

Partiel

Eléments de correction

Exercice 1 : (10 points)

On considère le modèle $(\mathcal{N}(\theta, \theta^2), \theta > 0)$. Soient \bar{X}_n et S_n^2 la moyenne et la variance empirique de l'échantillon (X_1, \dots, X_n) issu du modèle. Le but de l'exercice est de comparer les estimateurs \bar{X}_n et $T_n = \sqrt{S_n^2}$.

1. **Cours** :

Rappelez la définition de statistique exhaustive et de statistique complète.

2. Montrer que (\bar{X}_n, S_n^2) est une statistique exhaustive.

On utilise le théorème de factorisation avec la formule de Huygens $s_n^2 = \overline{x_n^2} - (\bar{x}_n)^2$ pour tout (X_1, \dots, x_n) :

$$\begin{aligned} f((x_1, \dots, x_n); \theta) &= (2\pi\theta^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\theta^2}\right) \\ &= (2\pi\theta^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n x_i^2 - 2\theta \sum_{i=1}^n x_i + n\theta^2}{2\theta^2}\right) \\ &= (2\pi\theta^2 e)^{-n/2} \exp\left(-\frac{n(\overline{x_n^2} - 2\theta\bar{x}_n)}{2\theta^2}\right) \\ &= (2\pi\theta^2 e)^{-n/2} \exp\left(-\frac{n(s_n^2 + (\bar{x}_n)^2 - 2\theta\bar{x}_n)}{2\theta^2}\right). \end{aligned}$$

Ici, on trouve $h(x) = 1$ et

$$g((v, w), \theta) = (2\pi\theta^2 e)^{-n/2} \exp\left(-\frac{n(w + v^2 - 2\theta v)}{2\theta^2}\right)$$

pour $(v, w) = (\bar{X}_n, S_n^2)$.

3. Calculer l'information de Fisher associée au modèle.

On suppose que les hypothèses usuelles d'un modèle régulier sont satisfaites. On calcule l'information de Fisher donnée par la formule $I(\theta) = -\mathbb{E}_\theta((\log f(X, \theta))'')$. On a

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

soit en dérivant une première fois

$$(\log f(X, \theta))' = -\frac{1}{\theta} + \frac{X - \theta}{\theta^2} + \frac{(X - \theta)^2}{\theta^3}$$

puis une deuxième fois

$$(\log f(X, \theta))'' = \frac{1}{\theta^2} - \frac{1}{\theta^2} - \frac{2(X - \theta)}{\theta^3} - \frac{2(X - \theta)}{\theta^3} - \frac{3(X - \theta)^2}{\theta^4}.$$

Par linéarité de l'espérance et en utilisant les relations $\mathbb{E}_\theta(X - \theta) = 0$ et $\mathbb{E}_\theta[(X - \theta)^2] = \text{Var}_\theta(X) = \theta^2$ on trouve facilement $I(\theta) = 3\theta^{-2}$.

4. Le théorème de Fisher assure que \bar{X}_n et S_n^2 sont 2 v.a. indépendantes. En déduire la loi du couple (\bar{X}_n, S_n^2) .

Rappel : La loi du χ_d^2 a pour densité $f(x) = \frac{1}{2^{d/2}\Gamma(d/2)}x^{d/2-1}\exp(-\frac{x}{2})1_{\{x \geq 0\}}$.

On sait que $\bar{X}_n \sim \mathcal{N}(\theta, \theta^2/n)$ et $n\theta^{-2}S_n^2 \sim \chi_{n-1}^2$. Par la méthode de la fonction muette, on trouve la densité de S_n^2

$$f_{S_n^2}(y) = \frac{n^{(n-1)/2}}{2^{(n-1)/2}\theta^{n-1}\Gamma((n-1)/2)}y^{(n-1)/2-1}\exp\left(-\frac{ny}{2\theta^2}\right)1_{\{y \geq 0\}}$$

On donne la loi du couple (\bar{X}_n, S_n^2) en déterminant sa densité. Par indépendance, en notant g la densité du couple (\bar{X}_n, S_n^2) , on trouve le produit des densités des marginales :

$$g((x, y), \theta) = \sqrt{\frac{n}{2\pi\theta^2}}\exp\left(-\frac{n(x - \theta)^2}{2\theta^2}\right)\frac{n^{(n-1)/2}y^{(n-1)/2-1}}{2^{(n-1)/2}\theta^{n-1}\Gamma((n-1)/2)}\exp\left(-\frac{ny}{2\theta^2}\right)1_{\{y \geq 0\}}.$$

