CC1 2024/2025 – Durée 1h30

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

Exercice 1 - Lois de Raab-Green

La loi de Raab-Green de paramètre $\theta > 0$, notée $RG(\theta)$, est définie par sa densité sur $\mathbb{R}$

$$f_{\theta}(x) = \frac{1 + \cos(\pi x/\theta)}{2\theta} \mathbb{1}_{|x| \leq \theta}.$$

1. Montrer que f_{θ} est bien une densité.
2. Soit $X \sim RG(\theta)$. Calculer $\mathbb{E}(X)$, et montrer que $\mathbb{E}(X^2) = v_2\theta^2$, $\mathbb{E}(X^4) = v_4\theta^4$, pour des constantes v_2 et v_4 que l'on ne calculera pas pour le moment.

On se donne $X_1, \dots, X_n$ i.i.d. de loi $RG(\theta)$, pour un paramètre $\theta > 0$ inconnu que l'on va chercher à estimer.

3. Donner un modèle pour cette expérience.
4. Trouver un estimateur $\hat{\theta}_{n,1}$ basé sur $\overline{X^2}_n = (1/n) \sum_{i=1}^n X_i^2$, et montrer sa consistance.
5. Déterminer le comportement asymptotique de $\hat{\theta}_{n,1}$, et en déduire un intervalle de niveau de confiance asymptotique $1 - \alpha$ sur θ , où $\alpha \in]0, 1[$. On le notera I_1 .
6. On pose maintenant $\hat{\theta}_{n,2} = \max_{i=1, \dots, n} |X_i|$. Montrer que, pour tout $\theta > 0$ et $0 \leq t \leq \theta$,

$$P_{\theta} \left(\hat{\theta}_{n,2} \leq \theta - t \right) = \left(1 - \frac{t}{\theta} + \frac{\sin(\pi t/\theta)}{\pi} \right)^n.$$

7. En déduire que $\hat{\theta}_{n,2}$ est consistant.
8. Montrer que $n(\theta - \hat{\theta}_{n,2})^3$ converge en loi.
9. En déduire un intervalle de niveau de confiance asymptotique $1 - \alpha$ sur θ , que l'on notera I_2 .
10. Comparer I_1 et I_2 sur la base de leurs longueurs lorsque $n \rightarrow +\infty$ (on pourra admettre les convergences presque sûrement des deux estimateurs).
11. Calculer v_2 et v_4 .

Solution 1 -

1. Soit $|x| \leq |\theta|$. Comme $\cos(\pi x/\theta) \geq -1$, f_θ est bien positive. On remarque que $f_\theta(x) = \frac{1}{\theta} f_1(x/\theta)$, et en déduit alors

$$\begin{aligned} \int_{-\theta}^{\theta} f_\theta(x) dx &= \int_{-1}^1 f_1(x) dx \\ &= \int_{-1}^1 \frac{1 + \cos(\pi x)}{2} dx = \int_{-1}^1 \frac{1}{2} dx = 1, \end{aligned}$$

en utilisant $\int_{-\pi}^{\pi} \cos(u) du = 0$. f_θ est donc bien une densité.

2. Soit $X \sim RG(\theta)$. On commence par remarquer que $|X|$ est bornée (par θ), donc tous les moments en question existent bien. Ensuite, f_θ étant paire, on a $\mathbb{E}(X) = 0$. Puis, on remarque que $X/\theta \sim RG(1)$ (se déduit de $f_\theta(x) = \frac{1}{\theta} f_1(x/\theta)$). On a alors, si $Y \sim RG(1)$,

$$\begin{aligned} \mathbb{E}(X^2) &= \theta^2 \mathbb{E}(Y^2), \\ \mathbb{E}(X^4) &= \theta^4 \mathbb{E}(Y^4), \end{aligned}$$

ces moments sont donc de la forme voulue, avec $v_j = \mathbb{E}(Y^j)$.

