Statistique Mathématique **Université de Rennes – M1**

CC3 2025/2026 – Durée 3h

Les documents et appareils électroniques (calculatrice, téléphone, ordinateur, ...) sont interdits. ***Toutes les réponses doivent être justifiées****.*

Exercice 1 - Longueurs de tâches

La longueur (en temps) d'une tâche attribuée à un ordinateur par un usager donné est modélisée par une loi de Pareto $\mathcal{P}(\alpha)$, définie par la densité f_α

$$f_\alpha(x) = (\alpha - 1)x^{-\alpha}\mathbb{1}_{x>1},$$

pour un $\alpha > 1$ inconnu. On collecte $X_1, \dots, X_n$, n longueurs de tâches i.i.d. (pour le même usager), pour $n \geq 3$.

1. Montrer que f_α définit bien une densité.
2. Donner le modèle de cette expérience.
3. Mettre ce modèle sous une forme exponentielle canonique (on n'oubliera pas d'en spécifier tous les éléments).
4. Est-il possible d'utiliser la méthode des moments pour construire un estimateur ?
5. Construire un estimateur $\hat{\alpha}$ de α par la méthode du maximum de vraisemblance.
6. Des tâches informatiques anormalement longues ($\alpha \leq 5$) peuvent être dues à un dysfonctionnement. On souhaite construire un test pour détecter ces éventuels dysfonctionnements. En partant du principe qu'il est plus grave de laisser passer un dysfonctionnement que de détecter un dysfonctionnement là où il n'y en a pas, quelles sont les hypothèses du test associé ?
7. Montrer que le rapport de vraisemblance s'exprime sous la forme $g(\overline{\log(X)})$, où g est décroissante. *Indication : on pourra commencer par exprimer le rapport de vraisemblance dans les cas où $\hat{\alpha} \leq 5$, $\hat{\alpha} > 5$, puis recoller les morceaux.*
8. Montrer que, si $Z \sim \mathcal{P}(\alpha)$, $\log(Z) \sim \mathcal{E}(\alpha - 1)$ (loi exponentielle). En déduire que $\overline{\log(X)}$ suit la loi $\gamma(n(\alpha - 1), n)$ (loi γ, on rappelle la densité d'une loi $\gamma(a, b)$: $t^{a-1}e^{-bt}b^a\mathbb{1}_{t>0}/\Gamma(a)$, pour $a, b > 0$).
9. Montrer que si $0 < b_1 \leq b_2$, pour tout $a > 0$ on a $\gamma(a, b_2) \preccurlyeq \gamma(a, b_1)$ (domination stochastique).
10. Construire le test du rapport de vraisemblance de niveau 5% pour ces hypothèses.
11. (*) Montrer que ce test est le plus puissant parmi les tests de niveau 5%.

