PROJET 14

Les statistiques d'ordre pour construire des estimateurs

sujet proposé par Katia Meziani

meziani@ceremade.dauphine.fr

Certaines questions théoriques sont facultatives, leur résolution apportera des points bonus.

1 Statistiques d'ordre

Soit $X_1, X_2, \cdots, X_n$ des variables aléatoires réelles (v.a.r.) définies sur un même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, indépendantes et de même loi absolument continue par rapport à la mesure de Lebesgue de densité f (de fonction de répartition noté F). Pour tout ω dans Ω, on peut ordonner les réels $X_1, \cdots, X_n$ sous la forme

$$X_{(1)}(\omega) \leq \cdots \leq X_{(i)}(\omega) \leq \cdots \leq X_{(n)}(\omega).$$

L'application

$$X_{(i)} \, : \, \omega \in \Omega \longmapsto X_{(i)}(\omega)$$

ainsi définie pour chaque i est une v.a.r. dite i-ème statistique d'ordre.

T1. Calculer la fonction de densité et la fonction de répartition de $X_{(n)} = \sup\left(X_1, \cdots, X_n\right)$ en fonction de f et F.

T2. Calculer la fonction de densité et la fonction de répartition de $X_{(1)} = \inf\left(X_1, \cdots, X_n\right)$ en fonction de f et F.

S1.

– Écrire une fonction qui calcule $X_{(n)}$ pour n tirage suivant une loi exponentielle de paramètre a.

– Puis pour $a = 2$ et pour les valeurs $n = (10, 10000)$, faire $N = 500$ répétitions et tracer l'histogramme des $X_{(n)}$ sur lequel on ajoutera la courbe de la densité théorique calculée précédemment.

S2.

– Écrire une fonction qui calcule $X_{(1)}$ pour n tirage suivant une loi exponentielle de paramètre a.

– Puis pour $a = 2$ et pour les valeurs $n = (10, 100, 1000)$, faire $N = 500$ répétitions et tracer l'histogramme des $X_{(n)}$ sur lequel on ajoutera la courbe de la densité théorique calculée précédemment. Que constatez-vous ?

T3. Calculer la densité du couple $(X_{(1)}, X_{(n)})$ que l'on exprimera en fonction de f et F.

T4. Déduire de la question T3 que la fonction de densité f_V de l'étendue $V = X_{(n)} - X_{(1)}$ est donnée par la formule:

$$f_V(v) = n(n-1) \int_{\mathbb{R}^+} f(u) f(u+v) \big(F(u+v) - F(u)\big)^{n-2} du.$$

T5. Calculer la fonction de répartition de $X_{(k)}$. *Indication:* introduire la v.a. $N_y = \sum_{i=1}^{n} 1_{X_i \leq y}$ puis comparer les événements $\{N_y \geq k\}$ et $\{X_{(k)} \leq y\}$.

T6. En déduire que le densité $f_{X_{(k)}}$ est donnée par la formule:

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} f(x) \big(F(x)\big)^{k-1} \big(1 - F(x)\big)^{n-k}.$$

S3.

– Écrire une fonction qui calcule $X_{(k)}$ pour n tirage suivant une loi exponentielle de paramètre a.

– Puis pour $a = 2$ et les valeurs $(k, n) = (75, 100)$, faire $N = 500$ répétitions et tracer l'histogramme des $X_{(k)}$ sur lequel on ajoutera la courbe de la densité théorique calculée précédemment. Faire de même avec $(k, n) = (100, 1000)$ et $(k, n) = (100, 10000)$.

– Que constatez-vous ?

T7. (*Question facultative*) Montrer que si $\mathbb{E}(X)$ existe alors $\mathbb{E}(X_{(k)})$ aussi.

T8. (*Question facultative*) Calculer la densité du vecteur $(X_{(1)}(\omega), \cdots, X_{(n)}(\omega))$. *Indication :* On pourra calculer

$$\mathbb{P}\big((X_{(1)}(\omega), \cdots, X_{(n)}(\omega)) \in B\big)$$

pour tout borélien B de $\mathcal{B}_{\mathbb{R}^n}$.

