Examen Partiel – Octobre 2018

Durée 2h00 – Sans documents

Important. Suivant les réglements en vigueur,

1. les enseignants présents lors de l'épreuve ne peuvent communiquer que sur les fautes d'énoncé potentielles. Toute autre question durant la composition ne sera pas acceptée.
2. les étudiants sont tenus de se lever au moment de l'annonce de fin de la composition. En cas de refus, le responsable de l'UE sera fondé à ne pas prendre en compte la copie incriminée.
3. l'identification des copies et intercalaires doit se faire au moment de la remise de chaque copie par les enseignants et surveillants. Il ne sera pas accordé de délai pour cette raison en fin d'épreuve.

Exercice 1 (6 pts). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires *i.i.d.* uniformes sur $[0,\theta]$, où $\theta > 0$ est un paramètre inconnu. On rappelle la densité d'une variable aléatoire uniforme sur $[0,\theta]$:

$$f(x|\theta) := \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x).$$

1. *(1 pt)* Soit

$$\widehat{\theta}_n := \frac{2}{n}\sum_{i=1}^{n} X_i.$$

Justifier la convergence presque sûre de cet estimateur.

2. *(0.5 pt)* En déduire un estimateur qui converge presque sûrement vers $\mathrm{Var}_\theta(X_1)$.

3. *(1.5 pts)* Montrer que

$$\sqrt{3n}\left(1-\frac{\theta}{\widehat{\theta}_n}\right) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,1).$$

4. *(1.5 pts)* Énoncer la méthode delta. Retrouver en utilisant cette méthode le résultat précédent.

5. *(1.5 pts)* Soit $\alpha \in]0,1[$.

(a) Trouver $q > 0$ tel que

$$\mathbb{P}_\theta\left(\theta \in \left[\left(1-\frac{q}{\sqrt{3n}}\right)\hat{\theta}_n, \left(1+\frac{q}{\sqrt{3n}}\right)\hat{\theta}_n\right]\right) \xrightarrow[n\to+\infty]{} \alpha.$$

(b) On choisit $\alpha = 0.9$. Donner la commande R qui renvoie la valeur de q.

Solution Exercice 1. On pose

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

1. On a $\mathbb{E}_\theta[X_1] = {}^{\theta}\!/_{2}$. Les variables aléatoires étant *i.i.d.* et intégrables, la loi forte des grands nombres donne que $\overline{X}_n$ converge presque sûrement vers ${}^{\theta}\!/_{2}$. La fonction $x \mapsto 2x$ étant continue, $\widehat{\theta}_n := 2\overline{X}_n$ converge presque sûrement vers θ.

2. On obtient par simple calcul $\mathrm{Var}_\theta[X_1] = {}^{\theta^2}\!/_{12}$. D'après la question précédente et par continuité de $x \mapsto {}^{x^2}\!/_{12}$, on obtient que ${}^{\hat\theta_n^2}\!/_{12}$ converge presque sûrement vers $\mathrm{Var}_\theta(X_1)$.

3. Les variables aléatoires étant *i.i.d.* et de carré intégrables, le théorème central limite donne

$$\sqrt{n}\left(\overline{X}_n - \frac{\theta}{2}\right) \xrightarrow[n\to+\infty]{loi} \mathcal{N}\left(0, \frac{\theta^2}{12}\right). \tag{1}$$

Par ailleurs ${}^{\hat\theta_n^2}\!/_{12}$ converge presque sûrement donc converge en loi vers la constante ${}^{\theta^2}\!/_{12}$. On a donc en utilisant le théorème de Slutsky :

$$\frac{1}{\sqrt{\hat\theta_n^2/12}} \times \sqrt{n}\left(\overline{X}_n - \frac{\theta}{2}\right) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,1).$$