Solution 1 -

1. f_α est mesurable, positive, et
$$\int_1^{+\infty} (\alpha-1)x^{-\alpha}dx = \left[-x^{-\alpha+1}\right]_1^{+\infty} = 1.$$
2. Le modèle est $(]1;+\infty[^n, \mathcal{B}(]1;+\infty[^n), (f_\alpha d\lambda)_{\alpha>1}^{\otimes n})$.
3. On commence par choisir la mesure dominante $\mu = \mathcal{L}\mathbb{1}_{]1;+\infty[}$ (Lebesgue restreinte). La densité de $X_1, \dots, X_n$ par rapport à $\mu^{\otimes n}$ s'écrit, pour $x_1, \dots, x_n > 1$,
$$\begin{aligned} f_\alpha(x_{1:n}) &= (\alpha-1)^n \left(\prod_{i=1}^n x_i\right)^{-\alpha} \\ &= \exp\left(n\left(\log(\alpha-1) - \alpha\overline{\log(x)}\right)\right), \end{aligned}$$
avec $\overline{\log(x)} = n^{-1}\sum_{i=1}^n \log(x_i)$. Pour tout $\alpha > 1$ on a $\int_{]1;+\infty[^n} (\prod_{i=1}^n x_i)^{-\alpha} = (\alpha-1)^{-n} < +\infty$, le domaine est donc $]1, +\infty[$. Le paramètre canonique est α, et la statistique exhaustive $T(X_{1:n}) = -n\overline{\log(X)}$. La fonction de partition est alors $Z(\alpha) = -n\log(\alpha-1)$.
4. Comme $E_\alpha|X_1| = +\infty$ pour $\alpha \leq 2$, la méthode des moments est donc hors de propos.
5. Pour un modèle exponentiel on peut se contenter de maximiser la log-vraisemblance. Pour $x_1, \dots, x_n > 1$, on a
$$\begin{aligned} \ell_{x_{1:n}}(\alpha) &= n\left(\log(\alpha-1) - \alpha\overline{\log(x)}\right), \\ \ell'_{x_{1:n}}(\alpha) &= n\left(\frac{1}{\alpha-1} - \overline{\log(x)}\right). \end{aligned}$$
$\ell_{x_{1:n}}$ étant concave, un point d'annulation de la dérivée est nécessairement un maximum. Or $\ell'_{x_{1:n}}(\alpha) = 0 \Leftrightarrow \alpha = 1 + \frac{1}{\overline{\log(x)}}$. On en déduit
$$\hat{\alpha} = 1 + \frac{1}{\overline{\log(X)}}.$$
6. On souhaite contrôler l'erreur la plus grave par le niveau du test. Le niveau majorant la probabilité de détecter H_1 lorsque l'on est sous H_0, les hypothèses de test sont alors
$$\begin{cases} H_0 & : \quad \alpha \leq 5 \\ H_1 & : \quad \alpha > 5. \end{cases}$$

7. Regardons les log-vraisemblances des hypothèses. Pour $x_1, \dots, x_n > 1$, en utilisant la stricte concavité de $\ell_{x_{1:n}}$,

$$\begin{aligned}\ell_1(x_{1:n}) = \sup_{\alpha>5} \ell_{x_{1:n}}(\alpha) &= n\left(\log(\hat{\alpha}-1) - \hat{\alpha}\overline{\log(x)}\right)\mathbb{1}_{\hat{\alpha}>5} + n(\log(4) - 5\overline{\log(x)})\mathbb{1}_{\hat{\alpha}\leq 5} \\ &= n\left(-\log(\overline{\log(x)}) - (\overline{\log(x)}+1)\right)\mathbb{1}_{\hat{\alpha}>5} + n(\log(4) - 5\overline{\log(x)})\mathbb{1}_{\hat{\alpha}\leq 5}.\end{aligned}$$

Pour la log-vraisemblance de H_0, on utilise en plus la continuité de $\ell_{x_{1:n}}$ en 5

$$\ell_0(x_{1:n}) = \sup_{\alpha\leq 5} \ell_{x_{1:n}}(\alpha) = n\left(-\log(\overline{\log(x)}) - (\overline{\log(x)}+1)\right)\mathbb{1}_{\hat{\alpha}\leq 5} + n(\log(4) - 5\overline{\log(x)})\mathbb{1}_{\hat{\alpha}>5}.$$

En désignant par ℓ_{RV} le log du rapport de vraisemblance, on a

$$\begin{aligned}\ell_{RV}(x_{1:n}) &= \ell_1(x_{1:n}) - \ell_0(x_{1:n}) \\ &= n\left(-\log(\overline{\log(x)}) + 4\overline{\log(x)} - 1 - \log(4)\right)(\mathbb{1}_{\hat{\alpha}>5} - \mathbb{1}_{\hat{\alpha}\leq 5}).\end{aligned}$$

Par ailleurs, $\hat{\alpha} \leq 5 \Leftrightarrow \overline{\log(x)} \geq 1/4$. Si on pose $h : u > 0 \mapsto -\log(u)+4u-1-\log(4)$, on a $h'(u) \leq 0 \Leftrightarrow u \leq 1/4$. On en déduit que h est décroissante sur $]0; 1/4]$ et croissante sur $[1/4; +\infty[$, donc que $v \mapsto n(-\log(v)+4v-1-\log(4))(\mathbb{1}_{v<1/4} - \mathbb{1}_{v\geq 1/4})$ est décroissante en v. On en déduit enfin que ℓ_{RV} et donc RV sont deux fonctions decroissantes de $\overline{\log(x)}$.

