Université Paris-Dauphine — Année 2008-2009
U.F.R. Mathématiques de la décision
L3 - Statistique Mathématique

Examen

Durée 2h00. Le barême est donné à titre indicatif.

Exercice 1 : *(5 points)* Soit $(X_1, \ldots, X_n)$ un échantillon de va iid de même loi que X, où X admet pour densité de probabilité $f(x, \theta)$ par rapport à la mesure Lebesgue, définie par :

$$f(x, \theta) = k x^{-(p+1)} \exp\left(-\frac{\theta}{x}\right) \mathbf{1}_{x>0},$$

avec $\theta \in]0, +\infty[$ un paramètre réel inconnu et $p > 0$ un nombre connu.

1. On pose $U = \dfrac{1}{X}$. Montrer que $U \sim Gamma(p, \theta)$ et en déduire la constante k, $\mathbb{E}U$ et $\operatorname{Var} U$.
2. Déterminer une statistique exhaustive complète pour θ.
3. Déterminer $\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ.
4. A l'aide de la question 1, déterminer la loi de $Z = \displaystyle\sum_{i=1}^{n} \frac{1}{X_i}$ et montrer que :

$$\mathbb{E}\widehat{\theta}_n = \frac{np\theta}{np-1}, \quad \operatorname{Var}\widehat{\theta}_n = \frac{(np\theta)^2}{(np-1)^2(np-2)}.$$

5. L'estimateur $\widehat{\theta}_n$ est-il sans biais ? convergeant ? efficace ?

Exercice 2 :*(5 points)* Soit $(X_1, \cdots, X_n)$ un échantillon de va iid de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$:

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

où $\theta \in \mathbb{R}$ est inconnu.

1. Calculer l'EMV de θ noté $\widehat{\theta}_n$.
2. Déterminer la forme de la zone de rejet d'un test des hypothèses $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$ utilisant l'EMV $\widehat{\theta}_n$.
3. Déterminer la statistique $\widehat{T}$ du test du rapport de vraisemblance pour les mêmes hypothèses H_0 et H_1. Qu'en déduire pour le test de la question 2 ?
4. Déterminer la fonction de répartition de $\widehat{T}$.
5. Expliciter la zone de rejet du test du rapport de vraisemblance pour les mêmes hypothèses H_0 contre H_1 de niveau α fixé et déterminer sa fonction puissance. La suite de tests associée est-elle consistante ?

Problème *(10 points)* Soit $(X_1, \dots, X_n)$ un échantillon de va iid de fonction de répartition F supposée continue et strictement croissante. On note F_n la fonction de répartition empirique

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{X_j \leq x}$$

La statistique de Cramer & Von Mises est définie par $\widehat{T}_n = \int_{\mathbb{R}} (F_n(x) - F(x))^2 \, dF(x)$. La présence de l'intégrale dans le terme précédent rend son utilisation ardue. En pratique, on l'approche par la statistique

$$\widehat{T}_n^* = \int_{\mathbb{R}} (F_n(x) - F(x))^2 \, dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} (F_n(X_i) - F(X_i))^2.$$

Le but de ce problème est de quantifier l'erreur d'approximation alors commise.

1. Montrez que $F(X_i)$ est une variable aléatoire uniforme et en déduire que $\mathbb{E}F(X_i) = \dfrac{1}{2}$ et $\operatorname{Var} F(X_i) = \dfrac{1}{12}$.
2. Montrez que $nF_n(x) \sim B(n, F(x))$ et en déduire l'espérance et la variance de $nF_n(x)$.
3. En déduire que $\mathbb{E}\widehat{T}_n = \dfrac{1}{6n}$.
4. Montrez que $\widehat{T}_n^\star = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \left(\frac{i}{n} - F(X_{(i)}) \right)^2$.
5. En admettant que
$$n\widehat{T}_n = \frac{1}{12n} + \sum_{k=1}^{n} \Big(\frac{2k-1}{2n} - F(X_{(k)}) \Big)^2$$
montrer que
$$\widehat{T}_n - \widehat{T}_n^\star = -\frac{1}{2n} - \frac{1}{6n^2} + \frac{1}{n^2} \sum_{i=1}^{n} F(X_i).$$
6. En déduire que $\lim_{n\to\infty} n^{3/2} \mathbb{E}(\widehat{T}_n^\star - \widehat{T}_n) = 0$ et que $\operatorname{Var}(\widehat{T}_n^\star - \widehat{T}_n) = \dfrac{1}{12n^3}$.
7. Montrer aussi que $n^{3/2}(\widehat{T}_n^\star - \widehat{T}_n) \to \mathcal{N}\Big(0, \dfrac{1}{12}\Big)$. Conclure.
8. (Facultatif) Montrer que
$$n\widehat{T}_n = \frac{1}{12n} + \sum_{k=1}^{n} \Big(\frac{2k-1}{2n} - F(X_{(k)}) \Big)^2.$$

