Statistique Mathématique **Université de Rennes – M1**

Exam 2023/2024 – Durée 3h

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

Exercice 1 - Loi de Pareto

Soit $\theta > 0$, et X une variable aléatoire de densité f_θ définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{2\theta^2}{x^3} \mathbb{1}_{x \geq \theta}.$$

On considère $X_1, \dots, X_n$ des variables aléatoires indépendantes et de même loi que X.

1. Donner le modèle de cette expérience.
2. Calculer $E_\theta(X)$. En déduire un estimateur de θ. Est-il consistant ? Le théorème central limite peut-il s'appliquer ?
3. Calculer $E_\theta\left(X^{-1}\right)$. En déduire un nouvel estimateur de θ, montrer qu'il est consistant et déterminer son comportement asymptotique.
4. On note $X_{(1)} = \min_{1 \leq i \leq n} X_i$.
   (a) Donner la fonction de répartition de $X_{(1)}$.
   (b) En déduire la loi limite de $n\left(X_{(1)} - \theta\right)$.
   (c) Pour $\alpha \in (0,1)$, construire un intervalle de confiance non-asymptotique pour θ de niveau $1-\alpha$. On le cherchera de la forme $I = [c_\alpha X_{(1)}, X_{(1)}]$.
   (d) Soit $\theta_0 > 0$. Construire un test T_n de niveau $\alpha \in (0,1)$ pour tester

   $$H_0 : \theta \geq \theta_0 \qquad \text{contre} \qquad H_1 : \theta < \theta_0 \,,$$

   i.e. un test T_n tel que $\sup_{\theta \geq \theta_0} P_\theta(T_n = 1) = \alpha$.
   (e) Montrer que ce test T_n est consistant, i.e. que pour tout $\theta < \theta_0$, on a $P_\theta(T_n = 1) \underset{n \to +\infty}{\longrightarrow} 1$.

Solution 1 -

1. On a

   $$E_\theta(X) = 2\theta^2 \int_\theta^{+\infty} \frac{1}{x^2} dx = 2\theta \,.$$

   On peut donc proposer l'estimateur $\frac{1}{2n}\sum_{i=1}^n X_i$, qui est consistant par la loi des grands nombres. Cependant, le TCL ne s'applique pas car

   $$E_\theta(X^2) = 2\theta^2 \int_\theta^{+\infty} \frac{1}{x} dx = +\infty \,.$$

2. On a
$$E_\theta\left(\frac{1}{X}\right) = 2\theta^2 \int_\theta^{+\infty} \frac{1}{x^4} dx = \frac{2}{3\theta},$$
ce qui conduit à considérer l'estimateur
$$\hat{\theta}_n = \frac{2}{3}\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{X_i}\right)^{-1}.$$
Par la loi forte des grands nombres, on a
$$\frac{1}{n}\sum_{i=1}^n \frac{1}{X_i} \xrightarrow[n\to+\infty]{p.s.} \frac{2}{3\theta},$$
et comme la fonction $g : x \longmapsto \frac{2}{3x}$ est continue en $2/3\theta \neq 0$, on a par le théorème de continuité $\hat{\theta}_n \xrightarrow[n\to+\infty]{p.s.} \theta$. De plus, on a $\mathrm{Var}_\theta\left(\frac{1}{X}\right) = E_\theta\left[\frac{1}{X^2}\right] - E_\theta\left[\frac{1}{X}\right]^2 = \frac{1}{2\theta^2} - \frac{4}{9\theta^2} = \frac{1}{18\theta^2}$. Le TCL nous donne donc
$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{X_i} - \frac{2}{3\theta}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{18\theta^2}\right).$$
De plus, g est dérivable sur $]0, +\infty[$, et pour tout $x > 0$, on a $g'(x) = \frac{-2}{3x^2}$. On en déduit
$$g'\left(\frac{2}{3\theta}\right) = \frac{-3}{2}\theta^2.$$
On obtient donc, par la méthode Delta,
$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\theta^2}{8}\right).$$