8. Soit $t > 0$, et $Z \sim \mathcal{P}(\alpha)$,

$$\begin{aligned}\mathbb{P}\left(\log(Z) > t\right) &= \mathbb{P}(Z > e^t) \\ &= \int_{e^t}^{+\infty} (\alpha-1)x^{-\alpha} = e^t(\alpha-1) = \mathbb{P}\left(\mathcal{E}(\alpha-1) > t\right).\end{aligned}$$

Si $Z_1 \sim \gamma(a,b)$ et $Z_2 \sim \gamma(a',b)$, avec $Z_1 \perp\!\!\!\perp Z_2$, on a $Z_1 + Z_2 \sim \gamma(a+a',b)$. On rappelle aussi que $\lambda\gamma(a,b) \sim \gamma(a, b/\lambda)$. Enfin, $\mathcal{E}(\lambda) \sim \gamma(1,\lambda)$. On déduit de tout cela que $\overline{\log(X)} \sim n^{-1}\gamma(n, \alpha-1) \sim \gamma(n, n(\alpha-1))$.

9. Soit $Z \sim \gamma(a,1)$, on a

$$\gamma(a,b_2) \sim \frac{1}{b_2}Z \leq \frac{1}{b_1}Z \sim \gamma(a,b_1),$$

d'où la domination stochastique.

10. D'après la question 7, la forme du test du rapport de vraisemblance, noté T_{RV} est

$$T_{RV}(X_{1:n}) = \mathbb{1}_{\overline{\log(X)}\leq t}.$$

Il n'y a plus qu'à calibrer t en résolvant l'inéquation

$$\sup_{\alpha\leq 5} P_\alpha(\overline{\log(X)} \leq t) \leq 5\% \quad \Leftrightarrow \quad \sup_{\alpha\leq 5} \mathbb{P}(\gamma(n, n(\alpha-1)) \leq t) \leq 5\%,$$

d'après la question 8. Or $\gamma(n, n(\alpha-1) \sim \frac{1}{\alpha-1}\gamma(n,n)$. On en déduit que si $\alpha_1 \leq \alpha_2$, $\gamma(n, n(\alpha_2-1) \preccurlyeq \gamma(n, n(\alpha_1-1)$ (domination stochastique), et donc que

$$\sup_{\alpha \leq 5} \mathbb{P}(\gamma(n, n(\alpha-1)) \leq t) = \mathbb{P}(\gamma(n,4n) \leq t).$$

En choisisant t comme le quantile d'ordre 5% d'une loi $\gamma(n,4n)$, on a notre test du rapport de vraisemblance (au niveau 5%).

11. Commençons par remarquer que notre test T_{RV} est aussi test du rapport de vraisemblance pout toutes les hypothèses de type $H_0 : \alpha = 5$, $H_1 : \alpha = \alpha_1$, avec $\alpha_1 > 5$, et est de niveau exact 5%. Soit T un test de niveau 5% pour les hypothèses $H_0 : \alpha \leq 5$, $H_1 : \alpha > 5$. T est en particulier de niveau 5% pour $H_0 : \alpha = 5$, $H_1 : \alpha > 5$. Soit $\alpha_1 > 5$. T est aussi de niveau 5% pour $H_0 : \alpha = 5$, $H_1 : \alpha = \alpha_1$. Le Theorème de Neyman Pearson donne alors

$$P_{\alpha_1}(T=1) \leq P_{\alpha_1}(T_{RV}=1).$$

L'inégalité du dessus valant pour tout α_1, on en déduit que T_{RV} est uniformément plus puissant que T.

Exercice 2 - Chant des baleines

Un océanographe souhaite déterminer la fréquence d'émission vocale des baleines, notée $\theta > 0$. Pour ce faire il étudie n enregistrements. La fréquence d'émission observée sur le i-ème enregistrement, notée X_i, se modélise par $X_i = \theta(1+\varepsilon_i)$, les $(\varepsilon_i)_{i=1,\dots,n}$ étant supposées i.i.d. de loi $\mathcal{N}(0,1)$.