Il suffit alors de remarquer

$$\frac{\sqrt{n}}{\sqrt{\hat\theta_n^2/12}}\sqrt{n}\left(\overline{X}_n - \frac{\theta}{2}\right) = \sqrt{12n}\left(\frac{1}{2} - \frac{\theta}{2\hat\theta_n}\right) = \sqrt{3n}\left(1 - \frac{\theta}{\hat\theta_n}\right).$$

4. Le théorème central limite (1) donne

$$\sqrt{n}\left(\hat\theta_n - \theta\right) \xrightarrow[n\to+\infty]{loi} \mathcal{N}\left(0, \frac{\theta^2}{3}\right).$$

Soit g une fonction de $\mathbb{R}$ dans $\mathbb{R}$ différentiable telle que $g'(\theta) \neq 0$. La méthode delta assure que

$$\sqrt{n}\left[g(\hat\theta_n) - g(\theta)\right] \xrightarrow[n\to+\infty]{loi} \mathcal{N}\left(0, \frac{\theta^2}{3}\left[g'(\theta)\right]^2\right).$$

On cherche g tel que

$$\int g'(t)\mathrm{d}t = \int \frac{\sqrt{3}}{t}\mathrm{d}t \quad\Leftrightarrow\quad g(t) = \sqrt{3}\ln t + \mathrm{cste}.$$

La méthode delta donne alors

$$\sqrt{n}\left(\sqrt{3}\ln\hat\theta_n - \sqrt{3}\ln\theta\right) = \sqrt{3n}\left(0 - \ln\frac{\theta}{\hat\theta_n}\right) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,1).$$

Pour obtenir le résultat souhaité, il suffit alors d'appliquer la méthode delta avec $h : x \mapsto e^x$:

$$\sqrt{3n}\left[h(0) - h\left(\ln\frac{\theta}{\hat\theta_n}\right)\right] \xrightarrow[n\to+\infty]{loi} \mathcal{N}\left(0, \left[h'(0)\right]^2\right) \quad\Leftrightarrow\quad \sqrt{3n}\left(1 - \frac{\theta}{\hat\theta_n}\right) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,1).$$

5. Soit $\alpha \in]0,1[$.

(a) On commence par écrire

$$\mathbb{P}_\theta\left(\theta \in \left[\left(1-\frac{q}{\sqrt{3n}}\right)\hat{\theta}_n, \left(1+\frac{q}{\sqrt{3n}}\right)\hat{\theta}_n\right]\right) = \mathbb{P}_\theta\left[-q \leq \sqrt{3n}\left(\frac{\theta}{\hat{\theta}_n}-1\right) \leq q\right].$$

Soit Z une variable aléatoire de loi normale centrée réduite. D'après la question 3, on a

$$\mathbb{P}_\theta\left[-q \leq \sqrt{3n}\left(\frac{\theta}{\hat{\theta}_n}-1\right) \leq q\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{P}[-q \leq Z \leq q].$$

On cherche donc q tel que $\mathbb{P}[-q \leq Z \leq q] = \alpha$, autrement dit q est le quantile d'ordre $1 - \frac{1-\alpha}{2} = \frac{1+\alpha}{2}$ de la loi normale $\mathcal{N}(0,1)$.

(b)
```
qnorm(0.95)
```

Exercice 2 (9 pts). Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires *i.i.d* de loi de Cauchy $\mathcal{C}(\beta)$. On rappelle que la densité de la loi de Cauchy est :

$$f_X(x) = \frac{1}{\pi(1+(x-\beta)^2)},$$

et sa fonction caractéristique : $\phi_\beta(t) = \mathbb{E}_\beta[\exp(itX)] = \exp(\beta it - |t|)$. L'objectif est de comparer deux estimateurs de β.

1. *(1 pt)* Déterminer la limite en loi de $\frac{1}{n}\sum_{i=1}^n X_i$.

2. *(0.5 pt)* Montrer que la médiane de $\mathcal{C}(\beta)$ est β.