3. (a) Pour tout $x < \theta$, on a $P_\theta(X_{(1)} \leq x) = 0$, et pour tout $x \geq \theta$,
$$P_\theta\left(X_{(1)} \leq x\right) = 1 - P_\theta\left(X_{(1)} > x\right) = 1 - P_\theta\left(X > x\right)^n = 1 - \frac{\theta^{2n}}{x^{2n}}.$$

(b) Pour tout $x \in \mathbb{R}$, on a
$$\begin{aligned} P_\theta\left(n(X_{(1)} - \theta) \leq x\right) &= \mathbb{1}_{x\geq 0}\left(1 - \left(\frac{\theta}{\theta + \frac{x}{n}}\right)^{2n}\right) \\ &= \mathbb{1}_{x\geq 0}\left(1 - \frac{1}{\left(1 + \frac{x}{\theta n}\right)^{2n}}\right) \\ &\xrightarrow[n\to+\infty]{} \mathbb{1}_{x\geq 0}\left(1 - e^{-\frac{2x}{\theta}}\right). \end{aligned}$$
Ainsi
$$n\left(X_{(1)} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{E}\left(\frac{2}{\theta}\right).$$

(c) Cherchons un intervalle de la forme

$$I = \left[X_{(1)}c_\alpha\,,\, X_{(1)}\right]\,,$$

avec $c_\alpha \in]0,1[$. On a

$$P_\theta\left(X_{(1)}c_\alpha \geq \theta\right) = \alpha \Leftrightarrow c_\alpha^{2n} = \alpha \Leftrightarrow c_\alpha = \alpha^{1/2n}.$$

On obtient donc l'intervalle $I = \left[X_{(1)}\alpha^{1/2n}\,,\, X_{(1)}\right]$.

(d) On cherche un test de la forme $T_n = \mathbb{1}_{X_{(1)}<c_\alpha\theta_0}$ avec $c_\alpha > 1$ (pour $c_\alpha \leq 1$, le test ne se trompe jamais sur H_0, i.e. la taille vaut 0). On a

$$\sup_{\theta\geq\theta_0} P_\theta\left(X_{(1)} < c_\alpha\theta_0\right) = \sup_{\theta\geq\theta_0}\left(1 - \left(\frac{\theta}{c_\alpha\theta_0}\right)^{2n}\right) = 1 - c_\alpha^{-2n}\,.$$

On prend donc $c_\alpha = (1-\alpha)^{-\frac{1}{2n}}$, et l'on obtient le test $T_n = \mathbb{1}_{X_{(1)}<(1-\alpha)^{-\frac{1}{2n}}\theta_0}$, qui est bien de taille α par construction.

(e) Soit $\theta < \theta_0$. On a

$$\begin{aligned} P_\theta(T_n = 0) &= P_\theta\left(X_{(1)} \geq (1-\alpha)^{-\frac{1}{2n}}\theta_0\right) \\ &= P_\theta\left(X \geq (1-\alpha)^{-\frac{1}{2n}}\theta_0\right)^n \\ &= (1-\alpha)\left(\frac{\theta}{\theta_0}\right)^{2n} \underset{n\to+\infty}{\longrightarrow} 0\,. \end{aligned}$$

La suite de test (T_n) est bien consistante.

Exercice 2 - Choc des savoirs - Bilan géographique

Suite à la mise en place du dispositif permettant de découper une classe de terminale en plusieurs niveaux (disons 4 : pas bon, passable, bon, très bon), un sociologue soupçonneux cherche à déterminer s'il existe un lien entre tronçonnage en niveaux et situation de l'établissement.

Pour ce faire, il collecte sur l'année 2024, pour les n élèves de classe Terminale de Rennes, l'appartenance à un des 4 groupe de niveau ainsi que la situation de leur établissement (on supposera qu'il y a 12 lycées différents à Rennes).

