

Maîtrises Maths-MASS - Statistiques générales

Corrigé du partiel du 26 novembre 2002

Ex 1.

1. Les variables X_i étant à valeurs dans \mathbb{N} , on choisit comme espace des observations du modèle statistique associé à (X_1, \dots, X_n) l'ensemble $\Omega = \mathbb{N}^n$, muni de la tribu de l'ensemble de ses parties $\mathcal{P}(\Omega)$. La famille de lois $\{\mathbb{P}_\theta, \theta \in \Theta = \mathbb{R}_+^*\}$ définies sur Ω est dominée par la mesure de comptage $N_{|\Omega}$ sur Ω , et est caractérisée par

$$\begin{aligned} \forall x \in \Omega, \quad \frac{d\mathbb{P}_\theta}{dN_{|\Omega}}(x) &= \mathbb{P}_\theta((X_1, \dots, X_n) = (x_1, \dots, x_n)) \\ &= \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) \quad \text{car les } X_i \text{ sont indépendants entre eux} \\ &= \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\theta} = \exp(Q(\theta)\bar{x}_n - \varphi(\theta)) h(x) \end{aligned}$$

avec

$$Q(\theta) = n \ln(\theta), \quad \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad \varphi(\theta) = n\theta \quad \text{et} \quad h(x) = \frac{1}{\prod_{i=1}^n x_i!}$$

On reconnaît donc un modèle exponentiel dont \bar{X}_n est une statistique exhaustive. Comme l'ensemble $Q(\Theta) = \mathbb{R}$ contient un ouvert de \mathbb{R} , on en déduit en outre que la statistique \bar{X}_n est complète et minimale.

2. Comme Q est une fonction dérivable, de dérivée $Q'(\theta) = n/\theta$ continue et ne s'annulant pas sur $\Theta = \mathbb{R}_+^*$ qui est un ouvert de \mathbb{R} , on en déduit que le modèle est régulier, et que \bar{X}_n est une statistique régulière, et un estimateur UVMB et efficace de son espérance

$$g(\theta) = \mathbb{E}_\theta[\bar{X}_n] = \mathbb{E}_\theta[X_1] = \theta$$

De plus, comme \bar{X}_n est un estimateur efficace, il y a égalité dans l'inégalité de Cramer-Rao, et on en déduit le calcul de l'information de Fisher $I_n(\theta)$ du modèle :

$$I_n(\theta) = \frac{(g'(\theta))^2}{\text{var}_\theta(\bar{X}_n)} = \frac{1}{\frac{1}{n} \text{var}_\theta(X_1)} = \frac{n}{\theta}$$

où l'on s'est servi du fait que $\text{var}_\theta(X_1) = \theta$.

3. On a

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

avec $q(u, v) = v - u^2$. La méthode de substitution consiste à remplacer les espérances par les espérances empiriques ; on en déduit l'estimateur (fortement consistant) de θ :

$$\hat{\theta}(X) = q\left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2$$

On reconnaît la variance empirique non corrigée.

4. Comme $\mathbb{E}_\theta [|X_1|] = \mathbb{E}_\theta [X_1] = \theta < +\infty$, on déduit de la loi forte des grands nombres que

$$\text{sous } \mathbb{P}_\theta, \quad \lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n X_i = \mathbb{E}_\theta [X_1] = \theta$$

Puisque \bar{X}_n tend \mathbb{P}_θ -p.s. vers θ et $1/n$ vers 0, quand n tend vers $+\infty$, on en déduit que $T = (\bar{X}_n + 1/n)/(1 + 1/n)$ tend aussi \mathbb{P}_θ -p.s. vers θ . Autrement dit, T est fortement consistant.

Calcul des fonctions de risque :

$$\begin{aligned} R(\theta, \bar{X}_n) &= \text{var}_\theta (\bar{X}_n) = \frac{\theta}{n} \\ R(\theta, T) &= \text{var}_\theta (T) + (\mathbb{E}_\theta [T] - \theta)^2 \\ &= \frac{\text{var}_\theta (\bar{X}_n)}{\left(1 + \frac{1}{n}\right)^2} + \left(\frac{\theta + \frac{1}{n}}{1 + \frac{1}{n}} - \theta\right)^2 \\ &= \frac{\theta}{n \left(1 + \frac{1}{n}\right)^2} + \left(\frac{1 - \theta}{1 + n}\right)^2 \end{aligned}$$

Comparaison des fonctions de risque :

$$\lim_{\theta \rightarrow 0^+} R(\theta, \bar{X}_n) = 0 \quad \text{et} \quad \lim_{\theta \rightarrow 0^+} R(\theta, T) = \frac{1}{(1+n)^2}$$

Dans un voisinage de 0^+ , le risque de \bar{X}_n est donc plus faible que celui de T .

$$R(1, \bar{X}_n) = \frac{1}{n} \quad \text{et} \quad R(1, T) = \frac{1}{n \left(1 + \frac{1}{n}\right)^2}$$

En 1, le risque de \bar{X}_n est donc plus élevé que celui de T . On en conclut que les fonctions de risque de \bar{X}_n et de T ne sont pas comparables.