1. Donner un modèle pour cette expérience. Est-on dans le cadre d'un modèle linéaire Gaussien du type de ceux vus en cours ? Ce modèle est-il dominé ?
2. Donner un estimateur de θ par la méthode des moments (on le notera $\hat{\theta}_1$). Calculer son risque quadratique.
3. Construire un intervalle de niveau de confiance **non-asymptotique** 92% sur θ, basé sur $\hat{\theta}_1$. On le notera I_1.
4. Montrer que l'information de Fisher (pour une observation), notée $I(\theta)$ est bien définie, et la calculer. Au besoin on pourra admettre que $I(\theta) = -E_\theta(\partial^2/\partial\theta^2 \ell_\theta(X))$ (cette formule ne peut être utilisée comme définition), ou que $\mathbb{E}(\varepsilon_1^4) = 3$.
5. $\hat{\theta}_1$ est-il efficace ?
6. En notant $\alpha = 1/\theta$, montrer que pour $x_{1:n} \in \mathbb{R}^n$, la log-vraisemblance $\ell_{x_{1:n}}(\theta)$ vérifie

$$\ell'_{x_{1:n}}(\theta) = n\alpha\left(\alpha^2\overline{x^2} - \alpha\bar{x} - 1\right),$$

avec $\overline{x^2} = n^{-1}\sum_{i=1}^n x_i^2$ et $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$.

7. En déduire l'expression de l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ (on veillera bien à justifier toutes les étapes) :

$$\hat{\theta}_{MV} = \frac{2\overline{X^2}}{\bar{X} + \sqrt{\bar{X}^2 + 4\overline{X^2}}}.$$

8. Montrer que $\hat{\theta}_{MV}$ est consistant.
9. On souhaite prouver que $\theta < 1$. Poser les hypothèses du test associé.
10. Donner la forme de la zone de rejet d'un test basé sur $\hat{\theta}_{MV}$, pour les hypothèses de la question précédente.
11. Pour $\theta > 0$, on note Q_θ la loi de $\hat{\theta}_{EMV}$ lorsque le paramètre inconnu vaut θ, et Z_θ une variable aléatoire de loi Q_θ. Montrer que $Z_\theta \sim \theta Z_1$.
12. En se basant sur les trois dernière questions, construire un test de niveau 1% (on admettra que l'on connaît la loi de $\hat{\theta}_{MV}$ lorsque $\theta = 1$).
13. Montrer que $\hat{\theta}_{MV}$ est asymptotiquement efficace. *Indication : on pourra admettre que* $g : (x, y) \mapsto 2y/(x + \sqrt{x^2 + 4y})$ *est dérivable en* $(\theta, 2\theta^2)$*, de gradient* $v(\theta) = (-1/3, 1/(3\theta))^T$.
14. En déduir un intervalle de niveau de confiance asymptotique 92% sur θ basé sur $\hat{\theta}_{MV}$.

Solution 2 -

1. Le modèle est $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\theta\mathcal{N}(1,1))_{\theta>0}^{\otimes n}$. On peut le mettre sous la forme $X = \mathbb{1}_n\theta + \theta\varepsilon$, c'est donc bien un modèle linéaire Gaussien du type de ceux vu en cours. Il est dominé par $\mathcal{L}_n$ (Lebesgue n-dimensionnelle).
2. Comme $E_\theta|X_1| \leq \theta\mathbb{E}(1 + |\mathcal{N}(0,1)|) < +\infty$, X_1 adment une espérance, qui vaut $E_\theta(X_1) = \theta$. La méthode des moments suggère donc de prendre

$$\hat{\theta}_1 = \bar{X}_n.$$

Pour le risque quadratique,

$$\begin{aligned} E_\theta((\hat{\theta}_1 - \theta)^2) &= \mathrm{Var}_\theta(\bar{X}_n) \qquad \text{(estimateur non biaisé)} \\ &= \frac{\mathrm{Var}_\theta(X_1)}{n} \\ &= \frac{\mathrm{Var}(\theta(1+\varepsilon_1))}{n} = \frac{\theta^2}{n}. \end{aligned}$$