3. Modèle; $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (RG(\theta)^{\otimes n})_{\theta > 0})$.

4. Soit $\theta > 0$. Comme $E_\theta(\overline{X^2}_n) = E_\theta X_1^2 = \theta^2 v_2$, on pose $\hat{\theta}_{n,1} := \sqrt{\overline{X^2}_n / v_2}$. Comme $E_\theta(X_1^2) < +\infty$, la loi des grands nombres indique que $\overline{X^2}_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} v_2 \theta^2$. Comme $f : x \mapsto \sqrt{x/v_2}$ est continue en θ^2 , $\hat{\theta}_{n,1}$ est consistant.

5. Comme $E_\theta X_1^4 < +\infty$, le Théorème Central Limite donne

$$\sqrt{n}(\overline{X^2}_n - v_2 \theta^2) \rightsquigarrow \mathcal{N}(0, v'_4 \theta^4),$$

avec $v'_4 = v_4 - v_2^2$. Comme f est différentiable en $v_2 \theta^2$, de dérivée $f'(v_2 \theta^2) = \frac{1}{2\theta v_2}$, on en déduit par la méthode Δ que

$$\sqrt{n}(\hat{\theta}_{n,1} - \theta) \rightsquigarrow \frac{1}{2\theta v_2} \mathcal{N}(0, v'_4 \theta^4) \sim \sigma \theta \mathcal{N}(0, 1),$$

où $\sigma = \sqrt{v'_4} / (2v_2)$ est connue. Soit alors q le quantile d'ordre $1 - \alpha/2$ d'une loi $\mathcal{N}(0, 1)$, on en déduit que

$$P_\theta \left(-q \leq \frac{\sqrt{n}}{\sigma} \left(\frac{\hat{\theta}_{n,1}}{\theta} - 1 \right) \leq q \right) \xrightarrow[n \rightarrow +\infty]{} 1 - \alpha,$$

et donc que $I_1 = \left[\frac{\hat{\theta}_{n,1}}{1 + \sigma q / \sqrt{n}}; \frac{\hat{\theta}_{n,1}}{1 - \sigma q / \sqrt{n}} \right]$ (bien défini pour n assez grand) est un intervalle de niveau de confiance asymptotique $1 - \alpha$.

6. On note $Y_i = |X_i|$. Soit $\theta > 0$, et $0 \leq t \leq \theta$. On a

$$\begin{aligned}
P_\theta \left(\hat{\theta}_{n,2} \leq \theta - t \right) &= P_\theta \left(\max_{i=1,\dots,n} Y_i / \theta \leq 1 - t/\theta \right) \\
&= P_1 \left(\max_{i=1,\dots,n} Y_i \leq 1 - t/\theta \right) \\
&= \left(\int_0^{1-t/\theta} (1 + \cos(\pi u)) du \right)^n \\
&= \left(1 - \frac{t}{\theta} + \frac{\sin \left[\pi \left(1 - \frac{t}{\theta} \right) \right]}{\pi} \right)^n \\
&= \left(1 - \frac{t}{\theta} + \frac{\sin(\pi t/\theta)}{\pi} \right)^n,
\end{aligned}$$

en utilisant $\sin(\pi - x) = \sin(x)$, pour $x \in [0, \pi]$.

7. Soit $0 < t \leq \theta$. Comme $\sin(\pi t/\theta) < \pi t/\theta$, on déduit que

$$\left(1 - \frac{t}{\theta} + \frac{\sin(\pi t/\theta)}{\pi} \right) < 1,$$

et donc que $P_\theta \left(\hat{\theta}_{n,2} \leq \theta - t \right) \xrightarrow{n \rightarrow +\infty} 0$. $\hat{\theta}_{n,2}$ est donc consistant.