On propose deux estimateurs $\hat{\beta}_n^1$ et $\hat{\beta}_n^2$ de β. Si $(X_{(i)})$ est la version ordonnée de (X_i), et $p \in]0, 1/2[$, on définit :

$$\hat{\beta}_n^1 = X_{\lceil n/2 \rceil} \quad \text{et} \quad \hat{\beta}_n^2 = \frac{1}{(1-2p)n}\sum_{i\in\lceil pn\rceil}^{\lceil n-pn\rceil} X_{(i)}.$$

Le premier estimateur est la médiane empirique, le second correspond à la *moyenne tronquée* (en anglais *trimmed mean*), c'est à dire la moyenne calculée sur le sous échantillon constitué des valeurs les moins extrêmes.

3. *(3.5 pts)* Soit $q_n(p)$ le quantile d'ordre p de l'échantillon $\{X_1, \ldots, X_n\}$, et $q(p)$ le quantile théorique d'ordre p de la loi $\mathcal{C}(\beta)$. On rappelle que

$$q_n(p) = \inf\left\{x, F_n(x) = \frac{1}{n}\sum \mathbb{1}_{X_i<x} \geq p\right\} \quad \text{et} \quad q(p) = \inf\{x, F(x) \geq p\},$$

avec F_n la fonction de répartition empirique associée à l'échantillon $\{X_1, \ldots, X_n\}$, et F la fonction de répartition associée à $\mathcal{C}(\beta)$.

(a) Montrer que $F_n(x) \xrightarrow[n\to+\infty]{p.s.} F(x)$ pour tout $x \in \mathbb{R}$.

(b) Soit $\delta > 0$, montrer que pour n assez grand et p.s. $F_n(q(p)-\delta) < p$ et en déduire que $q_n(p) \geq q(p) - \delta$.

(c) Conclure que $\liminf_n q_n(p) \geq q(p)$.

(d) Montrer que $F_n(q(p)+\delta) \geq p$, et en déduire que $\limsup_n q_n(p) \leq q(p)$.

(e) En conclure que $\hat{\beta}_n^1$ converge presque sûrement vers β.

4. *(1pt)* Étant donné un vecteur d'observations `x`, écrire une fonction `estimator1` qui retourne la médiane empirique, c'est à dire $\hat{\beta}_n^1$.

5. *(1 pt)* Parmi les fonctions R suivantes, donner, en justifiant, celle calculant une approximation du biais de $\hat{\beta}_n^1$.

```
bias.A <- function(X, k) {
        alpha.b <- numeric(k)
        for (i in 1:k) {
                alpha.b[i] <- estimator1(sample(X, replace = TRUE))
        }
        return(mean(alpha.b) - estimator1(X))
}
```

```
bias.B <- function(X, k) {
        alpha.b <- numeric(k)
        for (i in 1:k) {
                alpha.b[i] <- estimator1(sample(X,replace=TRUE))
        }
        return(estimator1(alpha.b) - estimator1(X))
}
```

```
bias.C <- function(X, k) {
        alpha.b <- numeric(k)
        for (i in 1:k) {
                alpha.b[i] <- estimator1(sample(X, replace = FALSE))
        }
        return(mean(alpha.b) - estimator1(X))
}
```

```
bias.D <- function(X,k) {
        alpha.b <- numeric(k)
        for (i in 1:k) {
                alpha.b[i] <- estimator1(sample(X, replace = FALSE))
        }
        return(estimator1(alpha.b) - estimator1(X))
}
```

6. *(2 pts)* La fonction calculant $\hat{\beta}_n^2$ est la suivante, où $p \in]0, 1/2[$:

```
estimator2 <- function(x,p) {
        return(mean(x, trim = p))
}
```

Écrire une fonction R nommée `IC.a2` qui calcule un intervalle de confiance pour $\hat{\beta}_n^2$ au niveau de confiance $1-\alpha$ en utilisant la méthode bootstrap de votre choix.

Solution Exercice 2.