3. Les X_i étant Gaussiennes et indépendantes, $\hat{\theta}_1 = \bar{X}_n$ l'est aussi, et on a $\hat{\theta}_1 \sim \mathcal{N}(\theta, \frac{\theta^2}{n})$. Soit q le quantile d'ordre 96% d'une loi $\mathcal{N}(0,1)$. On a alors, pour tout $\theta > 0$,

$$P_\theta\left(\sqrt{n}\left|\frac{\hat{\theta}_1 - \theta}{\theta}\right| \leq q\right) = 92\%.$$

Or,

$$\sqrt{n}\left|\frac{\hat{\theta}_1 - \theta}{\theta}\right| \leq q \quad \Leftrightarrow \quad \frac{\hat{\theta}_1}{1 + q/\sqrt{n}} \leq \theta \leq \frac{\hat{\theta}_1}{1 - q/\sqrt{n}}.$$

On en déduit que $I_1 = [\hat{\theta}_1/(1 + q/\sqrt{n}), \hat{\theta}_1/(1 - q/\sqrt{n})]$ est un intervalle de niveau de confiance 92%.

4. En notant f_θ la densité de X_1, on a que $f_\theta(x) > 0$ pour tout $x \in \mathbb{R}$. On peut donc définir, pour tout $x \in \mathbb{R}$,

$$\ell_\theta(x) = -(1/2)\log(2\pi) - \log(\theta) - (x - \theta)^2/(2\theta^2).$$

On a alors

$$\dot{\ell}_\theta(x) = -\frac{1}{\theta} - \frac{x}{\theta^2} + \frac{x^2}{\theta^3}.$$

Comme $E_\theta X^4 < +\infty$, $E_\theta \dot{\ell}_\theta(X)^2 < +\infty$, et l'information de Fisher est bien définie. En calculant

$$\ddot{\ell}_\theta(x) = \frac{1}{\theta^2} + \frac{2x}{\theta^3} - \frac{3x^2}{\theta^4},$$

on a

$$I(\theta) = -E_\theta(\ddot{\ell}_\theta(X)) = -\left(\frac{1}{\theta^2} + \frac{2}{\theta^2} - \frac{6}{\theta^2}\right) = \frac{3}{\theta^2}.$$

5. Commençons par remarquer que $\hat{\theta}_1$ est sans biais. Comme $E_\theta((\hat{\theta}_1 - \theta)^2) > (nI(\theta))^{-1}$, on déduit que $\hat{\theta}_1$ n'est pas efficace.

6. Soit $x_{1:n} \in \mathbb{R}^n$. Comme $f_\theta(x_{1:n}) > 0$ pour tout $\theta > 0$, la log-vraisemblance est bien définie par

$$\ell_{x_{1:n}}(\theta) = -n\log(\theta) - (n/2)\log(2\pi) - \frac{1}{2}\sum_{i=1}^n x_i^2/\theta^2 + \sum_{i=1}^n x_i/\theta - n/2.$$

On en déduit

$$\begin{aligned}\ell'_{x_{1:n}}(\theta) &= n\left(\overline{x^2}\theta^{-3} - \bar{x}\theta^{-2} - \theta^{-1}\right) \\ &= n\alpha\left(\alpha^2\overline{x^2} - \alpha\bar{x} - 1\right).\end{aligned}$$

7. Comme, pour tout $x \in \mathbb{R}^n$ pour tout $\theta > 0$ $f_\theta(x_{1:n}) > 0$, on peut se contenter de maximiser la log-vraisemblance. Soit P le polynôme $X^2\overline{x^2} - X\bar{x} - 1$. D'après ce qui précède $\ell'_{x_{1:n}}(\theta)$ a même signe que $P(\theta^{-1})$. P est à discriminant positif, sa première racine est négative, et la seconde vaut

$$\hat{\alpha} = \frac{\bar{x} + \sqrt{\bar{x}^2 + 4\overline{x^2}}}{2\overline{x^2}}.$$