2 Comparaison d'estimateurs dans un modèle de loi uniforme

On note $\mathcal{U}([0, a])$ la loi uniforme sur l'intervalle $[0, a]$. On considère le modèle (paramétrique) uniforme

$$\big\{\mathcal{U}([0, \theta]) \ : \ \theta > 0\big\}.$$

On considère un échantillon $X_1, \cdots, X_n$ et on note $X_{(1)} \leq \cdots \leq X_{(i)} \leq \cdots \leq X_{(n)}$ les statistiques d'ordre. On note $\overline{X}$ la moyenne empirique. On propose d'étudier les estimateurs suivants

$$\begin{aligned}
\widehat{\theta}_1 &= X_{(n)} \\
\widehat{\theta}_2 &= \frac{n+1}{n} X_{(n)} \\
\widehat{\theta}_3 &= X_{(1)} + X_{(n)} \\
\widehat{\theta}_4 &= 2\overline{X}
\end{aligned}$$

S4.

– Émettre une conjecture sur la pertinence/l'idée à la base de la proposition de l'estimateur $\widehat{\theta}_1$.

– Écrire une fonction qui calcule $\widehat{\theta}_1 = X_{(n)}$ pour n tirages suivant une loi uniforme $\mathcal{U}([0, \theta])$.

– Puis pour $\theta = 30$ et les valeurs $n = (100, 1000, 100000)$, faire $N = 500$ répétitions et tracer l'histogramme des $X_{(n)}$ sur lequel on ajoutera la courbe de la densité théorique.

– Commenter.

S5.

– Faire une conjecture sur la pertinence de $\widehat{\theta}_2$.

– Écrire une fonction qui calcule $\widehat{\theta}_2$ pour n tirages suivant une loi uniforme $\mathcal{U}([0, \theta])$.

– Puis pour $\theta = 30$ et les valeurs $n = (100, 10000)$, faire $N = 500$ répétitions et tracer l'histogramme des $\widehat{\theta}_2$ sur lequel on ajoutera la courbe de la densité théorique que l'on calculera.

– Commenter.

S6.

– Comparer par simulations la distance qui sépare $X_{(n)}$ de θ à celle qui sépare 0 de $X_{(1)}$, on choisira $\theta = 30$, $N = 500$ et $n = (100, 1000, 10000)$

– Conjecturer brièvement sur la pertinence/l'idée à la base de la proposition de l'estimateur $\widehat{\theta}_3$.

– Puis pour $\theta = 30$ et les valeurs $n = (100, 1000, 100000)$, faire $N = 500$ répétitions et tracer l'histogramme des $\widehat{\theta}_3$

– Commenter.

S7.

– Quel est l'intuition à la base de l'estimateur $\widehat{\theta}_4$?

– Proposer des simulations pour valider votre intuition.

– Puis pour $\theta = 30$ et les valeurs $n = (100, 1000, 100000)$, faire $N = 500$ répétitions et tracer l'histogramme des $\widehat{\theta}_4$.

– Commenter.

T9. Étudier la convergence (aussi appelée consistance) des 4 estimateurs $\widehat{\theta}_1$, $\widehat{\theta}_2$, $\widehat{\theta}_3$, $\widehat{\theta}_4$.

S8. Écrire un programme (graphiques) qui confirme vos résultats théoriques sur la consistance de ces 4 estimateurs. On prendra $\theta = 30$. On affichera l'évolution de la valeur de l'estimateur lorsque n varie:

– Sur un même graphique pour les 4 estimateurs, pour n allant jusqu'à $n = 100$.

– Sur des graphiques différents pour chaque estimateur pour n allant jusqu'à $n = 10000$.

Sur chaque graphique on tracera la droite $y = \theta$.