Eléments de correction

Exercice 1 : *(5 points)* Soit $(X_1, \ldots, X_n)$ un échantillon de va iid de même loi que X, où X admet pour densité de probabilité $f(x, \theta)$ par rapport à la mesure Lebesgue, définie par :

$$f(x, \theta) = kx^{-(p+1)} \exp\left(-\frac{\theta}{x}\right) \mathbf{1}_{x>0},$$

avec $\theta \in]0, +\infty[$ un paramètre réel inconnu et $p > 0$ un nombre connu.

1. On pose $U = \dfrac{1}{X}$. Montrer que $U \sim Gamma(p, \theta)$ et en déduire la constante k, $\mathbb{E}U$ et $\operatorname{Var} U$.
   On applique la méthode du changement de variable. Soit g une fonction continue bornée alors par changement de variable $y = 1/x$, on a

   $$\mathbb{E}(g(U)) = k \int_0^\infty g(1/x) x^{-(p+1)} \exp(-\theta/x)\, dx = k \int_0^\infty g(y) y^{p-1} \exp(-\theta y)\, dy.$$

   On reconnaît la densité d'une $Gamma(p, \theta)$ à la constante $k = \theta^p/\Gamma(p)$ près. On en déduit $\mathbb{E}U = p/\theta$ et $\operatorname{Var} U = p/\theta^2$.

2. Déterminer une statistique exhaustive complète pour θ.
   On est dans le cadre du modèle exponentiel à un paramètre avec

   $$\begin{aligned} c(\theta) &= \frac{\theta^p}{\Gamma(p)}, \\ h(x) &= x^{-(p+1)} \mathbf{1}_{x>0}, \\ \alpha_1(\theta) &= \theta, \\ T_1(x) &= -\frac{1}{x}. \end{aligned}$$

   Etant donné que $\alpha_1(\Theta) = \alpha_1(]0; +\infty[) =]0; +\infty[$, ouvert de dimension 1 et que T_1 est affinement indépendante, on en déduit que

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

   est une statistique exhaustive complète pour θ.

3. Déterminer $\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ.
   On est dans le cadre d'un modèle régulier exponentiel où $\beta(\theta) = p \ln(\theta)$ (notation de la page 79 du polycopié). Or β es deux fois continûment dérivable de dérivée seconde $-p/\theta^2 < 0$ donc l'EMV vérifie

   $$\beta'(\widehat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n \frac{1}{X_i} \Leftrightarrow \widehat{\theta}_n = \frac{np}{\sum_{i=1}^n \frac{1}{X_i}}$$

4. A l'aide de la question 1, déterminer la loi de $Z = \displaystyle\sum_{i=1}^n \frac{1}{X_i}$ et montrer que :

   $$\mathbb{E}\widehat{\theta}_n = \frac{np\theta}{np-1}, \quad \operatorname{Var} \widehat{\theta}_n = \frac{(np\theta)^2}{(np-1)^2(np-2)}.$$