On en déduit le tableau de variation suivant pour $\ell_{x_{1:n}}$:

θ	0		$\hat{\alpha}^{-1}$	
$\ell'_{x_{1:n}}(\theta)$	$+$	$+$	0	$-$

On en déduit que

$$\hat{\theta}_{MV} = \frac{2\overline{X^2}}{\bar{X} + \sqrt{\bar{X}^2 + 4\overline{X^2}}}.$$

8. On a $E_\theta|X_1|, E_\theta X_1^2 < +\infty$, la loi des grands nombres donne alors $\bar{X} \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$ et $\overline{X^2} \xrightarrow[n\to+\infty]{\mathbb{P}} E_\theta(X_1^2) = 2\theta^2$. Comme $g : (x, y) \mapsto 2y/(x + \sqrt{x^2 + 4y}$ est continue en $(\theta, 2\theta^2)$, on a que

$$\hat{\theta}_{MV} = g(\bar{X}, \overline{X^2}) \xrightarrow[n\to+\infty]{\mathbb{P}} g(\theta, 2\theta^2) = \frac{4\theta^2}{4\theta} = \theta,$$

$\hat{\theta}_{MV}$ est donc bien consistant.

9. On souhaite prouver $\theta < 1$, on le met donc dans l'hypothèse alternative. Les hypothèses sont donc $H_0 : \theta \geq 1$, $H_1 : \theta < 1$.
10. Sous H_1, on s'attend à ce que $\hat{\theta}_{MV}$ soit petit. La zone de rejet est donc de la forme $[0, t]$, où t est à calibrer.
11. Soit $Y_{1:n}$ un n-échantillon de $P_1^{\otimes n}$, alors $\theta Y_{1:n} \sim P_\theta^{\otimes n}$. On en déduit que

$$\hat{\theta}_{MV} = \frac{2\theta^2\overline{Y^2}}{\theta(\bar{Y} + \sqrt{\bar{Y}^2 + 4\overline{Y^2}})} = \theta\frac{\overline{Y^2}}{\bar{Y} + \sqrt{\bar{Y}^2 + 4\overline{Y^2}}} := \theta Z_1,$$

où par définition $Z_1 \sim Q_1$ et $\hat{\theta}_{MV} \sim Q_\theta$.

12. De la question précédente on déduit que $\theta_1 \leq \theta_2 \Rightarrow Q_{\theta_1} \preccurlyeq Q_{\theta_2}$ (ordre stochastique). On peut alors calibrer le test :

$$\begin{aligned}\sup_{\theta\geq 1} P_\theta(\hat{\theta}_{MV} \leq t) &= \sup_{\theta\geq 1} Q_\theta([0, t]) \\ &= Q_1([0, t]).\end{aligned}$$

En choisissant pour t le quantile d'ordre 1% de Q_1, on a notre test.

13. Comme g est différentiable en $(\theta, 2\theta^2)$, la méthode Δ donne

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) = \sqrt{n}(g(\bar{X}, \overline{X^2}) - g(\theta, 2\theta^2) \rightsquigarrow \langle v(\theta), \mathcal{N}(0, \Sigma(\theta)) \rangle ,$$

avec

$$\Sigma(\theta) = \begin{pmatrix} \mathrm{Var}_\theta(X_1) & \mathrm{Cov}_\theta(X_1, X_1^2) \\ \mathrm{Cov}_\theta(X_1, X_1^2) & \mathrm{Var}_\theta(X_1^2) \end{pmatrix}.$$

Une succession de calculs donne $E_\theta(X_1^2) = 2\theta^2$, $E_\theta X_1^3 = 4\theta^2$ et $E_\theta(X_1^4) = 10\theta^4$, ce dont on déduit $\mathrm{Var}_\theta(X_1) = \theta^2$, $\mathrm{Cov}_\theta(X_1, X_1^2) = 2\theta^3$, $\mathrm{Var}_\theta(X_1^2) = 6\theta^4$. On en déduit que

$$v(\theta)^T \Sigma(\theta) v(\theta) = \frac{\theta^2}{3} = I(\theta)^{-1},$$

et donc que

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) \rightsquigarrow \mathcal{N}(0, I(\theta)^{-1}).$$

$\hat{\theta}_{MV}$ est donc asymptotiquement efficace.

