PSSD **Université de Rennes – L3**

CC3 2023/2024 – Durée 1h30

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

N.B. : Le sujet comporte 2 pages.

Exercice 1 - Surbooking dans les avions

Pour remplir au maximum ses avions, une compagnie aérienne veut évaluer la probabilité d'absence d'un passager. Pour chaque vol, la compagnie vend K billets, où K est connu. On suppose que les K clients correspondants se comportent de manière indépendante, et que chaque client a une probabilité $p \in]0,1[$ inconnue d'être absent à l'embarquement. Pour faire son estimation, la compagnie collecte le nombre de passagers embarqués sur les n vols précédents.

1. En notant X_i le nombre de passagers présents sur le vol i, montrer que $X_i \sim \mathcal{B}(K,(1-p))$.
2. En déduire le modèle de cette expérience.
3. Donner la densité associée.
4. Montrer que l'estimateur du maximum de vraisemblance, noté $\hat{p}_n$, est défini par

$$\hat{p}_n = 1 - \frac{\bar{X}_n}{K},$$

   où $\bar{X}_n = (\sum_{i=1}^n X_i/n)$.
5. Montrer que $\hat{p}_n$ est consistant.
6. Montrer que $\hat{p}_n$ est asymptotiquement normal.
7. En déduire un intervalle de confiance de niveau **asymptotique** 95% sur p, de la forme $[s;1[$.
8. **Application** : la compagnie calcule $s = 0.02$. Au vu de cette estimation, quel est le nombre de billets qu'elle peut vendre pour qu'en moyenne ses avions soient totalement remplis, chaque avion comportant 196 places.

Solution 1 -

1. On note $Y_{i,j}$ la variable qui vaut 1 si le passager j du i-ème vol se présente, 0 sinon. D'après l'énoncé, on peut supposer les $(Y_{i,j})_{j=1,\dots,K}$ indépendantes. Par ailleurs, $Y_{i,j} \sim \mathcal{B}(1-p)$. Comme $X_i = \sum_{j=1}^K Y_{i,j}$, on en déduit $X_i \sim \mathcal{B}(K,(1-p))$.

2. La compagnie observe $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{B}(K, 1-p)$. Le modèle est donc

$$\left(\llbracket 0, K \rrbracket^n, (\mathcal{B}(K, 1-p)^{\otimes n})_{p \in]0,1[}\right).$$

3. Soit $x_1, \dots, x_n \in \llbracket 1, K \rrbracket$, la densité est donnée par

$$f_p(x_{1:n}) = \prod_{i=1}^n \binom{K}{x_i} (1-p)^{x_i} p^{K-x_i} = \left(\prod_{i=1}^n \binom{K}{x_i}\right) (1-p)^{\sum_{i=1}^n x_i} p^{nK - \sum_{i=1}^n x_i}.$$

4. Soit $x_1, \dots, x_n \in \llbracket 0, K \rrbracket$, et $p \in]0,1[$. La vraisemblance est donnée par

$$V_{x_{1:n}}(p) = f_p(x_{1:n}) = \left(\prod_{i=1}^n \binom{K}{x_i}\right) (1-p)^{\sum_{i=1}^n x_i} p^{nK - \sum_{i=1}^n x_i}.$$

Comme $V_{x_{1:n}}(p) > 0$ pour tout $p \in]0,1[$, on peut se contenter de maximiser la log-vraisemblance définie par

$$\ell_{x_{1:n}}(p) = \log(V_{x_{1:n}}(p)) = \log\left(\prod_{i=1}^n \binom{K}{x_i}\right) + S\log(1-p) + (nK - S)\log(p),$$

avec $S = \sum_{i=1}^n x_i$. Cette fonction étant concave en p, si $p \in]0,1[$ vérifie $\ell'_{x_{1:n}}(p) = 0$, ce sera nécessairement un maximum. Or, pour $p \in]0,1[$,

$$\begin{aligned} \ell'_{x_{1:n}}(p) = 0 \quad &\Leftrightarrow \quad \frac{nK - S}{p} - \frac{S}{1-p} = 0 \\ &\Leftrightarrow \quad -pS + (1-p)(nK - S) = 0 \\ &\Leftrightarrow \quad p = 1 - \frac{S}{nK}. \end{aligned}$$

On en déduit que l'estimateur du maximum de vraisemblance vaut bien $\hat{p}_{n,1} = 1 - \bar{X}_n/K$.

5. Comme $E_p(|X_1|) < +\infty$, la loi des grands nombres donne

$$\bar{X}_n \xrightarrow[n \to +\infty]{\mathbb{P}} E_p(X_1) = K(1-p).$$

Comme $g : x \mapsto 1 - x/K$ est continue en $K(1-p)$, on en déduit que

$$\hat{p}_{n,1} = g(\bar{X}_n) \xrightarrow[n \to +\infty]{\mathbb{P}} g(K(1-p)) = p.$$

$\hat{p}_{n,1}$ est donc consistant.