8. Soit $t \geq 0$.

$$\begin{aligned}
P_\theta \left(n(\theta - \hat{\theta}_{n,2})^3 \geq t \right) &= P_\theta \left(\hat{\theta}_{n,2} \leq \theta - (t/n)^{1/3} \right) \\
&= \left(1 - \frac{1}{\theta} (t/n)^{1/3} + \frac{\sin(\pi (t/n)^{1/3} / \theta)}{\pi} \right) \mathbb{1}_{t \leq n\theta^3}.
\end{aligned}$$

Comme $\mathbb{1}_{t \leq n\theta^3} \xrightarrow{n \rightarrow +\infty} 1$, on s'intéresse au premier terme. On remarque que

$$\begin{aligned}
\frac{\sin(\pi (t/n)^{1/3} / \theta)}{\pi} &= \frac{1}{\pi} \left(\frac{\pi}{\theta} (t/n)^{1/3} - \frac{1}{6} \frac{\pi^3 t}{n\theta^3} + o(1/n) \right) \\
&= \frac{(t/n)^{1/3}}{\theta} - \frac{\pi^2 t}{6n\theta^3} + o(1/n).
\end{aligned}$$

On en déduit alors que

$$\begin{aligned}
\left(1 - \frac{1}{\theta} (t/n)^{1/3} + \frac{\sin(\pi (t/n)^{1/3} / \theta)}{\pi} \right) &= \left(1 - \frac{\pi^2 t}{6n\theta^3} + o(1/n) \right)^n \\
&\xrightarrow{n \rightarrow +\infty} \exp \left(-\frac{\pi^2 t}{6\theta^3} \right).
\end{aligned}$$

On en déduit alors que $n(\theta - \hat{\theta}_{n,2})^3 \rightsquigarrow \mathcal{E} \left(\frac{\pi^2}{6\theta^3} \right)$.

9. Posons alors $t_\alpha = \frac{6 \log(1/\alpha)}{\pi^2}$, de telle sorte que

$$\mathbb{P} \left(\mathcal{E} \left(\frac{\pi^2}{6\theta^3} \right) \geq \theta^3 t_\alpha \right) = \alpha.$$

On déduit de ce qui précède que

$$P_\theta \left(n(\theta - \hat{\theta}_{n,2})^3 \geq \theta^3 t_\alpha \right) \xrightarrow{n \rightarrow +\infty} \alpha,$$

et donc que $I_2 = \left[\hat{\theta}_{n,2}; \frac{\hat{\theta}_{n,2}}{1 - (t_\alpha/n)^{1/3}} \right]$ est un intervalle de niveau de confiance asymptotique $1 - \alpha$.

10. On note L la longueur. En admettant la consistance forte de $\hat{\theta}_{n,1}$ et $\hat{\theta}_{n,2}$, on a

$$\begin{aligned} L(I_1) &= \frac{2\sigma q \hat{\theta}_{n,1}}{\sqrt{n}} + o(1/\sqrt{n}) \sim \frac{2\sigma q \theta}{\sqrt{n}} \quad p.s. \\ L(I_2) &= \hat{\theta}_{n,2} (t_\alpha/n)^{1/3} + o(1/n^{1/3}) \sim \theta t_\alpha^{1/3} n^{-1/3} \quad p.s.. \end{aligned}$$

On en déduit que $L(I_1) = o(L(I_2))$ p.s., et donc que I_1 est asymptotiquement meilleur.

11. On a

$$\begin{aligned} v_2 &= \mathbb{E}(Y^2) = \int_0^1 (x^2 + x^2 \cos(\pi x)) dx \\ &= \frac{1}{3} + \int_0^1 x^2 \cos(\pi x) dx. \end{aligned}$$

Par ailleurs, en intégrant par parties,

$$\begin{aligned} \int_0^1 x^2 \cos(\pi x) dx &= \left[x^2 \frac{\sin(\pi x)}{\pi} \right]_0^1 - \int_0^1 2x \frac{\sin(\pi x)}{\pi} dx \\ &= \frac{2}{\pi} \left[x \frac{\cos(\pi x)}{\pi} \right]_0^1 - \frac{2}{\pi} \int_0^1 \frac{\cos(\pi x)}{\pi} dx \\ &= -\frac{2}{\pi^2}. \end{aligned}$$