D'après la question 1, $1/X_i$ est une suite iid de va suivant la loi $Gamma(p,\theta)$ donc $Z \sim Gamma(np,\theta)$. On en déduit que

$$\mathbb{E}(\widehat{\theta}_n) = np\mathbb{E}\left(\frac{1}{Z}\right) = np\int_0^\infty \frac{1}{z}\frac{\theta^{np}}{\Gamma(np)}z^{np-1}\exp(-\theta z)dz = \frac{np\theta}{\Gamma(np)}\int_0^\infty u^{np-1-1}\exp(-u)du$$

en effectuant le changement de variable $u = \theta z$. On reconnaît la quantité $\Gamma(np-1)$ et la relation $\Gamma(np) = (np-1)\Gamma(np-1)$ permet de conclure pour le calcul de l'espérance. Reste à calculer $\mathbb{E}(1/Z^2)$. Avec le même changement de variable que précédemment, on obtient $\mathbb{E}(1/Z^2) = \theta^2\Gamma(np-2)/\Gamma(np)$ d'où

$$\mathrm{Var}\,\widehat{\theta}_n = (np)^2\mathbb{E}(1/Z^2) - \frac{(np\theta)^2}{(np-1)^2} = (np\theta)^2\left(\frac{1}{(np-1)(np-2)} - \frac{1}{(np-1)^2}\right)$$

et le résultat souhaité suit facilement.

5. L'estimateur $\widehat{\theta}_n$ est-il sans biais ? convergeant ? efficace ?
   L'estimateur $\widehat{\theta}_n$ est biaisé d'après la question précédente. Il ne peut donc pas être efficace. Il est toutefois convergeant car asymptotiquement sans biais et de variance tendat vers 0.

Exercice 2 :*(5 points)* Soit $(X_1,\cdots,X_n)$ un échantillon de va iid de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$:

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

où $\theta \in \mathbb{R}$ est inconnu.

1. Calculer l'EMV de θ noté $\widehat{\theta}_n$.
   On écrit la vraisemblance

   $$L_n(\theta) = e^{n\theta-\sum_{i=1}^n X_i}\mathbf{1}_{\min(X_1,\ldots,X_n)\geq\theta}$$

   qui est maximale lorsque $\widehat{\theta}_n = \min(X_1,\ldots,X_n)$.
2. Déterminer la forme de la zone de rejet d'un test des hypothèses $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$ utilisant l'EMV $\widehat{\theta}_n$.
   La zone de rejet sera de la forme

   $$W = \{\widehat{\theta}_n > k\}$$

   où k est un seuil à calculer en fonction du seuil souhaité.
3. Déterminer la statistique $\widehat{T}$ du test du rapport de vraisemblance pour les mêmes hypothèses H_0 et H_1. Qu'en déduire pour le test de la question 2 ?
   La vraisemblance $L_n(\theta)$ vérifie le critère de factorisation pour l'EMV $\widehat{\theta}_n$ avec

   $$h(\mathbf{x}) = e^{-\sum_{i=1}^n x_i} \text{ et } g(\theta,\widehat{\theta}_n) = e^{n\theta}\mathbf{1}_{\widehat{\theta}_n\geq\theta}.$$

   Donc l'EMV est une statistique exhaustive. De plus, pour tout $\theta_0 < \theta_1$ on a

   $$V_{\theta_0,\theta_1} = e^{\theta_0-\theta_1}\mathbf{1}_{\widehat{\theta}_n\geq\theta_1}$$

   qui est une fonction croissante en $\widehat{\theta}_n$. On est donc dans un modèle à RVM pour $\widehat{T} = \widehat{\theta}_n$ et le test de la question 2 est UPP parmi tous les tests de niveau $\leq \mathbb{P}_{\theta_0}(W)$ (Karlin-Rubin).