T10. Étudier les biais des 4 estimateurs $\widehat{\theta}_1$, $\widehat{\theta}_2$, $\widehat{\theta}_3$, $\widehat{\theta}_4$.

T11. Le biais d'un estimateur n'est pas un bon indicateur de la précision d'un estimateur. en effet être bon en moyenne ne garantie pas que l'on estime correctement le paramètre inconnu θ. Il faut étudier son risque. On prendra ici le risque quadratique d'un estimateur $\widehat{\theta}$, défini comme

$$\theta \longmapsto R(\theta, \widehat{\theta}) = \mathbb{E}_\theta[(\theta - \widehat{\theta})^2].$$

Montrer que ce risque peut être décomposé comme la somme de 2 termes (le biais et la variance)

$$R(\theta, \widehat{\theta}) = b^2(\theta, \widehat{\theta}) + \mathbb{V}ar_\theta(\widehat{\theta}),$$

où $b^2(\theta, \widehat{\theta})$ sera explicité (*c.f.* Définition 8.14 du poly MAP361).

T12. Montrer les formules suivantes pour les risques quadratiques des estimateurs $\widehat{\theta}_1$, $\widehat{\theta}_2$ et $\widehat{\theta}_4$:

$$\begin{aligned} R(\theta, \widehat{\theta}_1) &= \frac{2\theta^2}{(n+2)(n+1)}, \\ R(\theta, \widehat{\theta}_2) &= \frac{\theta^2}{n(n+2)}, \\ R(\theta, \widehat{\theta}_4) &= \frac{\theta^2}{3n}. \end{aligned}$$

T13. (*Question facultative*) Montrer que le risque quadratique de $\widehat{\theta}_3$ est donnée par la formule

$$R(\theta, \widehat{\theta}_3) = \frac{2\theta^2}{(n+2)(n+1)}.$$

T14. On dit qu'un estimateur est admissible si il n'existe pas d'estimateurs qui admet un risque strictement plus petit que lui pour θ. Comparer/discuter la qualité de ces estimateurs. Quel est le meilleur des 4? Lesquels sont (possiblement) admissibles, non admissibles ?

S9. Tracer sur un même graphique les courbes $(n, R(\theta, \widehat{\theta}_i))$ pour les 4 estimateurs, on n prendra là $\theta = 30$ et on fera varier n de 0 à 100, puis de 0 à 10. Commenter.

T15. Définissons pour $\alpha > 0$ fixé, l'estimateur $\widehat{\theta}_\alpha = \alpha X_{(n)} = \alpha\widehat{\theta}_1$. Calculer $R(\theta, \widehat{\theta}_\alpha)$.

S10.
– Écrire une fonction qui calcule $\widehat{\theta}_\alpha$ pour n tirages suivant une loi uniforme $\mathcal{U}([0, \theta])$.
– Puis pour $\theta = 30$ et les valeurs $n = (10, 100)$, tracer $(\alpha, \widehat{\theta}_\alpha)$ sur l'intervalle $\alpha \in [0, 2]$.
– Commenter.

T16. (*Question facultative*) Montrer que la valeur α_{opt} qui minimise le risque $R(\theta, \widehat{\theta}_\alpha)$ vaut

$$\alpha_{opt} = \frac{n+2}{n+1}$$

et qu'alors

$$R(\theta, \widehat{\theta}_{\alpha_{opt}}) = \frac{\theta^2}{(n+1)^2}.$$

S11.
– Tracer sur un même graphique les courbes $(n, R(\theta, \widehat{\theta}_i))$ pour les 5 estimateurs précédents, on prendra $\theta = 30$ et on fera varier n de 0 à 10. Commenter.
– Tracer en fonction de n la différence entre le risque de $\widehat{\theta}_{\alpha_{opt}}$ et $\widehat{\theta}_{i^*}$ (le meilleur des 4 premiers estimateurs). On prendra $\theta = 30$ et on fera varier n de 0 à 10. Commenter.