Construire un test pour déterminer s'il y a dépendance entre appartenance à un groupe de niveau et établissement. On n'oubliera pas de poser un modèle, les hypothèses de tests correspondantes, la statistique de test, sa loi sous H_0 et le test final correspondant. On pourra aussi noter $n_{i,j}$ le nombre d'élèves de niveau $i \in [\![1,4]\!]$ dans l'établissement $j \in [\![1,12]\!]$.

Solution 2 -

On suppose les élèves indépendants, et on note Z_ℓ la variable aléatoire sur $[\![1,4]\!] \times [\![1,12]\!]$ qui donne le niveau X_ℓ et l'établissement Y_ℓ de l'èlève ℓ.

Le modèle est alors $(([\![1,4]\!] \times [\![1,12]\!])^n, \mathcal{P}(([\![1,4]\!] \times [\![1,12]\!])^n), (P_p^{\otimes n})_{p\in[0,1]^{48}})$, où P_p est la loi sur $[\![1,4]\!] \times [\![1,12]\!]$ définie par

$$P_p(\{(i,j)\}) = P_p(\{X=i\} \cap \{Y=j\}) = p_{i,j}.$$

Il s'agit alors de déterminer si, pour tous i,j, $P_p(\{X=i\} \cap \{Y=j\}) = P_p(X=i)P_p(Y=j)$, soit encore si $p_{i,j} = p_{i,.}p_{.,j}$, avec $p_{i,.} = \sum_j p_{i,j}$, $p_{.,j} = \sum_i p_{i,j}$. On va effectuer cela à l'aide d'un test du chi-deux d'indépendance, d'hypothèses

$$H_0 : X \perp\!\!\!\perp Y,$$
$$H_1 : X \text{ et } Y \text{ ne sont pas indépendantes.}$$

En notant $n_{i,j}$ le nombre d'élèves de niveau i dans l'établissement j, $\hat{p}_{i,.} = (\sum_j n_{i,j})/n$ la proportion totale observée d'élèves de niveau i, et $\hat{p}_{.,j} = (\sum_i n_{i,j})/n$ la proportion totale observées d'élèves dans l'établissement j, on définit l'effectif théorique

$$n_{t,i,j} = n\hat{p}_{i,.}\hat{p}_{.,j},$$

qui correspond à l'effectif auquel on s'attendrait si H_0 était vraie. La statistique de Pearson correspondant à ce test s'écrit alors

$$S = \sum_{i=1}^{4}\sum_{j=1}^{12} \frac{(n_{i,j} - n_{t,i,j})^2}{n_{t,i,j}}.$$

Sous H_0, on a que $S \rightsquigarrow \chi^2((4-1)(12-1)) \sim \chi^2(33)$ lorsque $n \to +\infty$. Dans la suite on supposera cette approximation vraie (loisible si pour tout i,j, $n_{t,i,j} \geq 5$). En notant q le quantile d'ordre $1-\alpha$ d'une telle loi, on a que

$$T = \mathbb{1}_{S \geq q}$$

est un test de niveau (asymptotique) α pour statuer sur l'éventuelle inéquité géographique d'une telle réforme.

Exercice 3 - Financement de campagne

Préliminaire technique

Soient $a,b > 0$. La loi Bêta de paramètres (a,b), notée $\beta(a,b)$, est la loi sur $]0,1[$ de densité

$$p_{a,b}(u) = \frac{u^{a-1}(1-u)^{b-1}}{B(a,b)} \mathbb{1}_{]0,1[}(u),$$

où $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ (et Γ est la fonction Gamma usuelle).