5. Le risque maximum de \bar{X}_n sur $]0, 1]$ est

$$\sup_{\theta \in]0, 1]} R(\theta, \bar{X}_n) = \sup_{\theta \in]0, 1]} \left(\frac{\theta}{n}\right) = \frac{1}{n}$$

Pour calculer la valeur maximale prise par $R(\theta, T)$, remarquons que cette fonction est une parabole convexe, et qu'elle atteint donc son maximum aux bornes de l'intervalle sur lequel on l'étudie ; ainsi

$$\sup_{\theta \in]0, 1]} R(\theta, T) = \max(R(0, T), R(1, T)) = \max\left(\frac{1}{(1+n)^2}, \frac{1}{n \left(1 + \frac{1}{n}\right)^2}\right) = \frac{1}{n \left(1 + \frac{1}{n}\right)^2}$$

Il en ressort que T a le plus faible risque maximum, et qu'il est donc meilleur que \bar{X}_n au sens du risque minimax.

6. Calcul des risques bayésiens (à l'aide de l'indication) :

$$\begin{aligned} R(\nu, \bar{X}_n) &= \int R(\theta, \bar{X}_n) d\nu(\theta) = \int_0^{+\infty} R(\theta, \bar{X}_n) e^{-\theta} d\theta \\ &= \frac{1}{n} \int_0^{+\infty} \theta e^{-\theta} d\theta = \frac{1}{n} \\ R(\nu, T) &= \int R(\theta, T) d\nu(\theta) = \int_0^{+\infty} R(\theta, T) e^{-\theta} d\theta \\ &= \int_0^{+\infty} \left(\frac{\theta}{n \left(1 + \frac{1}{n}\right)^2} + \left(\frac{1-\theta}{1+n}\right)^2 \right) e^{-\theta} d\theta \\ &= \frac{1}{n \left(1 + \frac{1}{n}\right)^2} + \frac{1-2+2}{(1+n)^2} = \frac{1}{n+1} \end{aligned}$$

Le risque bayésien de T relatif à la loi a priori ν est plus faible que celui de \bar{X}_n (T est d'ailleurs l'estimateur bayésien de θ relatif à ν).

Ex 2.

1. Les variables X_i étant à valeurs dans \mathbb{R} , on choisit comme espace des observations du modèle statistique associé à (X_1, \dots, X_n) l'ensemble $\Omega = \mathbb{R}^n$, muni de la tribu borélienne $\mathcal{B}(\Omega)$. La famille de lois $\{\mathbb{P}_\theta, \theta \in \Theta = \mathbb{R}\}$ définies sur Ω est dominée par la mesure de Lebesgue $\lambda|_\Omega$ sur Ω , et est caractérisée par

$$\forall x \in \Omega, \quad \frac{d\mathbb{P}_\theta}{d\lambda|_\Omega}(x) = \prod_{i=1}^n f_\theta(x_i) = e^{-\sum_{i=1}^n x_i + n\theta} \prod_{i=1}^n \mathbf{1}_{x_i \geq \theta}$$

Posons $x_{(1)} = \min(x_1, \dots, x_n)$; on remarque que $\prod_{i=1}^n \mathbf{1}_{x_i \geq \theta} = \mathbf{1}_{x_{(1)} \geq \theta}$, d'où il vient

$$\forall x \in \Omega, \quad \frac{d\mathbb{P}_\theta}{d\lambda|_\Omega}(x) = e^{-\sum_{i=1}^n x_i + n\theta} \mathbf{1}_{x_{(1)} \geq \theta} = g_\theta(x_{(1)}) h(x)$$

avec

$$g_\theta(u) = e^{n\theta} \mathbf{1}_{u \geq \theta} \text{ et } h(x) = e^{-\sum_{i=1}^n x_i}$$

On déduit alors du théorème de factorisation que $X_{(1)}$ est une statistique exhaustive.