4. Déterminer la fonction de répartition de $\widehat{T}$.
   On a
   $$\mathbb{P}_\theta(\widehat{T} \leq x) = 1 - \mathbb{P}_\theta(\widehat{T}_n > x) = 1 - \prod_{i=1}^{n} \mathbb{P}_\theta(X_i > x),$$
   par indépendance des X_i. Mais pour $x > \theta$ on a
   $$\mathbb{P}_\theta(X_i > x) = \int_x^\infty e^{-(u-\theta)} du = \exp(-(x-\theta))$$
   pour tout i donc
   $$\mathbb{P}_\theta(\widehat{T} \leq x) = 1 - (\exp(-(x-\theta)))^n = 1 - \exp(-n(x-\theta))$$
   pour $x > \theta$ et $\mathbb{P}_\theta(\widehat{T} \leq x) = 0$ autrement.
5. Expliciter la zone de rejet du test du rapport de vraisemblance pour les mêmes hypothèses H_0 contre H_1 de niveau α fixé et déterminer sa fonction puissance. La suite de tests associée est-elle consistante ?
   Etant donné que $W = \{\widehat{T} > k\}$ on cherche $k > \theta_0$ tel que
   $$\alpha = \mathbb{P}_{\theta_0}(\widehat{T} > k) = \exp(-n(k-\theta_0))$$
   et on trouve
   $$k = \theta_0 + \frac{1}{n} \ln\left(\frac{1}{\alpha}\right).$$
   La fonction puissance vaut
   $$\beta(\theta) = \mathbb{P}_\theta(W) = \mathbb{P}_\theta(\widehat{T} > k) = \left(\int_{\theta \vee k}^\infty e^{-(u-\theta)} du\right)^n = 1 \wedge \exp(n(\theta - k))$$
   $$= 1 \wedge \alpha \exp\left(n\left(\theta - \theta_0\right)\right)$$
   en remplaçant k par sa valeur pour tout $\theta > \theta_0$.
   Quand $n \to \infty$ et comme $\theta - \theta_0 > 0$ alors $\exp(n(\theta - \theta_0)) \to \infty$ donc $\beta(\theta) \to 1$. La suite de test est bien consistante.

Problème *(10 points)* Soit $(X_1, \ldots, X_n)$ un échantillon de va iid de fonction de répartition F supposée continue et strictement croissante. On note F_n la fonction de répartition empirique
$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{X_j \leq x}$$
La statistique de Cramer & Von Mises est définie par $\widehat{T}_n = \int_{\mathbb{R}} (F_n(x) - F(x))^2 \, dF(x)$. La présence de l'intégrale dans le terme précédent rend son utilisation ardue. En pratique, on l'approche par la statistique
$$\widehat{T}_n^* = \int_{\mathbb{R}} (F_n(x) - F(x))^2 \, dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} (F_n(X_i) - F(X_i))^2.$$
Le but de ce problème est de quantifier l'erreur d'approximation alors commise.

1. Montrez que $F(X_i)$ est une variable aléatoire uniforme et en déduire que $\mathbb{E}F(X_i) = \dfrac{1}{2}$ et $\operatorname{Var} F(X_i) = \dfrac{1}{12}$.
   Une fonction continue et strictement croissante est inversible. Soit F^{-1} son inverse allant de $[0;1]$ dans $\mathbb{R}$. Pour tout $x \in [0,1]$ on a

$$P(F(X_i) \geq x) = \mathbb{P}(X_i \geq F^{-1}(x)) = F(F^{-1}(x)) = x.$$

   Donc $F(X_i)$ suit une loi uniforme sur $[0,1]$ et on en déduit la valeur de son espérance et de sa variance.

2. Montrez que $nF_n(x) \sim B(n, F(x))$ et en déduire l'espérance et la variance de $nF_n(x)$.
   On a $\mathbf{1}_{X_j \leq x}$ qui suit une loi de Bernoulli de paramètre $F(x) = \mathbb{P}(\mathbf{1}_{X_j \leq x} = 1)$ donc $nF_n(x) = \sum_{i=1}^n \mathbf{1}_{X_j \leq x} \sim B(n, F(x))$ par indépendance des X_i d'où $\mathbb{E}(nF_n(x)) = nF(x)$ et $\operatorname{Var}(nF_n(x)) = nF(x)(1 - F(x))$.