1. Montrer que si $X \sim \beta(a,b)$, alors $\mathbb{E}(X) = \frac{a}{a+b}$ et $\mathbb{E}(X^2) = \frac{a(a+1)}{(a+b)(a+b+1)}$. *Indication : d'après l'énoncé, on sait que $\int_0^1 u^{a-1}(1-u)^{b-1} = B(a,b)$, pour tous $a,b>0$.*

Élections européennes 2024

Le candidat A aux élections européennes 2024 demande un prêt à sa banque pour financer sa campagne. Du point de vue du banquier, le score d'un candidat à une élection peut être modélisé par une loi $\beta(r, 1-r)$, où $r \in]0,1[$, et les règles de la banque sont strictes : on ne finance que les candidats dont le r dépasse 5% (seuil de remboursement des frais de campagne).

Pour statuer sur le dossier, le banquier a accès aux scores du candidat aux élections précédentes $X_1, \dots, X_n$, supposés i.i.d. suivant la même loi $\beta(r, 1-r)$.

2. Donner le modèle de cette expérience.
3. Proposer un estimateur de r basé sur la méthode des moments (on le notera $\hat{r}_1$).
4. Montrer que le risque quadratique de $\hat{r}_1$ vaut $r(1-r)/2$.
5. Déterminer le comportement asymptotique de $\hat{r}_1$.
6. En déduire un intervalle de niveau de confiance **asymptotique** 98% de type $[\hat{r}_{1,-}, +\infty[$.
7. Le banquier veut tester si r dépasse 5% ou non, et veut être sûr de son coup en finançant le candidat. Quelles sont les hypothèses de son test ?
8. Déduire de ce qui précède un test d'erreur **asymptotique** 2% visant à déterminer si le candidat mérite son prêt ou non.
9. Montrer la consistance du test proposé.

Le banquier étant pointilleux par nature, il cherche à savoir si l'estimateur de r utilisé plus haut est optimal.

10. Le modèle est-il exponentiel ? Si oui donner au moins sa mesure dominante et sa statistique exhaustive.
11. On admet le résultat suivant : pour tout $r \in]0,1[$, $\Gamma(r)\Gamma(1-r) = \frac{\pi}{\sin(\pi r)}$. Trouver l'estimateur du maximum de vraisemblance pour r (on le notera $\hat{r}_2$).
12. Rappeler la définition de l'information de Fisher $I(r)$.
13. En admettant que, pour $n = 1$, $I(r) = \frac{\pi^2}{\sin^2(\pi r)}$, déterminer le comportement asymptotique de $\hat{r}_2$.
14. Montrer que $\hat{r}_2$ est **asymptotiquement** plus efficace que $\hat{r}_1$, et en déduire que, pour tout $r \in]0,1[$, $\frac{\sin^2(\pi r)}{\pi^2} \leq \frac{r(1-r)}{2}$.
15. **Bonus** : Montrer que $I(r) = -E_r(\ddot{\ell}_n(r))$ (espérance de la dérivée seconde en r). En déduire que $I(r) = \frac{\pi^2}{\sin^2(\pi r)}$.

Solution 3 -

1. Comme $a > -1$ et $b-1 > -1$, $\int_0^1 u^a(1-u)^{b-1}$ est finie. On a alors

$$\begin{aligned}\mathbb{E}(X) &= \int_0^1 \frac{u^a(1-u)^{b-1}}{B(a,b)}du \\ &= \frac{B(a+1,b)}{B(a,b)} = \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+1)} = \frac{a}{a+b}.\end{aligned}$$

   Comme $a+1 > -1$ et $b-1 > -1$, $\int_0^1 u^{a+1}(1-u)^{b-1}$ est finie. On a alors

$$\mathbb{E}(X^2) = \frac{B(a+2,b)}{B(a,b)} = \frac{\Gamma(a+2)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+2)} = \frac{a(a+1)}{(a+b)(a+b+1)}.$$

2. Modèle : $\left(]0,1[^n, \mathcal{B}(]0,1[^n), (\beta(r,1-r)^{\otimes n})_{r\in]0,1[}\right)$.
3. D'après la question 1, on a $E_r(X_1) = r$. Un estimateur par moments est donc