1. Posons $S_n = \sum_{i=1}^n X_i$. Les variables $X_1, \ldots, X_n$ étant *i.i.d* de loi de Cauchy $\mathcal{C}(\beta)$, on pour tout réel t

$$\phi_{S_n}(t) = \phi_\beta(t)^n = \exp(n\beta it - |t|) \quad \Rightarrow \quad \phi_{S_n/n}(t) = \phi_{S_n}\left(\frac{t}{n}\right) = \exp(\beta it - |t|).$$

On en déduit d'après le théorème de Lévy que $\frac{1}{n}\sum_{i=1}^n X_i$ converge en loi vers la loi de Cauchy $\mathcal{C}(\beta)$.

2. En appliquant le changement de variable $t \mapsto t-\beta$ qui réalise un $\mathcal{C}^1$-difféomorphisme de $\mathbb{R}$ dans $\mathbb{R}$, on obtient

$$\mathbb{P}_\beta(X \leq \beta) = \int_{-\infty}^0 \frac{1}{\pi(1+t^2)}\mathrm{d}t = \frac{1}{\pi}\left[\arctan(t)\right]_{-\infty}^0 = \frac{1}{2}.$$

Donc la médiane de X est bien β.

Alternative. La densité est symétrique par rapport à $x = \beta$, *i.e.*, $f_X(\beta+x) = f_X(\beta-x)$ donc $\mathbb{P}_\beta(X \leq \beta) = \mathbb{P}_\beta(X \geq \beta) = 0.5$.

3. (a) Soit $x \in \mathbb{R}$. La fonction indicatrice étant mesurable et les variables $X_1, \ldots, X_n$ étant *i.i.d*, $\mathbb{1}_{X_1<x}, \ldots, \mathbb{1}_{X_n<x}$ forme une suite de variables *i.i.d* et intégrables. La loi forte des grands nombres donne

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i<x} \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}_\beta[\mathbb{1}_{X_1<x}] = \mathbb{P}_\beta[X_1 < x] = F(x).$$

(b) Soit $\delta > 0$. D'après la question précédente, on a

$$\forall \epsilon,\ \exists N_\epsilon,\ \forall n \geq N_\epsilon, |F_n(q(p)-\delta) - F(q(p)-\delta)| < \epsilon \quad \text{presque sûrement.} \quad (2)$$

Par ailleurs, on déduit de la définition de $q(p)$ que $F(q(p)-\delta) < p$. En effet, si $q(p)-\delta \in \{x, F(x) \geq p\}$, on obtientrait une contradiction sur la défintion de l'inf. En posant $\epsilon_1 = p - F(q(p)-\delta) > 0$, on a donc en particulier avec (2)

$$\exists N_1,\ \forall n \geq N_1, F_n(q(p)-\delta) < p \quad \text{presque sûrement.}$$

Il en résulte que $q(p)-\delta \notin \{x, F_n(x) \geq p\}$ et comme F_n est croissante

$$q(p) - \delta < \inf\{x, F_n(x) \geq p\} = q_n(p).$$

(c) En passant à la $\liminf$ et en faisant tendre δ vers 0, on obtient le résultat souhaité.

(d) La fonction de répartition empirique étant croissante, on a par définition de $q(p)$

$$F(q(p)+\delta) \geq F(q(p)) \geq q.$$

En posant $\epsilon_2 = F(q(p)+\delta) - p \geq 0$, on a donc en particulier avec (2)

$$\exists N_2,\ \forall n \geq N_2, F_n(q(p)+\delta) \geq p \quad \text{presque sûrement.}$$

Il en résulte par définition de l'inf que $q(p)+\delta \geq \inf\{x, F_n(x) \geq p\} = q_n(p)$. En passant à la $\limsup$ et en faisant tendre δ vers 0, on obtient le résultat souhaité.