5. Calculer l'information de Fisher associée à la statistique (\bar{X}_n, S_n^2) . Pouvait-on s'attendre au résultat ?

On suppose que le modèle issu du couple (\bar{X}_n, S_n^2) est régulier donc que l'information de Fisher est donnée par la formule $I_{(\bar{X}_n, S_n^2)}(\theta) = -\mathbb{E}_\theta((\log(g((\bar{X}_n, S_n^2), \theta)))')$. On trouve

$$\log(g((\bar{X}_n, S_n^2), \theta)) = -\log(\theta) - \frac{n(\bar{X}_n - \theta)^2}{2\theta^2} - (n-1)\log(\theta) - \frac{nS_n^2}{2\theta^2} + K$$

où K est une constante ne dépendant pas de θ , soit, en dérivant une première fois

$$(\log(g((\bar{X}_n, S_n^2), \theta)))' = -\frac{n}{\theta} + \frac{n(\bar{X}_n - \theta)}{\theta^2} + \frac{n(\bar{X}_n - \theta)^2}{\theta^3} + \frac{nS_n^2}{\theta^3}$$

et en dérivant une deuxième fois

$$(\log(g((\bar{X}_n, S_n^2), \theta)))'' = \frac{n}{\theta^2} - \frac{n}{\theta^2} - \frac{2n(\bar{X}_n - \theta)}{\theta^3} - \frac{2n(\bar{X}_n - \theta)}{\theta^3} - \frac{3n(\bar{X}_n - \theta)^2}{\theta^4} - \frac{3nS_n^2}{\theta^4}$$

Par linéarité de l'espérance, en utilisant les relations $\mathbb{E}_\theta(\bar{X}_n - \theta) = 0$, $n\mathbb{E}_\theta(\bar{X}_n - \theta)^2 = \theta^2$ et $n\mathbb{E}_\theta(S_n^2) = (n-1)\theta^2$ on trouve $I_{(\bar{X}_n, S_n^2)}(\theta) = 3n/\theta^2$. Ce résultat était attendu car une statistique exhaustive a la même information de Fisher que celle de l'échantillon $I_n(\theta) = nI(\theta) = 3n/\theta^2$.

6. Donner les valeurs de $\mathbb{E}_\theta((\bar{X}_n)^2)$ et $\mathbb{E}_\theta(S_n^2)$ et en déduire que (\bar{X}_n, S_n^2) n'est pas une statistique complète.

D'après la formule de Huygens, on trouve $\mathbb{E}_\theta((\bar{X}_n)^2) = \mathbb{E}_\theta(\bar{X}_n^2) - \mathbb{E}_\theta(S_n^2)$. Or $\mathbb{E}_\theta(\bar{X}_n^2) = \mathbb{E}_\theta(X^2)$ par linéarité soit $\mathbb{E}_\theta(\bar{X}_n^2) = 2\theta^2$ en réutilisant la formule de Huygens. Le cours donne $\mathbb{E}_\theta(S_n^2) = (n-1)\theta^2/n$ donc finalement $\mathbb{E}_\theta((\bar{X}_n)^2) = (n+1)\theta^2/n$. On déduit que $\mathbb{E}_\theta(f(\bar{X}_n, S_n^2)) = 0$ avec $f(x, y) = (n-1)x^2 - (n+1)y$. Or $f(\bar{X}_n, S_n^2) \neq 0$, ce qui est en contradiction avec la définition de statistique complète et (\bar{X}_n, S_n^2) n'en est pas une.

7. En utilisant le TCL, montrer que \bar{X}_n est un estimateur asymptotiquement normal et donner sa variance asymptotique.

On trouve directement en appliquant le TCL avec $\mathbb{E}_\theta(X^2) = \theta^2 < \infty$ que

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta^2).$$

On reconnaît un estimateur asymptotiquement normal de θ de variance asymptotique θ^2 .

