Université de Pau et des Pays de l'Adour | **Semestre printemps 2023–2024**
L3 Math | Introduction à la Statistique

CC 1 - Estimation statistique Correction

Question de cours (3pts)

Soit $(\mathcal{H}^n, (Q_\theta)_{\theta\in\Theta})$ un modèle statistique. On note $\hat{\theta}_n$ un estimateur du véritable paramètre inconnu θ^*.

1. (1pt) Qu'est-ce qu'un estimateur sans biais ?

Ne pas oublier le $\forall\theta \in \Theta$ dans la définition !

2. (2pts) Pour le modèle $(\{0,1\}^n, (\mathcal{B}(p)^{\otimes n})_{p\in(0,1)})$, montrez que l'estimateur suivant est sans biais

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Par linéarité.

Exercice 1 Méthode des moments (3pts)

Soit $(\mathbb{R}_+^n, (\mathbb{P}^{\otimes n}_{a\in\mathbb{R}, b\in\mathbb{R}, c>0}))$ un modèle statistique vérifiant pour $X \sim \mathbb{P}_{a,b,c}$

$$\mathbb{E}[X] = a + b - c^2;\ \mathrm{Var}(X) = a - b + c^2;\ \mathbb{E}[X^4] = c^2.$$

Déterminer un estimateur de (a^*, b^*, c^*) à l'aide de la méthode des moments.

Il fallait résoudre le système

$$\begin{cases} \bar{X}_n = \hat{a}_{MM} + \hat{b}_{MM} - \hat{c}^2_{MM} \\ S_n^2 = \hat{a}_{MM} - \hat{b}_{MM} + \hat{c}^2_{MM} \\ U_4(n) = \hat{c}^2_{MM} \end{cases}$$

Exercice 2 Estimation du paramètre de la loi de Weibull (15pts)

Soit X une variable aléatoire réelle suivant la loi de Weibull de densité : pour tout $x > 0$

$$f_\beta(x) = \frac{e^{-\frac{\sqrt{x}}{\beta}}}{2\beta\sqrt{x}}$$

où β est un paramètre strictement positif inconnu.

Partie 1 : Estimation classique du paramètre

1. (1pt) Écrire le modèle statistique dans le cas où les observations sont indépendantes.

On travaille dans la suite sur le modèle $((\mathbb{R}_+^*)^n, (\mathbb{P}_\beta^{\otimes n})_{\beta>0})$ avec $\mathbb{P}_\beta$ pour $\beta > 0$ donné par l'énoncé.

2. (2pts) Montrer que pour $\beta > 0$, on a

$$\mathbb{E}_\beta[X] = 2\beta^2.$$

En déduire un estimateur $\hat{\beta}_{MM}$ par la méthode des moments.

Soit $\beta > 0$, on a

$$\begin{aligned}
\mathbb{E}_\beta[X] &= \int_0^{+\infty} x f_\beta(x)\ \mathrm{d}x = \int_0^{+\infty} \frac{\sqrt{x}}{2\beta} \mathrm{e}^{-\sqrt{x}/\beta}\ \mathrm{d}x \\
&= \beta^2 \int_0^{+\infty} \frac{1}{2\sqrt{x}\beta} \left(\frac{\sqrt{x}}{\beta}\right)^2 \mathrm{e}^{-\sqrt{x}/\beta}\ \mathrm{d}x \text{ et donc avec un changement de variable avec } u = \sqrt{x}/\beta, \\
&= \beta^2 \int_0^{+\infty} u^2 \mathrm{e}^{-u}\ \mathrm{d}u.
\end{aligned}$$

Il reste à effectuer deux IPP. Bien vérifier l'existence de chaque termes lorsqu'on effectue une IPP avec une intégrale généralisée ! Maintenant pour trouver l'estimateur des moments, il suffit de résoudre

$$\bar{X}_n = 2\hat{\beta}_{MM}^2.$$

3. (1.5pt) Soit $x_1, \dots, x_n \in \mathbb{R}_+^*$, déterminer la vraisemblance du modèle et montrer que pour tout $\beta \in \mathbb{R}_*^+$:

$$\frac{\partial \log L_n}{\partial \beta}(x_1, \dots, x_n; \beta) = -\frac{n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n \sqrt{x_i}.$$

On rappelle qu'on utilise la notation log pour le logarithme népérien.