(e) Les questions précédentes permettent d'établir que $q_n(p)$ converge presque-sûrement vers $q(p)$ pour tout $p \in]0,1[$. En effet, il existe des ensembles $\mathbb{P}$-négigeable E_1 et E_2 tels que

$$\begin{cases} q(p) \leq \liminf_n q_n(p) \leq \limsup_n q_n(p) & \text{sur } \Omega \setminus E_1, \\ \liminf_n q_n(p) \leq \limsup_n q_n(p) \leq q(p) & \text{sur } \Omega \setminus E_2. \end{cases}$$

On en déduit que sur $\Omega \setminus E_1 \cup E_2$, on a $\liminf_n q_n(p) = \limsup_n q_n(p) = q(p)$, *i.e.*, $\lim_n q_n(p) = q(p)$. Or l'ensemble $E_1 \cup E_2$ est $\mathbb{P}$-négigeable, la convergence est donc presque sûre et en particulier que $\hat{\beta}_n^1 = q_n(0.5)$ converge presque sûrement vers $q(0.5) = \beta$.

4. Il suffit d'utiliser la fonction `median`

```
estimator1 <- function(x){
  return(median(x))
}
```

ou éventuellement la fonction `quantile`

```
estimator1 <- function(x){
  return(quantile(x, 0.5))
}
```

5. L'estimateur boostrap du biais $\mathbb{E}_\beta\left[\hat{\beta}_n^1\right] - \beta$ s'écrit

$$\frac{1}{k}\sum_{j=1}^{k} \text{estimator1}\left(X_1^{(j)}, \ldots, X_n^{(j)}\right) - \text{estimator}(X_1, \ldots, X_n),$$

où $X_1^{(j)}, \ldots, X_n^{(j)}$ est un échantillon simulé suivant la fonction de répartition empirique. Autrement dit un échantillon $X_1^{(j)}, \ldots, X_n^{(j)}$ est obtenu en effectuant un tirage uniforme avec remise dans $X_1, \ldots, X_n$. La bonne fonction est donc `bias.A`.

6. On peut prendre pour intervalle de confiance les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ de l'approximation bootstrap de la distribution de $\hat{\beta}_n^2$.

```
IC.a2 <- function(X, k, alpha){
  beta.b <- numeric(k)
  for (i in 1:k) {
              beta.b[i] <- estimator2(sample(X, replace = TRUE))
  }
  return(quantile(beta.b, probs = c(alpha/2, 1 - alpha/2)))
}
```

Une autre possibilité est d'utiliser une approximation gaussienne de la distribution de $\hat{\beta}_n^2$, *i.e.*, $\hat{\beta}_n^2 \sim \mathcal{N}(\beta, \eta(F)^2)$. On a alors l'intervalle de confiance

$$\left[\hat{\beta}_n^2 - q_{1-\alpha/2}\eta(F_n), \hat{\beta}_n^2 + q_{1-\alpha/2}\eta(F_n)\right],$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ la loi normale centrée réduite et $\eta(F_n)$ un estimateur bootstrap de $\eta(F)$.

```
IC.a2 <- function(X, k, alpha){
  beta <- estimator2(X)
  beta.b <- numeric(k)
  for (i in 1:k) {
              beta.b[i] <- estimator2(sample(X, replace = TRUE))
  }
  sd.b <- sd(beta.b)
  q <- qnorm(1- alpha/2)
  return(c(beta - q*sd.b, beta + q*sd.b))
}
```

Exercice 3 (6 pts). Soit X une variable aléatoire admettant la densité de probabilité suivante :

$$f_X(x;\theta) = \frac{\theta\exp(\theta x)}{\exp(\theta^2)-1}\mathbb{1}_{[0,\theta]}(x)$$

où $\theta > 0$. On considère un n-échantillon $X_1,\dots,X_n$ de X.

1. *(2 pts)* Montrer que la densité est bien normalisée, c'est-à-dire, intègre à 1. Donner la fonction génératrice des moments associée à $f_X(\cdot\,;\theta)$, c'est-à-dire $\mathbb{E}[\exp(tX)]$, $t\in\mathbb{R}$. En déduire le premier moment de X.