8. Calculer $\mu_4 = \mathbb{E}_\theta(X - \theta)^4$ en utilisant une IPP.

On calcule

$$\mathbb{E}_\theta(X - \theta)^4 = \int_{\mathbb{R}} (x - \theta)^4 \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x - \theta)^2}{2\theta^2}\right)$$

en faisant l'IPP $(x - \theta)^3 \rightarrow 3(x - \theta)^2$ et

$$(x - \theta)(2\pi\theta^2)^{-1/2} \exp(-(x - \theta)^2/(2\theta^2)) \rightarrow \theta(2\pi)^{-1/2} \exp(-(x - \theta)^2/(2\theta^2))$$

et on trouve

$$\mathbb{E}_\theta(X - \theta)^4 = 3\theta^2 \int_{\mathbb{R}} (x - \theta)^2 \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x - \theta)^2}{2\theta^2}\right).$$

On reconnaît la formule de la variance $\text{Var}_\theta(X) = \mathbb{E}_\theta[(X - \theta)^2] = \theta^2$ d'où finalement $\mu_4 = 3\theta^2$.

9. En utilisant le comportement asymptotique de S_n^2 vu en cours, montrer que

$$\sqrt{n}(S_n^2 - \theta^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\theta^4).$$

On a vu en cours que

$$\sqrt{n}(S_n^2 - \theta^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_4 - \theta^4).$$

D'après la question précédente, $\mu_4 - \theta^4 = 2\theta^4$ d'où le résultat.

10. En déduire que T_n est asymptotiquement normale et donner sa variance asymptotique. Les estimateurs \bar{X}_n et T_n sont-ils asymptotiquement efficaces? Lequel choisir?

Comme $T_n = \sqrt{S_n^2}$, on utilise la δ -méthode appliquée à la fonction $x \rightarrow \sqrt{x}$ bien définie sur \mathbb{R}_+^* de dérivée $1/(2\sqrt{x})$. On trouve

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1/(2\sqrt{\theta^2}))^2 2\theta^4) = \mathcal{N}(0, \theta^2/2).$$

On reconnaît un estimateur T_n asymptotiquement normal de variance asymptotique $\theta^2/2$. Aucun des 2 estimateurs n'est asymptotiquement efficace car la borne de Cramer-Rao asymptotique vaut $\theta^2/3$. Toutefois T_n est préférable à \bar{X}_n car sa variance asymptotique est plus petite.

Exercice 2 : (10 points)

On considère le modèle Gaussien multidimensionnel $(\mathcal{N}_d(\theta, I_d), \theta \in \mathbb{R}^d)$ où I_d est la matrice identité de dimension $d \times d$. Dans ce problème, nous allons mettre en évidence le phénomène (paradoxe) de Stein qui dit que la moyenne empirique \bar{X}_n de l'échantillon (X_1, \dots, X_n) issu du modèle est un estimateur inadmissible pour $d \geq 3$. Dans tout l'exercice on note $\theta = (\theta_1, \dots, \theta_d)^T$, $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ et $\bar{X}_n = (\bar{X}_n^{(1)}, \dots, \bar{X}_n^{(d)})^T$.

1. **Cours :**

Rappeler la définition d'un estimateur inadmissible.

2. Donner la densité de \bar{X}_n .

On sait que $\bar{X}_n \sim \mathcal{N}_d(\theta, n^{-1}I_d)$ de densité

$$f(x) = \left(\frac{n}{2\pi}\right)^{d/2} \exp(-n\|x - \theta\|^2/2)$$

car $\det(n^{-1}I_d) = n^{-1}$ et $x^T I_d^{-1} x = \|x\|^2$.

3. Calculer le risque quadratique associé à \bar{X}_n en utilisant la décomposition biais-variance.

On a la décomposition biais-variance du risque vue en cours $R(\bar{X}_n, \theta) = \text{Tr}(\text{Var}_\theta(\bar{X}_n)) + \|b_n(\theta)\|^2$. Ici, $b_n(\theta) = 0$, $\text{Var}_\theta(\bar{X}_n) = n^{-1}I_d$ et $\text{Tr}(n^{-1}I_d) = d/n$ d'où $R(\bar{X}_n, \theta) = d/n$.

4. Calculer l'information de Fisher associée à l'échantillon (X_1, \dots, X_n) et vérifier que \bar{X}_n est bien l'estimateur sans biais de variance minimale.