2. Calculons la fonction de répartition de la loi de $X_{(1)}$: quel que soit $t \in \mathbb{R}$,

$$\begin{aligned} F_{X_{(1)}}(t) &= \mathbb{P}_\theta(X_{(1)} \leq t) = 1 - \mathbb{P}_\theta(X_{(1)} > t) \\ &= 1 - \mathbb{P}_\theta(\forall i = 1, \dots, n, X_i > t) \\ &= 1 - \prod_{i=1}^n \mathbb{P}_\theta(X_i > t) \text{ car les } X_i \text{ sont indépendants entre eux} \\ &= 1 - \mathbb{P}_\theta(X_1 > t)^n \end{aligned}$$

Or

$$\mathbb{P}_\theta (X_1 > t) = \begin{cases} 1 & \text{si } t \leq \theta \\ \int_t^{+\infty} e^{-(u-\theta)} du = e^{-(t-\theta)} & \text{si } t \geq \theta \end{cases}$$

On en déduit la fonction de répartition

$$F_{X_{(1)}}(t) = \begin{cases} 0 & \text{si } t \leq \theta \\ 1 - e^{-n(t-\theta)} & \text{si } t \geq \theta \end{cases}$$

puis, en dérivant, la densité de la loi de $X_{(1)}$ par rapport à la mesure de Lebesgue sur \mathbb{R} :

$$\forall t \in \mathbb{R}, f_\theta^{X_{(1)}}(t) = \begin{cases} 0 & \text{si } t \leq \theta \\ ne^{-n(t-\theta)} & \text{si } t \geq \theta \end{cases} = ne^{-n(t-\theta)} \mathbf{1}_{t \geq \theta}$$

Pour montrer que $X_{(1)}$ est complète, supposons qu'il existe φ borélienne intégrable telle que $\mathbb{E}_\theta [\varphi(X_{(1)})] = 0$ quel que soit $\theta \in \mathbb{R}$. On peut supposer de plus que φ est continue. Alors l'hypothèse implique

$$\forall \theta \in \mathbb{R}, \int_\theta^{+\infty} \varphi(t) ne^{-n(t-\theta)} dt = 0$$

ce qui implique

$$\forall \theta \in \mathbb{R}, \int_\theta^{+\infty} \varphi(t) e^{-nt} dt = 0$$

On dérive par rapport à θ l'expression ci-dessus, d'où il appert que

$$\forall \theta \in \mathbb{R}, \varphi(\theta) e^{-n\theta} = 0$$

et l'on en déduit que φ est la fonction nulle. On a ainsi montré que $X_{(1)}$ est une statistique complète. Comme toute statistique exhaustive et complète est aussi minimale, il s'ensuit que $X_{(1)}$ est aussi minimale.

3. Soit $x \in \Omega$. Cherchons la valeur du paramètre θ qui maximise la vraisemblance du modèle

$$L(\theta, x) = e^{-\sum_{i=1}^n x_i + n\theta} \mathbf{1}_{x_{(1)} \geq \theta}$$

Sur l'intervalle $] -\infty, x_{(1)}]$, la fonction $\theta \mapsto L(\theta, x) = e^{-\sum_{i=1}^n x_i + n\theta}$ est croissante et atteint donc son maximum en $x_{(1)}$; sur l'intervalle $]x_{(1)}, +\infty[$, cette fonction est nulle. On en déduit

$$\sup_{\theta \in \Theta} L(\theta, x) = L(x_{(1)}, x)$$

Autrement dit, l'estimateur du maximum de vraisemblance de θ est la statistique $X_{(1)}$.

4. Calcul de $\mathbb{E}_\theta [X_{(1)}]$:

$$\begin{aligned} \mathbb{E}_\theta [X_{(1)}] &= \int_\theta^{+\infty} t ne^{-n(t-\theta)} dt = \int_0^{+\infty} \left(\frac{u}{n} + \theta\right) e^{-u} du \text{ en posant } u = n(t-\theta) \\ &= \frac{1}{n} + \theta \text{ d'après l'indication de l'exercice 1 question 6} \end{aligned}$$

On en déduit que $X_{(1)} - \frac{1}{n}$ est un estimateur sans biais de θ , fonction d'une statistique exhaustive et complète; c'est donc l'estimateur UVMB de θ , d'après le corollaire du théorème de Lehmann-Scheffé.