2. *(1 pt)* La famille de densités $f_X(\cdot\,;\theta)$ indicée par θ constitue-t-elle une famille exponentielle ? Si oui, donnez le paramètre naturel $\mu=\mu(\theta)$ et la représentation cannonique en $\exp\{\mu\cdot T(x)-\psi(\mu)\}$. Si non, indiquez la partie de $f_X(\cdot\,;\theta)$ qui empêche une représentation exponentielle.

3. *(1 pt)* Montrer que la densité du n-échantillon factorise en

$$g(T(x_1,\dots,x_n);\theta)\times h(x_1,\dots,x_n)$$

où $T(x_1,\dots,x_n)$ est une statistique de dimension deux.

4. *(2 pts)* Calculer la moyenne des deux composantes de $T(X_1,\dots,X_n)$. Pouvez vous en déduire un estimateur sans biais de θ ?

Solution Exercice 3.

1. On a

$$\int_{\mathbb{R}} f_X(x;\theta)\mathrm{d}x = \frac{\theta}{e^{\theta^2}-1}\int_0^\theta e^{\theta x}\mathrm{d}x = \frac{\theta}{e^{\theta^2}-1}\left[\frac{1}{\theta}e^{\theta x}\right]_0^\theta = 1.$$

Pour $t\in\mathbb{R}$, la fonction génératrice des moments est donnée par

$$m_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_0^\theta \frac{\theta e^{(\theta+t)x}}{\exp(\theta^2)-1}dx = \begin{cases} \frac{\theta}{(\theta+t)}\frac{e^{(\theta+t)\theta}-1}{e^{\theta^2}-1} & \text{si } t\neq-\theta,\\ \frac{\theta^2}{e^{\theta^2}-1} & \text{sinon.}\end{cases}$$

Le premier moment de X est $\mathbb{E}[X]=m'_X(0)$. Ici

$$\frac{\mathrm{d}}{\mathrm{d}t}m_X(t) = -\frac{\theta}{(\theta+t)^2}\frac{e^{(\theta+t)\theta}-1}{e^{\theta^2}-1} + \frac{\theta}{(\theta+t)}\frac{\theta e^{(\theta+t)\theta}}{e^{\theta^2}-1},$$

et on en déduit

$$\mathbb{E}[X] = \frac{\theta}{1 - e^{-\theta^2}} - \frac{1}{\theta}.$$

2. La famille n'est pas exponentielle car le support de la fonction indicatrice dépend de θ.

3. La loi du n-échantillon est

$$\frac{\theta^n}{(\exp(\theta^2) - 1)^n} \exp\left(\theta\{x_1 + \ldots + x_n\}\right) \underbrace{\mathbb{I}_{x_1 \leq \theta} \times \ldots \times \mathbb{1}_{x_n \leq \theta}}_{\mathbb{1}_{\max(x_1,\ldots,x_n) \leq \theta} = \mathbb{1}_{x_{(n)} \leq \theta}}$$

Cette densité se factorise donc sous la forme demandée avec $h \equiv 1$ et

$$T(X_1, \ldots, X_n) = (n\bar{X}_n, X_{(n)}), \quad \text{où} \quad \begin{cases} \overline{X}_n & = \frac{1}{n}\sum_{k=1}^n X_i \\ X_{(n)} & = \max\{X1, \ldots, X_n\}. \end{cases}$$

4. La densité de la statistique d'ordre n est

$$f_{X_{(n)}}(x\,;\,\theta) = nf(x)\left(\int_0^x f_X(x\,;\,\theta)\mathrm{d}x\right)^{n-1}.$$

En utilisant la densité de $X_{(n)}$, on trouve que

$$\mathbb{E}[X_{(n)}] = \frac{(n-1)!}{e^{\theta^2} - 1}\left[(\theta - 1)(e^{\theta^2} - 1)^n - 1\right].$$

On ne peut donc pas en déduire un estimateur sans biais de θ.