On est dans un modèle Gaussien donc régulier et l'information de Fisher est donnée par la formule $I(\theta) = -\mathbb{E}_\theta(\mathbb{H}_\theta(\log f(X, \theta)))$ avec ici

$$\log f(X, \theta) = -\frac{d}{2} \log(2\pi) - \frac{\|x - \theta\|^2}{2} = -\frac{d}{2} \log(2\pi) - \frac{\sum_{i=1}^d (x_i - \theta_i)^2}{2}.$$

En dérivant par rapport à θ_i , on trouve

$$\frac{\partial \log f(X, \theta)}{\partial \theta_i} = x_i - \theta_i$$

puis une seconde fois par rapport à θ_j :

$$\frac{\partial^2 \log f(X, \theta)}{\partial \theta_j \partial \theta_i} = 0 \quad \text{si } i \neq j \quad \frac{\partial^2 \log f(X, \theta)}{\partial \theta_j \partial \theta_i} = -1 \quad \text{sinon.}$$

D'où l'information de Fisher du modèle $I(\theta) = I_d$. On en déduit l'information de Fisher de l'échantillon $I_n(\theta) = nI_d$ et la borne de Cramer-Rao $I_n^{-1}(\theta) = n^{-1}I_d$. Comme c'est aussi la variance de \bar{X}_n qui est sans biais, l'estimateur est efficace donc de variance minimale.

5. Dans toute la suite $d \geq 3$. Soit l'estimateur de James-Stein défini par

$$T_n = \left(1 - \frac{d-2}{n\|\bar{X}_n\|^2}\right) \bar{X}_n.$$

Quelle est la loi suivie par $n\|\bar{X}_n\|^2$ dans le cas $\theta = 0_d$? En déduire que l'estimateur de James-Stein est bien défini P_θ -p.s. dans ce cas-là.

Dans le cas $\theta = 0_d$ le théorème de Cochran donne directement $n\|\bar{X}_n\|^2 = \sum_{i=1}^d (\sqrt{n}\bar{X}_n^{(i)})^2 \sim \chi_d^2$ car c'est la norme au carré du vecteur gaussien $(\sqrt{n}\bar{X}_n^{(1)}, \dots, \sqrt{n}\bar{X}_n^{(d)}) \sim \mathcal{N}_d(0_d, I_d)$ qui est aussi la projection de lui-même sur l'espace vectoriel \mathbb{R}^d en entier.

6. Montrer que le risque quadratique de T_n vaut

$$R(T_n, \theta) = R(\bar{X}_n, \theta) + 2 \frac{d-2}{n} \mathbb{E}_\theta \left[\frac{(\theta - \bar{X}_n)^T \bar{X}_n}{\|\bar{X}_n\|^2} \right] + \frac{(d-2)^2}{n^2} \mathbb{E}_\theta \left[\frac{1}{\|\bar{X}_n\|^2} \right].$$

On écrit

$$R(T_n, \theta) = \mathbb{E}_\theta \|T_n^\theta\|^2 = \mathbb{E}_\theta (T_n - \theta)^T (T_n - \theta).$$

Puis, comme $T_n - \theta = \bar{X}_n - \theta - (d-2)/\|\bar{X}_n\|^2 \bar{X}_n$ on trouve le résultat en développant.

7. Soit h une fonction de $\mathbb{R}^d \mapsto \mathbb{R}$ telle que $(\theta - \bar{X}_n)^T h(\bar{X}_n)$ et $\nabla h(\bar{X}_n)$ soient intégrables. En utilisant une IPP, montrer que

$$\mathbb{E}_\theta [(\theta_i - \bar{X}_n^{(i)}) h(\bar{X}_n)] = -n^{-1} \mathbb{E}_\theta \left[\frac{\partial h}{\partial x_i}(\bar{X}_n) \right].$$

D'après la première question, on peut écrire

$$\mathbb{E}_\theta [(\theta_i - \bar{X}_n^{(i)}) h(\bar{X}_n)] = \left(\frac{n}{2\pi} \right)^{d/2} \int_{\mathbb{R}^d} (\theta_i - x_i) h(x) \exp(-n\|x - \theta\|^2/2) dx.$$