14. C'est cadeau. De l'efficacité asymptotique et de la consistance de $\hat{\theta}_{MV}$ on déduit en utilisant le Lemme de Slutsky :

$$\frac{\sqrt{3n}}{\hat{\theta}_{MV}} \left(\hat{\theta}_{MV} - \theta \right) \rightsquigarrow \mathcal{N}(0, 1),$$

et donc on déduit que $[\hat{\theta}_{MV}(1 \pm \sqrt{3}q/\sqrt{n})]$ est un intervalle de niveau de confiance asymptotique 92%, où q est le quantile d'ordre 96% d'une loi $\mathcal{N}(0, 1)$.

Exercice 3 - Un test

On suppose que l'on a accès à un échantillon $X_1, \ldots, X_n$ i.i.d. de loi P_θ définie par

$$\forall k \in \mathbb{N}^* \quad P_\theta(\{k\theta\}) = \frac{1}{k(k+1)},$$

où $\theta > 0$ est inconnu. On souhaite tester les hypothèses suivantes

$$\begin{cases} H_0 & : \quad \theta \geq 1, \\ H_1 & : \quad \theta < 1. \end{cases}$$

1. Montrer que, pour tout $\theta > 0$, P_θ définit bien une loi de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
2. Montrer que, pour tout $\theta > 0$, si $X \sim P_\theta$, alors $X/\theta \sim P_1$.
3. Donner le modèle de cette expérience.

4. Peut-on construire un test du rapport de vraisemblance (on rappelle qu'une justification est obligatoire) ?
5. Pour $i = 1, \dots, n$, on note $Y_i = 1/X_i$. Montrer que $E_\theta(Y_1) = \frac{\mu_1}{\theta}$, avec $\mu_1 = (\pi^2/6 - 1)$. *(On rappelle que $\sum_{k\geq 1} k^{-2} = \pi^2/6$)*.
6. On souhaite bâtir un test pour les hypothèses H_0 et H_1 sur la statistique $\bar{Y}_n$. Donner la forme de la région de rejet associée (on ne cherchera pas à la calculer exactement pour le moment).
7. Énoncer l'inégalité de Hoeffding. En déduire que, si $\theta = 1$, pour tout $\alpha \in]0; 1[$,

$$P_1\left(\bar{Y}_n \geq \mu_1 + \sqrt{\frac{\log(1/\alpha)}{2n}}\right) \leq \alpha.$$

8. En déduire un test de niveau α pour ces deux hypothèses (il ne suffit pas de le donner, il faut montrer qu'il est bien de niveau α).

Solution 3 -

1. P_θ est bien σ-additive et vérifie de manière évident $P_\theta(\emptyset) = 0$. Il reste à vérifier $P_\theta(\mathbb{R}) = 1$. On a

$$\begin{aligned} P_\theta(\mathbb{R}) &= \sum_{k\geq 1} \frac{1}{k(k+1)} \\ &= \lim_{N\to+\infty} \sum_{k=1}^{N} \frac{1}{k(k+1)} \qquad (1/(k(k+1)) \text{ sommable}) \\ &= \lim_{N\to+\infty} \sum_{k=1}^{N} \left(\frac{1}{k} - \frac{1}{k+1}\right) \\ &= \lim_{N\to+\infty} 1 - \frac{1}{N+1} = 1. \end{aligned}$$

2. Comme X prend ses valeurs dans $\theta\mathbb{N}^*$, X/θ prend ses valeurs dans $\mathbb{N}^*$. Pour $\theta > 0$ et $k \in \mathbb{N}^*$, on a

$$P_\theta\left(X/\theta = k\right) = P_\theta(X = k\theta) = \frac{1}{k(k+1)} = P_1(X = k).$$

   Donc $X/\theta \sim P_1$.
3. Modèle : $\left(]0; +\infty[^n, \mathcal{B}(]0; +\infty[^n), (P_\theta^{\otimes n})_{\theta>0}\right)$.