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

$$\sqrt{n}(\bar{X}_n - K(1-p)) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}_p(X_1)),$$

avec $\mathrm{Var}_p(X_1) = Kp(1-p)$. Comme g est dérivable en $K(1-p)$, la méthode Delta donne alors

$$\sqrt{n}(\hat{p}_{n,1} - p) = \sqrt{n}(g(\bar{X}_n) - g(K(1-p))) \rightsquigarrow g'(K(1-p))\mathcal{N}(0, Kp(1-p)) \sim \mathcal{N}(0, p(1-p)/K).$$

On en déduit que $\hat{p}_{n,1}$ est asymptotiquement normal.

7. D'après la question précédente, on a

$$\sqrt{\frac{Kn}{p(1-p)}}(\hat{p}_{n,1} - p) \rightsquigarrow \mathcal{N}(0,1).$$

Comme $\hat{p}_{n,1}$ est consistant, le Lemme de Slutsky donne alors

$$\sqrt{\frac{Kn}{\hat{p}_{n,1}(1-\hat{p}_{n,1})}}(\hat{p}_{n,1} - p) \rightsquigarrow \mathcal{N}(0,1).$$

Soit q le quantile d'ordre 95% d'une loi $\mathcal{N}(0,1)$. On en déduit que

$$P_p\left(\sqrt{\frac{Kn}{\hat{p}_{n,1}(1-\hat{p}_{n,1})}}(\hat{p}_{n,1} - p) \leq q\right) \underset{n\to+\infty}{\longrightarrow} 95\%.$$

Or,

$$\sqrt{\frac{Kn}{\hat{p}_{n,1}(1-\hat{p}_{n,1})}}(\hat{p}_{n,1} - p) \leq q \quad \Leftrightarrow \quad p \geq \hat{p}_{n,1} - \frac{q\sqrt{\hat{p}_{n,1}(1-\hat{p}_{n,1})}}{Kn}.$$

On en déduit que

$$\left[\hat{p}_{n,1} - \frac{q\sqrt{\hat{p}_{n,1}(1-\hat{p}_{n,1})}}{Kn}; 1\right[$$

est un intervalle de niveau de confiance asymptotique 95%.

8. Au vu de l'intervalle de confiance précédent, la compagnie peut miser sur un taux d'absentéisme d'au moins $p = 0.02$. Pour ce p, pour K billets vendus, le nombre moyen de places occupées va être $K(1-p)$. La compagnie va donc choisir K pour que

$$K(1-p) = 196,$$

donc 200 billets.

Exercice 2 - Surmulots parisiens

La concentration de surmulots observés par mètre carré dans une zone de Paris tirée au hasard est modélisée par la loi P_r de densité sur $\mathbb{R}^+$ donnée par

$$f_r(x) = rx^{-2}\mathbb{1}_{x\geq r},$$

pour $x \in \mathbb{R}$ et $r > 0$ inconnu. Les service de la mairie collectent n observations de concentrations de surmulots dans des zones tirées au hasard de manière indépendante.

1. Montrer que f_r est bien une densité.
2. Donner le modèle de l'expérience.
3. Peut-on utiliser la méthode des moments pour estimer r ? (*À titre de rappel, une justification est nécessaire*).
4. Donner un estimateur de r par la méthode du maximum de vraisemblance. On le notera $\hat{r}_n$.
5. Montrer que, pour tout $t \geq r$,
$$P_r(\hat{r}_n > t) = \left(\frac{r}{t}\right)^n.$$
6. En déduire que $\hat{r}_n$ est consistant.
7. En déduire que $n(\hat{r}_n - r) \rightsquigarrow \mathcal{E}(1/r)$ (loi exponentielle de paramètre $1/r$).
8. Montrer que la médiane de P_r vaut $2r$.
9. Un opposant politique à la mairie veut prouver que la concentration de surmulots médiane est plus grande (strictement) que 10. Quelles sont les hypothèses du test associé ?
10. Construire un test (non asymptotique) de niveau 5% pour ces hypothèses, basé sur $\hat{r}_n$.
11. **Application** : Sur 10 observations la mairie observe une fréquence minimale de surmulots de 6. L'opposition peut-elle crier au scandale ? (On pourra utiliser le fait que $(5/6)^9 \geq 0.19$).

Solution 2 -

1. Soit $r > 0$. f_r est bien positive, et vérifie
$$\int_0^{+\infty} f_r(u)du = r\int_r^{+\infty} u^{-2}du = r/r = 1.$$
C'est donc bien une densité.