Pour $\beta > 0$ et $x_1, \dots, x_n \in \mathbb{R}_+^*$, on a

$$L_n(x_1, \dots, x_n; \beta) = \prod_{i=1}^n f_\beta(x_i) = \frac{1}{\beta^n} \prod_{i=1}^n \frac{\mathrm{e}^{-\sqrt{x_i}/\beta}}{2\sqrt{x_i}}.$$

Et donc la log-vraisemblance vaut

$$\log L_n(x_1, \dots, x_n; \beta) = -n \log \beta - \sum_{i=1}^n \log(2\sqrt{x_i}) - \frac{\sum_{i=1}^n \sqrt{x_i}}{\beta}.$$

D'où en dérivant par rapport à β,

$$\frac{\partial \log L_n}{\partial \beta}(x_1, \dots, x_n; \beta) = -\frac{n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n \sqrt{x_i}.$$

4. (1.5pts) Montrer que l'estimateur du maximum de vraisemblance du modèle est

$$\hat{\beta}_{MV} = \frac{1}{n} \sum_{i=1}^n \sqrt{X_i}.$$

On justifiera bien le fait qu'il s'agit d'un maximum.

Pour $x_1, x \dots, x_n \in \mathbb{R}_+^*$, on peut vérifier par croissance comparée que

$$\lim_{\beta \to 0} \log L_n(x_1, \dots, x_n; \beta) = \lim_{\beta \to +\infty} \log L_n(x_1, \dots, x_n; \beta) = -\infty,$$

donc la fonction possède un maximum global. Comme la log-vraisemblance possède un unique point critique, ce point critique est forcément le maximum global.

5. (3pts) Déterminer le biais de l'EMV, puis sans calculer explicitement le biais, montrer que l'estimateur $\hat{\beta}_{MM}$ est biaisé.

Par un calcul similaire à la question 2, on peut montrer que $\hat{\beta}_{MV}$ est sans biais. Pour l'estimateur des moments, comme $u \mapsto \sqrt{u}$ est strictement concave (donc son opposée est strictement convexe), on a par l'inégalité de Jensen,

$$\mathbb{E}_\beta[\hat{\beta}_{MM}] = \mathbb{E}_\beta\left[\sqrt{\frac{\bar{X}_n}{2}}\right] < \sqrt{\mathbb{E}_\beta[\bar{X}_n]/2} = \sqrt{\mathbb{E}_\beta[X]/2} = \beta.$$

6. (1pt) Montrer que les estimateurs $\hat{\beta}_{MM}$ et $\hat{\beta}_{MV}$ sont fortement consistants.
Comme nos observations $(X_i)_i$ et $(\sqrt{X_i})_i$ sont i.i.d. et possèdent un moment d'ordre 1, on peut appliquer la LGN, pour $\beta > 0$, $\mathbb{P}_\beta$-p.s.

$$\bar{X}_n \to \mathbb{E}_\beta[X_1] = 2\beta^2 \text{ et } \hat{\beta}_{MV} = \frac{1}{n}\sum_{i=1}^n \sqrt{X_i} \to \mathbb{E}_\beta[\sqrt{X_1}] = \beta.$$

Donc l'EMV est fortement consistant et comme $\hat{\beta}_{MM} = \sqrt{\bar{X}_n/2}$, on obtient aussi sa forte consistance par la continuité des opérations.
7. (1pt) Déterminer le risque de l'EMV.
Comme l'EMV est sans biais, on a d'après la décomposition biais-variance, pour $\beta > 0$,

$$\mathcal{R}(\beta; \hat{\beta}_{MV}) = \mathrm{Var}_\beta(\hat{\beta}_{MV}) = \mathrm{Var}_\beta\left(\frac{1}{n}\sum_{i=1}^n \sqrt{X_i}\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_\beta(\sqrt{X_i}),$$

par l'indépendance. Et comme nos variables sont identiquement distribuées, il suffit de calculer la variance de $\sqrt{X_1}$,

$$\mathrm{Var}_\beta(\sqrt{X_1}) = \mathbb{E}_\beta[X_1] - \mathbb{E}_\beta[\sqrt{X_1}]^2 = 2\beta^2 - \beta^2 = \beta^2.$$