4. On va montrer que ce modèle ne peut être dominé par une mesure σ-finie. Soit μ_n une telle mesure dominante (que l'on peut supposer de probabilité), et μ_1 sa première composante ($\mu_1(A) = \mu_n(A\times]0;+\infty[^{n-1})$). On a alors, pour tout $\theta > 0$, $P_\theta \ll \mu_1$, et donc, pour tout $(k,\theta) \in \mathbb{N}^*\times]0;+\infty[$, $\mu_1(\{k\theta\}) > 0$. On en déduit alors que, pour tout $x > 0$, $\mu_1(\{x\}) > 0$, μ_1 ne peut donc pas être σ-finie. Le modèle n'étant pas dominé, on ne peut définir de vraisemblance, et encore moins construire un test du rapport de vraisemblance.

5. Commençons par remarquer que, sous P_θ, $Y_1 \in]0;1/\theta[$, et donc que Y_1 est intégrable sous P_θ, pour tout $\theta > 0$. On a alors, pour $\theta > 0$,

$$\begin{aligned} E_\theta(Y_1) &= E_\theta(1/X_1) = \theta^{-1}E_1(X_1) \\ &= \theta^{-1}\sum_{k\geq 1}\frac{1}{k^2(k+1)} \\ &= \theta^{-1}\left(\sum_{k\geq 1}\frac{1}{k^2} - \sum_{k\geq 1}\frac{1}{k(k+1)}\right) \qquad \text{(les deux séries de droite sont sommables)} \\ &= \theta^{-1}\left(\pi^2/6 - 1\right). \end{aligned}$$

6. En se donnant $U_1,\dots,U_n$ i.i.d. de loi $1/X_1$, où $X_1 \sim P_1$, pour $\theta > 0$ on a

$$\bar{Y}_n \sim \frac{1}{\theta}\bar{U}_n.$$

Sous H_1, on s'attend donc à ce que $\bar{Y}_n$ soit grand (on rappelle que $\bar{U}_n > 0$) . La région de rejet associée est de type $[t_\alpha;+\infty[$, où t_α reste à calibrer.

7. Inégalité de Hoeffding : si $Z_1,\dots,Z_n$ sont indépendantes, et $Z_i \in [a_i;b_i]$ presque sûrement, pour tout $i \in [\![1;n]\!]$, alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n Z_i - \mathbb{E}(Z_i) \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2v}\right),$$

avec $v = \frac{1}{4}\sum_{i=1}^n (b_i - a_i)^2$. Pour $\theta = 1$, $Y_i \in]0;1[$ et $E_1(Y_i) = \mu_1$, pour tout $i \in [\![1;n]\!]$. On en déduit

$$P_1\left(\bar{Y}_n \geq \mu_1 + \sqrt{\frac{\log(1/\alpha)}{2n}}\right) \leq \exp\left(-2n\left[\sqrt{\frac{\log(1/\alpha)}{2n}}\right]^2\right) = \alpha.$$

8. Au vu de la question $6-$, le test sera de la forme

$$T(X_{1:n}) = \mathbb{1}_{\bar{Y}_n \geq t_\alpha},$$

où t_α est à calibrer de telle sorte que

$$\forall \theta \geq 1 \quad P_\theta\left(\bar{Y}_n \geq t_\alpha\right) \leq \alpha.$$

On pose $t_\alpha = \mu_1 + \sqrt{\frac{\log(1/\alpha)}{2n}}$. Pour tout $\theta \geq 1$ on a

$$\begin{aligned} P_\theta\left(\bar{Y}_n \geq t_\alpha\right) &= P_1\left(\frac{\bar{Y}}{\theta} \geq t_\alpha\right) \\ &\leq P_1\left(\bar{Y} \geq t_\alpha\right) \qquad \text{(car } \theta \text{ et } \bar{Y} \text{ sont positifs)} \\ &\leq \alpha, \end{aligned}$$

d'après la question précédente. On a donc bien un test de niveau α pour le t_α choisi.