On remarque que $\partial \exp(-n\|x - \theta\|^2/2) / \partial x_i = n(\theta_i - x_i) \exp(-n\|x - \theta\|^2/2)$ donc on effectue l'IPP $h(x) \rightarrow \partial h(x) / \partial x_i$ et $n(\theta_i - x_i) \exp(-n\|x - \theta\|^2/2) \rightarrow \exp(-n\|x - \theta\|^2/2)$ qui donne

$$\begin{aligned} & \left(\frac{n}{2\pi} \right)^{d/2} \int_{\mathbb{R}^d} (\theta_i - x_i) h(x) \exp(-n\|x - \theta\|^2/2) \\ &= \left[h(x) \frac{n^{d/2-1}}{(2\pi)^{d/2}} \exp(-n\|x - \theta\|^2/2) \right]_{x_i \rightarrow -\infty}^{x_i \rightarrow +\infty} - \int_{\mathbb{R}^d} \frac{\partial h(x)}{\partial x_i} \frac{n^{d/2-1}}{(2\pi)^{d/2}} \exp(-n\|x - \theta\|^2/2) dx \\ &= -n^{-1} \int_{\mathbb{R}^d} \frac{\partial h(x)}{\partial x_i} \frac{n^{d/2}}{(2\pi)^{d/2}} \exp(-n\|x - \theta\|^2/2) dx \end{aligned}$$

d'où le résultat souhaité.

8. Pour tout $1 \leq i \leq d$, montrer que $h(x) = x_i / \|x\|^2$ vérifie la relation

$$\frac{\partial h}{\partial x_i}(x) = \frac{1}{\|x\|^2} - \frac{2x_i^2}{\|x\|^4}.$$

On écrit

$$h(x) = \frac{x_i}{\|x\|^2} = \frac{x_i}{\sum_{i=1}^d x_i^2}$$

et le résultat suit de la dérivation.

9. En admettant que $h(x) = x_i / \|x\|^2$ satisfait les conditions d'intégrabilité de la question 7. pour tout $1 \leq i \leq d$, montrer que

$$\mathbb{E}_\theta \left[\frac{(\theta - \bar{X}_n)^T \bar{X}_n}{\|\bar{X}_n\|^2} \right] = -\frac{d-2}{n} \mathbb{E}_\theta \left[\frac{1}{\|\bar{X}_n\|^2} \right].$$

On remarque tout d'abord que par linéarité de l'espérance

$$\mathbb{E}_\theta \left[\frac{(\theta - \bar{X}_n)^T \bar{X}_n}{\|\bar{X}_n\|^2} \right] = \sum_{i=1}^d \mathbb{E}_\theta [(\theta_i - \bar{X}_n^{(i)}) h_i(\bar{X}_n)]$$

avec $h_i(x) = x_i/\|x\|^2$ pour tout $1 \leq i \leq d$. En utilisant le résultat de la question 7 simultanément pour chaque h_i , on trouve

$$\mathbb{E}_\theta \left[\frac{(\theta - \bar{X}_n)^T \bar{X}_n}{\|\bar{X}_n\|^2} \right] = -n^{-1} \sum_{i=1}^d \mathbb{E}_\theta \left[\frac{\partial h_i}{\partial x_i}(\bar{X}_n) \right].$$

Puis en utilisant le calcul de la question 8, on a

$$\mathbb{E}_\theta \left[\frac{(\theta - \bar{X}_n)^T \bar{X}_n}{\|\bar{X}_n\|^2} \right] = -n^{-1} \sum_{i=1}^d \mathbb{E}_\theta \left[\frac{1}{\|\bar{X}_n\|^2} - 2 \frac{(\bar{X}_n^{(i)})^2}{\|\bar{X}_n\|^4} \right].$$

Par linéarité de l'espérance on obtient le résultat souhaité en remarquant que $\|\bar{X}_n\|^2 = \sum_{i=1}^d (\bar{X}_n^{(i)})^2$.

10. En déduire le signe de $R(T_n, \theta) - R(\bar{X}_n, \theta)$ pour tout $\theta \in \mathbb{R}^d$ et conclure.

En utilisant l'expression de la question 6, on obtient directement la simplification

$$R(T_n, \theta) = R(\bar{X}_n, \theta) - \frac{(d-2)^2}{n^2} \mathbb{E}_\theta \left[\frac{1}{\|\bar{X}_n\|^2} \right].$$

On en déduit que $R(T_n, \theta) - R(\bar{X}_n, \theta) < 0$ pour tout $\theta \in \mathbb{R}^d$ et donc que \bar{X}_n est inadmissible.