5. Calcul de $\mathbb{E}_\theta [\bar{X}_n]$:

$$\begin{aligned}\mathbb{E}_\theta [\bar{X}_n] &= \mathbb{E}_\theta [X_1] = \int_\theta^{+\infty} t e^{-(t-\theta)} dt \\ &= \int_0^{+\infty} (u + \theta) e^{-u} du \text{ en posant } u = t - \theta \\ &= 1 + \theta\end{aligned}$$

Il en résulte que $\bar{X}_n - 1$ est un estimateur sans biais de θ , et que $\mathbb{E}_\theta [\bar{X}_n - 1 | X_{(1)}]$ est l'estimateur UVMB de θ d'après le théorème de Lehmann-Scheffé. D'où $\mathbb{E}_\theta [\bar{X}_n - 1 | X_{(1)}] = X_{(1)} - \frac{1}{n}$. On en déduit

$$\mathbb{E}_\theta [\bar{X}_n | X_{(1)}] = X_{(1)} + 1 - \frac{1}{n}$$

6. Montrons que $X_{(1)}$ est faiblement consistant : soit $\varepsilon > 0$ quelconque,

$$\begin{aligned}\mathbb{P}_\theta (|X_{(1)} - \theta| > \varepsilon) &= \mathbb{P}_\theta (X_{(1)} - \theta > \varepsilon) \text{ car } X_{(1)} \geq \theta \\ &= P_\theta (X_{(1)} > \theta + \varepsilon) = \int_{\theta+\varepsilon}^{+\infty} n e^{-n(t-\theta)} dt = e^{-n\varepsilon}\end{aligned}$$

Il en découle que $\lim_{n \rightarrow +\infty} \mathbb{P}_\theta (|X_{(1)} - \theta| > \varepsilon) = 0$ quel que soit $\varepsilon > 0$, ce qui montre la convergence en probabilité de $X_{(1)}$ vers θ . La statistique $X_{(1)}$ est donc faiblement consistante.

Montrons maintenant qu'elle est aussi fortement consistante. On sait, d'après le lemme de Borel-Cantelli, que sous l'hypothèse

$$\forall \varepsilon > 0, \sum_{n \geq 1} \mathbb{P}_\theta (|X_{(1)} - \theta| > \varepsilon) < +\infty$$

alors $X_{(1)}$ converge \mathbb{P}_θ -p.s. vers θ quand n tend vers l'infini, autrement dit $X_{(1)}$ est fortement consistant. Or quel que soit $\varepsilon > 0$,

$$\sum_{n \geq 1} \mathbb{P}_\theta (|X_{(1)} - \theta| > \varepsilon) = \sum_{n \geq 1} e^{-n\varepsilon} = \frac{1}{1 - e^{-\varepsilon}} - 1 < +\infty$$

On en déduit le résultat attendu.

7. Déterminons la loi de $n(X_{(1)} - \theta)$: soit φ fonction borélienne bornée quelconque ; alors

$$\begin{aligned}\mathbb{E}_\theta [\varphi (n(X_{(1)} - \theta))] &= \int_\theta^{+\infty} \varphi (n(t - \theta)) n e^{-n(t-\theta)} dt \\ &= \int_0^{+\infty} \varphi (u) e^{-u} du \text{ en posant } u = n(t - \theta)\end{aligned}$$

La variable aléatoire $n(X_{(1)} - \theta)$ suit donc une loi exponentielle de paramètre 1, pour tout n . Il en résulte bien évidemment qu'elle converge en loi, quand n tend vers $+\infty$, vers cette loi exponentielle.

L'estimateur $X_{(1)}$ est-il asymptotiquement sans biais ?

- Premier sens : $X_{(1)}$ est asymptotiquement sans biais car son biais tend vers 0

$$\mathbb{E}_\theta [X_{(1)}] = \theta + \frac{1}{n} \xrightarrow{n \rightarrow +\infty} \theta$$

- Deuxième sens : $X_{(1)}$ n'est pas asymptotiquement sans biais car le rapport de son risque et de sa variance ne tend pas vers 1

$$\begin{aligned} \frac{R(\theta, X_{(1)})}{\text{var}_\theta (X_{(1)})} &= 1 + \frac{(\mathbb{E}_\theta [X_{(1)}] - \theta)^2}{\text{var}_\theta (X_{(1)})} \\ &= 1 + \frac{(\theta + \frac{1}{n} - \theta)^2}{\frac{1}{n^2} \text{var}_\theta (n(X_{(1)} - \theta))} \\ &= 1 + 1 = 2 \end{aligned}$$

en utilisant le fait que la variance d'une variable exponentielle de paramètre 1 est 1.

- Troisième sens : $X_{(1)}$ n'est pas asymptotiquement sans biais car $n(X_{(1)} - \theta)$ converge en loi vers une loi d'espérance 1, non nulle.