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

4. $\hat{r}_1$ étant sans biais, on a

$$E_r((\hat{r}_1 - r)^2) = \mathrm{Var}_r(\hat{r}_1) = \frac{\mathrm{Var}_r(X_1)}{n}.$$

   Par ailleurs,

$$\mathrm{Var}_r(X_1) = E_r(X_1)^2 - (E_r(X_1))^2 = \frac{r(r+1)}{2} - r^2 = \frac{r(1-r)}{2}.$$

5. Comme $E_r(X_1)^2 < +\infty$, le théorème central limite donne

$$\sqrt{n}\,(\hat{r}_1 - r) \rightsquigarrow \mathcal{N}\left(0, \frac{r(1-r)}{2}\right).$$

6. La question précédente donne

$$\mathbb{P}\left(r \geq \hat{r}_1 - \frac{\sqrt{(r(1-r)/2)}}{\sqrt{n}}q\right) \underset{n\to+\infty}{\longrightarrow} 98\%,$$

   où q est le quantile d'ordre 98% d'une loi $\mathcal{N}(0,1)$. Comme $r(1-r) \leq 1/4$, on en déduit que

$$I_1 = \left[\hat{r}_1 - \frac{1}{2\sqrt{2n}}q, +\infty\right[$$

   est un intervalle de niveau de confiance asymptotique 98%.

7. Si le banquier veut être sûr de son coup en finançant le candidat, il doit poser $H_1 : r > 5\%$, et donc $H_0 : r \leq 5\%$.

8. Regardons le test

$$T = \mathbb{1}_{\hat{r}_1 - \frac{1}{2\sqrt{2n}}q > 5\%}.$$

Supposons $r < 5\%$, d'après la question précédente on a

$$\begin{aligned} P_r\left(5\% < \hat{r}_1 - \frac{1}{2\sqrt{2n}}q\right) &\leq P_r\left(r < \hat{r}_1 - \frac{1}{2\sqrt{2n}}q\right) \\ &\leq P_r(r \notin I_1). \end{aligned}$$

On en déduit alors que $\limsup_n P_r(T = 1) \leq 2\%$, et donc que T est d'erreur asymptotique moins de 2%.

9. Supposons maintenant que $r > 5\%$. Pour n assez grand tel que $q/(2\sqrt{2n}) < (r - 5\%)/2$, on a

$$\begin{aligned} P_r(T = 1) &\geq P_r\left(\hat{r}_1 - r + (r - 5\%)/2 > 0\right) \\ &\underset{n\to+\infty}{\longrightarrow} 1, \end{aligned}$$

car $\hat{r}_1 \to r$ en probabilité (loi des grands nombres).

10. Regardons pour $n = 1$. Si on prend comme mesure dominante $\mu(dx) = x^{-1}\mathbb{1}_{]0,1[}x\lambda(dx)$, où λ est la mesure de Lebesgue sur $\mathbb{R}$, on a bien une mesure dominante σ-finie, et une loi $\beta(r, 1-r)$ admet pour densité

$$f_r(x) = \exp\left(r\log\left(\frac{x}{1-x}\right) - \log(B(r, 1-r))\right)$$

par rapport à cette mesure. On en déduit que le modèle pour $n = 1$ est exponentiel, de statistique exhaustive $T(X) = \log\left(\frac{X}{1-X}\right)$. Pour un n quelconque, le modèle reste exponentiel, avec pour mesure dominante $\mu^{\otimes n}$, et statistique exhaustive $\overline{\log\left(\frac{X}{1-X}\right)}$. $X \sim \beta(r, 1-r)$ étant une variable bornée, $E_r e^{tX}$ est finie pour tout $t \in \mathbb{R}$, le domaine est donc $]0, 1[$.