On en déduit que $v_2 = \frac{1}{3} - \frac{2}{\pi^2}$. Pour v_4 , le raisonnement est le même.

$$\begin{aligned} v_4 &= \int_0^1 x^4 dx + \int_0^1 x^4 \cos(\pi x) dx \\ &= \frac{1}{5} + \int_0^1 x^4 \cos(\pi x) dx. \end{aligned}$$

Et on calcule encore avec intégrations successives par partie

$$\begin{aligned} \int_0^1 x^4 \cos(\pi x) dx &= \left[x^4 \frac{\sin(\pi x)}{\pi} \right]_0^1 - \int_0^1 4x^3 \frac{\sin(\pi x)}{\pi} \\ &= \frac{4}{\pi} \left[x^3 \frac{\cos(\pi x)}{\pi} \right]_0^1 - \frac{4}{\pi} \int_0^1 3x^2 \frac{\cos(\pi x)}{\pi} \\ &= -\frac{12}{\pi^2} \int_0^1 x^2 \cos(\pi x) dx - \frac{4}{\pi^2}. \end{aligned}$$

Comme $\int_0^1 x^2 \cos(\pi x) dx = -2/\pi^2$, on en déduit que

$$v_4 = \frac{1}{5} + \frac{24}{\pi^4} - \frac{4}{\pi^2}.$$

Exercice 2 - Prix IgNobel 2024 de Probabilités

En 2007, P. Diaconis et ses coauteurs faisaient la prévision suivante : "si on tire à pile ou face avec une pièce équilibrée, en partant d'un côté (disons pile) sur le pouce, alors la probabilité que la pièce retombe du même côté (pile donc) n'est pas 50% mais p , où $p > 50\%$ ". Ils calculaient ensuite qu'une valeur plausible de p est autour de 51%. Vous avez à votre disposition une pièce et beaucoup de temps.

N.B. : *Les questions sont volontairement laissées ouvertes, il vous faudra décrire des procédures de test **complètes**. Vous remarquerez aussi qu'aucune valeur de quantile n'est fournie, il est donc recommandé de ne pas en utiliser.*

1. Proposer une expérience et un test permettant de prouver la conjecture de P. Diaconis, avec certitude d'au moins 95% (de manière non asymptotique). On notera ce test T_n .
2. Supposons que P. Diaconis a raison, et que $p \geq 50,5\%$ (pour se donner de la marge). Calculer le nombre de lancers n nécessaire pour être sûr à 95% que le test précédent T_n détecte bien ce biais. *Indication pour le calcul final : regarder ce que ça donne pour $n \geq 2 \log(20)/(5\%)^2$.*
3. **Application numérique :** On donne les valeurs approchées suivantes : $\log(20) \approx 3$, $(0,5\%)^{-2} = 40000$. Donner un ordre de grandeur du nombre de lancers nécessaire pour prouver la conjecture de P. Diaconis dans le cadre des questions 1 et 2.
4. (Pour 0 points) : Le prix IgNobel 2024 a été attribué à une équipe de 50 chercheurs (non mathématiciens) qui ont prouvé que la conjecture de P. Diaconis était vraie, au niveau de confiance 95%, sur la base d'environ 350000 lancers, la vraie valeur étant située entre 0.506 et 0.509 au même niveau de confiance. Ces chercheurs sont ils trop payés ?