Chaque espérance avait déjà été calculée, d'où

$$\mathcal{R}(\beta; \hat{\beta}_{MV}) = \frac{\beta^2}{n}.$$

Partie 2 : Estimation à l'aide des statistiques d'ordre

8. (1.5pts) Déterminer la fonction de répartition F_X de la variable X. En déduire que

$$F_X^{-1}(0.5) = (\beta \log 2)^2.$$

Interpréter le terme $F_X^{-1}(0.5)$.
Soit $\beta > 0$, pour $t \leqslant 0$, on a $F_X(t) = 0$ car notre variable est à valeurs dans $\mathbb{R}_+^*$, et pour $t > 0$,

$$F_X(t) = \int_0^t f_\beta(x)\,\mathrm{d}x = 1 - \mathrm{e}^{-\sqrt{t}/\beta}.$$

En résolvant $\frac{1}{2} = 1 - \mathrm{e}^{-\sqrt{t}/\beta}$, on retrouve la valeur de l'énoncé. Il s'agit de la médiane de notre variable. X a autant de chance d'être au dessus de cette valeur que en dessous.
9. (1.5pts) On range de manière croissante notre échantillon, et notons

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)},$$

le nouveau réarrangement. Montrer que presque sûrement, ces inégalités sont strictes.

Pour $\beta > 0$, on a

$$\begin{aligned}
Q_\beta(\text{une de ces inégalités est une égalité}) &= Q_\beta(\exists i \neq j,\ X_{(i)} = X_{(j)}) \\
&= Q_\beta(\exists i \neq j,\ X_i = X_j) \\
&\leqslant \sum_{i \neq j} Q_\beta(X_i = X_j) = \sum_{i\neq j} \int_{\{x=y\}} f_\beta(x) f_\beta(y)\ \mathrm{d}x\mathrm{d}y = 0,
\end{aligned}$$

par l'indépendance entre X_i et X_j et le fait que $\{x = y\}$ est de mesure nulle pour la mesure de Lebesgue (il s'agit d'une droite dans le plan). On pourrait aussi dire que $X_i - X_j$ est à densité car X_i et X_j sont à densité et **indépendantes**. C'est assez important de revenir à X_i et non garder $X_{(i)}$ car la loi du réarrangement est un peu compliquée.

10. (1pt) À l'aide du réarrangement, déterminer un nouvel estimateur de β.

On peut estimer la vraie médiane par la médiane de notre échantillon, et comme la médiane vaut $(\beta \log 2)^2$, on peut proposer comme estimateur de β,

$$\hat{\beta}_{med} = \frac{\sqrt{X_{(\lfloor n/2 \rfloor)}}}{\log 2}.$$

Rappel du cours de probabilité : On pourra utiliser sans preuve les résultats suivants :

1. Si $X \sim \mathcal{B}(p)$ pour $p \in [0, 1]$ alors $\mathbb{E}[X] = p$ et $\mathrm{Var}(X) = p(1-p)$.
2. **La Loi des Grands Nombres** : Soit X une variable aléatoire possèdant un moment d'ordre 1 et soit $(X_n)_{n\in\mathbb{N}}$ une suite i.i.d. de variables aléatoires de même loi X, on a
$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} \mathbb{E}[X].$$
3. **Le Théorème Central Limite** : Si de plus, X possède un moment d'ordre 2 alors
$$\sqrt{n}\frac{\bar{X}_n - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} \xrightarrow{loi} Z \sim \mathcal{N}(0, 1).$$
4. **Inégalité de Jensen** : Soit X une variable aléatoire non constante ayant un moment d'ordre 1, pour φ une fonction strictement convexe, si $\varphi(X)$ possède un moment d'ordre 1 alors
$$\varphi(\mathbb{E}[X]) < \mathbb{E}[\varphi(X)].$$