3. En déduire que $\mathbb{E}\widehat{T}_n = \dfrac{1}{6n}$.
   En appliquant Fubini, on obtient

$$\begin{aligned} \mathbb{E}\widehat{T}_n &= \int_{\mathbb{R}} \mathbb{E}(F_n(x) - F(x))^2 \, dF(x) = \frac{1}{n} \int_{\mathbb{R}} F(x)(1 - F(x)) \, dF(x) \\ &= \frac{1}{n} \int_0^1 t(1-t) \, dt = \frac{1}{6n} \end{aligned}$$

4. Montrez que $\widehat{T}_n^\star = \dfrac{1}{n} \displaystyle\sum_{i=1}^n \left(\frac{i}{n} - F(X_{(i)}) \right)^2$.
   On remarque que l'expression de $\widehat{T}_n^\star$ est indépendante de l'ordre dans lesquels les X_i sont sommés. Donc

$$\widehat{T}_n^* = \frac{1}{n} \sum_{i=1}^n (F_n(X_{(i)}) - F(X_{(i)}))^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{i}{n} - F(X_{(i)}) \right)^2.$$

5. En admettant que

$$n\widehat{T}_n = \frac{1}{12n} + \sum_{k=1}^n \left(\frac{2k-1}{2n} - F(X_{(k)}) \right)^2$$

   montrer que

$$\widehat{T}_n - \widehat{T}_n^\star = -\frac{1}{2n} - \frac{1}{6n^2} + \frac{1}{n^2} \sum_{i=1}^n F(X_i).$$

On calcule

$$\begin{aligned}
\widehat{T}_n - \widehat{T}_n^\star &= -\frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^n \left(\left(\frac{i}{n} - F(X_{(i)}) \right)^2 - \left(\frac{2i-1}{2n} - F(X_{(i)}) \right)^2 \right) \\
&= -\frac{1}{12n^2} + \frac{1}{2n^2}\sum_{i=1}^n \left(\frac{4i-1}{2n} - 2F(X_{(i)}) \right) \\
&= -\frac{1}{12n^2} + \frac{1}{2n^2}\sum_{i=1}^n \frac{4i-1}{2n} - \frac{1}{n^2}\sum_{i=1}^n F(X_{(i)}) \\
&= -\frac{1}{12n^2} + \frac{1}{2n^2}\sum_{i=1}^n \frac{4i-1}{2n} - \frac{1}{n^2}\sum_{i=1}^n F(X_i) \\
&= \frac{1}{2n} + \frac{1}{6n^2} - \frac{1}{n^2}\sum_{i=1}^n F(X_i)
\end{aligned}$$

6. En déduire que $\lim_{n\to\infty} n^{3/2}\mathbb{E}(\widehat{T}_n^\star - \widehat{T}_n) = 0$ et que $\mathrm{Var}\,(\widehat{T}_n^\star - \widehat{T}_n) = \dfrac{1}{12n^3}$.
Le fait qu'une variable uniforme comme $F(X_i)$ ait pour moyenne $1/2$ et pour variance $1/12$ implique que

$$\widehat{T}_n - \widehat{T}_n^\star = \frac{1}{n^2}\sum_{i=1}^n (F(X_i) - \mathbb{E}F(X_i)) - \frac{1}{6n^2}$$

d'où

$$\mathbb{E}(\widehat{T}_n^\star - \widehat{T}_n) = \frac{1}{2n} + \frac{1}{6n^2} - \frac{1}{2n} = \frac{1}{6n^2} \text{ et } \mathrm{Var}\,(\widehat{T}_n^\star - \widehat{T}_n) = \frac{1}{n^3}\mathrm{Var}\,F(X_1) = \frac{1}{12n^3}$$

7. Montrer aussi que $n^{3/2}(\widehat{T}_n^\star - \widehat{T}_n) \to \mathcal{N}\Big(0, \dfrac{1}{12}\Big)$. Conclure.
L'identité précédente

$$\widehat{T}_n - \widehat{T}_n^\star = \frac{1}{n^2}\sum_{i=1}^n (F(X_i) - \mathbb{E}F(X_i)) - \frac{1}{6n^2}$$

et le théorème de limite centrale implique le résultat. On en déduit que l'erreur d'approximation est de l'ordre de $n^{-3/2}$.

8. (Facultatif) Montrer que

$$n\widehat{T}_n = \frac{1}{12n} + \sum_{k=1}^n \Big(\frac{2k-1}{2n} - F(X_{(k)}) \Big)^2.$$