2. On note X_i la i-ème observation de fréquence de surmulots. On peut supposer que les X_i sont i.i.d. de loi P_r, pour $r > 0$ inconnu. Le modèle est donc
$$((\mathbb{R}^+)^n, (P_r^{\otimes n})_{r>0}).$$
3. Comme $E_r(|X_1|) = +\infty$, la méthode des moments n'est pas utilisable (enfin pour des moments d'ordre plus grands que 1).
4. Soit $x_1, \dots, x_n \geq 0$, et $r > 0$. La vraisemblance s'écrit
$$V_{x_{1:n}}(r) = \prod_{i=1}^n r x_i^{-2} \mathbb{1}_{x_i \geq r} = r^n \mathbb{1}_{\min_{i=1,\dots,n} x_i \geq r} (\prod_{i=1}^n x_i^{-2}).$$
Cette vraisemblance pouvant s'annuler, on ne peut pas passer au logarithme. Cela dit, $V_{x_{1:n}}$ est strictement croissante sur $]0; \min_{i=1,\dots,n} x_i]$ et vaut 0 après. On en déduit alors que le maximum est atteint en $\min_{i=1,\dots,n} x_i$, et que l'estimateur par maximum de vraisemblance vaut alors
$$\hat{r}_n = \min_{i=1,\dots,n} X_i.$$
5. Soit $r > 0$ et $t \geq r$. On a
$$\begin{aligned} P_r(\hat{r}_n > t) = P_r(\min_{i=1,\dots,n} X_i > t) &= (P_r(X_1 > t))^n \qquad \text{(par indépendance)} \\ &= \left(r \int_t^{+\infty} u^{-2} du \right)^n = \left(\frac{r}{t} \right)^n. \end{aligned}$$
6. Soit $\varepsilon > 0$. Comme $\hat{r}_n \geq r$, on a
$$P_r(|\hat{r}_n - r| > \varepsilon) = P_r(\hat{r}_n > r + \varepsilon) = \left(\frac{r}{r+\varepsilon} \right)^n \underset{n \to +\infty}{\longrightarrow} 0,$$
d'après la question précédente et en remarquant que $r/(r+\varepsilon) < 1$. On en déduit que $\hat{r}_n$ est consistant.
7. Soit $t \geq 0$. On a alors $r + t/n \geq r$. En utilisant la question $5-$, on a
$$\begin{aligned} P_r\left(n(\hat{r}_n - r) > t\right) &= P_r\left(\hat{r}_n > r + \frac{t}{n}\right) \\ &= \left(\frac{r}{r + t/n} \right)^n = \left(\frac{1}{1 + t/(nr)} \right)^n \underset{n \to +\infty}{\longrightarrow} \exp\left(-t/r\right). \end{aligned}$$
On en déduit alors que $n(\hat{r}_n - r) \rightsquigarrow \mathcal{E}(1/r)$.
8. Comme
$$P_r(X_1 \leq 2r) = \int_0^{2r} f_r(u) du = \left[-r u^{-1} \right]_r^{2r} = 1/2,$$
la médiane de P_r vaut bien $2r$.

9. L'opposant politique veut prouver que $2r > 10$. Les hypothèses de test sont donc

$$\begin{cases} H_0 & : \quad r = 5, \\ H_1 & : \quad r > 5. \end{cases}$$

10. Sous H_1 on s'attend à ce que $\hat{r}_n$ soit grand. On choisit donc une zone de rejet de la forme $[t_{5\%}; +\infty[$, où $t_{5\%}$ doit vérifier

$$P_5(\hat{r}_n \geq t_{5\%}) \leq 5\%.$$

Or, d'après la question $5-$, on a, pour $t \geq 5$,

$$P_5\left(\hat{r}_n \geq t\right) = (5/t)^n.$$

En choisissant $t_{5\%} = \frac{5}{(5\%)^{1/n}}$, on a bien

$$P_5(\hat{r}_n \geq t_{5\%}) = 5\%.$$

Le test

$$T(X_{1:n}) = \mathbb{1}_{\hat{r}_n \geq \frac{5}{(5\%)^{1/n}}}$$

est donc bien de niveau 5% pour ces hypothèses.

11. On a $n = 10$ et $\hat{r}_n = 6$. Donc ici

$$\begin{aligned} \hat{r}_n \geq \frac{5}{(5\%)^{1/n}} \quad &\Leftrightarrow \quad 6^{10} \geq 5^9 \times 100 \\ &\Leftrightarrow (5/6)^9 \leq 6/100. \end{aligned}$$

Comme $(5/6)^9 \geq 0.19 > 0.06$, l'opposition ne peut qu'accepter l'hypothèse nulle et donc s'abstenir de crier au scandale (dans un monde idéal).