Solution 2 -

1. On va sans surprise lancer la pièce n fois en partant du côté pile, noter X_i le i -ème tirage, et supposer que les X_i sont i.i.d. de loi $\mathcal{B}(p)$. Le modèle correspondant s'écrit alors

$$(\{0, 1\}^n, \mathcal{P}(\{0, 1\}^n), (\mathcal{B}(p)^{\otimes n})_{p \in]0;1[}).$$

Pour ce modèle, on veut prouver $p > 1/2$. On pose alors les hypothèses

$$\begin{cases} H_0 & : p \leq 1/2, \\ H_1 & : p > 1/2. \end{cases}$$

Une statistique suivant le comportement de p est $\hat{p}_n = \bar{X}_n$. Sous H_1 on s'attend à de plus grandes valeurs de $\hat{p}_n$ que sous H_0 , on prendra donc une région de rejet de la forme $[t; 1]$. Calibrons le t de manière non-asymptotique pour obtenir un test de niveau 5%. Il faut

$$\sup_{p \leq 1/2} P_p(\hat{p}_n \geq t) \leq 5\%.$$

Comme $\sum_{i=1}^n X_i \sim \mathcal{B}(n, p)$, et que pour $p_1 \leq p_2$, $\mathcal{B}(n, p_1) \preceq \mathcal{B}(n, p_2)$ (domination stochastique), on a, pour tout $t \in [0, 1]$,

$$\sup_{p \leq 1/2} P_p(\hat{p}_n \geq t) = P_{1/2}(\hat{p}_n \geq t).$$

On n'a pas accès aux tables de la loi Binomiale, et va donc passer par l'inégalité de Hoeffding. Pour $\varepsilon > 0$, on a

$$P_{1/2}\left(\hat{p}_n \geq \frac{1}{2} + \varepsilon\right) \leq \exp[-2n\varepsilon^2].$$

On prenant $\varepsilon = \sqrt{\log(1/(5\%))/(2n)} = \sqrt{\log(20)/(2n)}$, on a alors que

$$T_n = \mathbb{1}_{\bar{X}_n \geq 1/2 + \sqrt{\log(20)/(2n)}}$$

est un test de niveau 5% pour ces hypothèses.

2. On suppose que P. Diaconis a raison, et que $p \geq p_1 = 50, 5\%$. On veut être sûrs à 95% que T_n détecte ce biais. Il faut alors

$$\inf_{p \geq p_1} P_p(T_n = 1) = \inf_{p \geq p_1} P_p\left(\hat{p}_n \geq 1/2 + \sqrt{\log(20)/(2n)}\right) \geq 95\%.$$

En utilisant l'ordre stochastique sur les $\mathcal{B}(n, p)$ là encore, on a

$$\inf_{p \geq p_1} P_p(T_n = 1) = P_{p_1}\left(\hat{p}_n \geq 1/2 + \sqrt{\log(20)/(2n)}\right).$$

Notons $\delta = 0.5\%$. On a

$$\begin{aligned} P_{p_1} \left(\hat{p}_n < 1/2 + \sqrt{\log(20)/(2n)} \right) &= P_{p_1} \left(\hat{p}_n - p_1 < -\delta + \sqrt{\log(20)/(2n)} \right) \\ &\leq \exp \left[-2n(\delta - \sqrt{\log(20)/(2n)})_+^2 \right], \end{aligned}$$

en utilisant encore l'inégalité de Hoeffding. Il reste alors à résoudre

$$\exp \left[-2n(\delta - \sqrt{\log(20)/(2n)})_+^2 \right] \leq 5\%$$

On remarque que cette inégalité est satisfaite lorsque $n \geq 2 \log(20)/\delta^2$ (dans ce cas $(\delta - \sqrt{\log(20)/(2n)})_+^2 \geq \delta^2/4$ et on résoud $\exp(-n\delta^2/2) \leq 5\%$).

3. En utilisant les valeurs proposées, pour prouver que P. Diaconis a raison et si on veut être sûrs à 95% de le détecter, il faut prendre $n \gtrsim 240000$.
4. Notre approximation du nombre de lancers nécessaire étant pessimiste (utilisation de Hoeffding là où Bernstein, les tables de la loi Binomiale, ou encore l'approximation normale auraient donné moins, + erreurs d'arrondis toujours au sup), ces chercheurs ont effectué moitié plus de lancers que nécessaire. Ils sont donc trop payés (d'un facteur 3/2 à peu près).