11. Soit $x_{1:n} \in]0,1[^n$, et notons $\ell_n(r) = nr\overline{\log\left(\frac{x}{1-x}\right)} - n\log(B(r, 1-r))$ la log-vraisemblance associée. D'après le cours, si l'équation $\dot{\ell}_n(r) = 0$ admet une solution dans $]0, 1[$, alors cette solution est le maximum de vraisemblance.
En utilisant $B(r, 1-r) = \frac{\pi}{\sin(\pi r)}$, on a

$$\dot{\ell}_n(r) = 0 \Leftrightarrow n\frac{\pi\cos(\pi r)}{\sin(\pi r)} + n\overline{\log\left(\frac{x}{1-x}\right)} = 0,$$

qui admet pour solution $\hat{r}_2 = \frac{1}{\pi}\cot^{-1}\left(\overline{\log\left(\frac{1-x}{x}\right)}\right) \in]0,1[$, où $\cot$ est la fonction cotangente. On en déduit

$$\hat{r}_2 = \frac{1}{\pi}\cot^{-1}\left(\overline{\log\left(\frac{1-X}{X}\right)}\right) \quad \text{p.s..}$$

12. Pour $n = 1$, l'information de Fisher est définie par

$$I(r) = E_r\left(\dot{\ell}_r(X)^2\right),$$

lorsque cette intégrale est bien définie (ce qui est toujours le cas dans les modèles exponentiels). L'information de Fisher pour $n \geq 1$ est donnée par $I_n(r) = nI(r)$.

13. Notre modèle étant exponentiel donc régulier, on a

$$\sqrt{n}\,(\hat{r}_2 - r) \rightsquigarrow \mathcal{N}(0, I(r)^{-1}) \sim \mathcal{N}\left(0, \frac{\sin^2(\pi r)}{\pi^2}\right).$$

14. Le modèle étant régulier et $\hat{r}_1$ non biaisé, la borne de Cramer-Rao s'applique : on a $n\mathrm{Var}_r(\hat{r}_1) \geq I(r)^{-1}$. On en déduit alors que $\hat{r}_2$ est asymptotiquement plus efficace que $\hat{r}_1$. Cela revient à comparer les variances asymptotiques renormalisées, soit

$$\frac{r(1-r)}{2} \geq \frac{\sin^2(\pi r)}{\pi^2}.$$

15. C'est un résultat général que dans les modèles suffisamment réguliers, $I(\theta) = -E_\theta(\ddot{\ell}(\theta))$ (non vu en cours). Prouvons le dans notre cadre d'un modèle exponentiel réel : $r \in]0,1[$, $n = 1$, $f_r(x) = \exp(rT(x) - \log(Z(r))$. On repart de

$$E_\mu(f_r(X)) = 1,$$

ce qui en dérivant (les interversions sont garanties par les conditions de régularité des modèles exponentiels) donne

$$0 = E_\mu(\partial_r f_r(X)) = E_r(\dot{\ell}(r)).$$

En dérivant encore une fois, on obtient

$$\begin{aligned} 0 &= \partial_r\left(E_\mu(\dot{\ell}(r) f_r)\right) \\ &= E_\mu\left(\ddot{\ell}(r) f_r\right) + E_\mu(\dot{\ell}(r)\partial_r f) \\ &= E_r(\ddot{\ell}(r)) + E_r(\dot{\ell}(r)^2). \end{aligned}$$

On en déduit $I(r) = E_r(\dot{\ell}(r)^2) = -E_r(\ddot{\ell}(r))$. Dans le cas $n = 1$, dans notre modèle on a

$$\dot{\ell}(r) = \frac{\pi\cos(\pi r)}{\sin(\pi r)} + n\log\left(\frac{x}{1-x}\right),$$

et donc

$$\ddot{\ell}(r) = \frac{-\pi^2 \sin^2(\pi r) - \pi^2 \cos^2(\pi r)}{\sin^2(\pi r)} = -\frac{\pi^2}{\sin^2(\pi r)},$$

ce dont on déduit $I(r) = \frac{\pi^2}{\sin^2(\pi r)}$.