\theta\in\Theta}))$, où $(\mathbb{Q}_\theta)_{\theta\in\Theta}$ est une famille de mesures de probabilité sur $(E, \mathcal{E})$, qui sont absolument continues par rapport à une mesure σ-finie ν. Pour tout $\theta \in \Theta$, soit $f(\theta, \cdot)$ la densité de $\mathbb{Q}_\theta$ par rapport à ν. Alors la densité $p(\theta, .)$ de $\mathbb{P}_\theta$ par rapport à $\mu = \nu^{\otimes n}$ est donnée par la relation

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

Le principe de maximum de vraisemblance consiste à maximiser la fonction de vraisemblance

$$(\theta \in \Theta) \longmapsto \ell(\theta, x_1, \ldots, x_n) = \frac{1}{n} \ln p(\theta, x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n \ln f(\theta, x_i).$$

On suppose l'existence d'une application mesurable $\theta^* : E^n \to \Theta$ telle que

$$\ell(\theta^*(\boldsymbol{x}), \boldsymbol{x}) = \sup_{\theta\in\Theta} \ell(\theta, \boldsymbol{x}).$$

On désigne par $\widehat{\theta}$ l'estimateur $\theta^*(Y)$, où $Y = (Y_1, \ldots, Y_n)$ est la statistique d'observation.

Exemple 4.5.1. — On considère le cas où $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et $\mathbb{Q}_{(\mu,\sigma)} = N(\mu, \sigma^2)$. Dans ce cas-là la fonction de vraisemblance est donnée par

$$\ell(\mu, \sigma, \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \Big(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \Big).$$

En particulier, on a

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma, \boldsymbol{x}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}.$$

Donc

$$\mu^*(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n).$$

L'estimateur du maximum de vraisemblance de μ est donc

$$\widehat{\mu} = \frac{1}{n}(Y_1 + \cdots + Y_n),$$

qui est un estimateur sans biais. De façon similaire, on a

$$\frac{\partial}{\partial \sigma} \ell(\mu, \sigma, \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \Big(-\frac{1}{\sigma} + \frac{1}{\sigma^3}(x_i - \mu)^2 \Big)$$

D'où

$$\sigma^*(\boldsymbol{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu^*(\boldsymbol{x}))^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j$$

En particulier, on a

$$\mathbb{E}[\widehat{\sigma}^2] = \frac{n-1}{n} \sigma^2.$$

Cela montre que l'estimateur $\widehat{\sigma}^2$ de σ^2 n'est pas sans biais.

Exemple 4.5.2. — On considère le cas où $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et $\mathbb{Q}_\sigma = N(0, \sigma^2)$. Dans ce cas-là la fonction de vraisemblance est donnée par

$$\ell(\sigma, \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \Big(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} x_i^2 \Big).$$

On obtient

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} Y_i^2,$$

qui est un estimateur san biais. La variance de cet estimateur est alors

$$\mathrm{var}(\widehat{\sigma}^2) = \frac{1}{n^2}(2\sigma^4 n) = \frac{2\sigma^4}{n}$$

Rappelons que l'information de Fisher est donnée par

$$I(\sigma) = \mathbb{E}^{\mathbb{P}_\sigma}[\partial_\sigma \ln p(\sigma, Y)^2] = \sum_{i=1}^{n} \mathbb{E}^{\mathbb{Q}_\sigma}\Big[\Big(-\frac{1}{\sigma} + \frac{1}{\sigma^3} Y_i^2\Big)^2\Big] = \frac{2n}{\sigma^2}.$$

On voit que la variance de $\widehat{\sigma}$ atteint la borne prédite par l'inégalité de Cramér-Lao. Cependant, si on considère l'estimateur

$$T = \frac{1}{n+2} \sum_{i=1}^{n} Y_i^2,$$

dont l'espérance est $\frac{n\sigma^2}{n+2}$. On a

$$\mathrm{var}(T) = \frac{2n\sigma^4}{(n+2)^2}.$$

Donc on a

$$R_M(T) = \frac{2\sigma^4}{n+2},$$

qui est plus petit que $R_M(\widehat{\sigma}^2)$. On en déduit que T est quadratiquement préférable à $\widehat{\sigma}^2$.