Sorbonne Université
Master Mathématiques et Applications
MU4MA015 : Statistique

11 janvier 2024
Aucun document autorisé
Durée : 3 heures

Examen

Rappels

— Il sera tenu compte du soin apporté à la présentation et à la rédaction. Si vous n'arrivez pas à démontrer un résultat, vous pouvez l'admettre pour la suite de l'exercice.

— Pour $a > 0$ et $b > 0$, la loi $\mathrm{Gamma}(a, b)$ a pour densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ la fonction

$$x \mapsto \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbb{1}_{x>0}.$$

La loi inverse-Gamma $\mathrm{IG}(a, b)$ est la loi de l'inverse d'une variable aléatoire de loi $\mathrm{Gamma}(a, b)$, et a pour densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ la fonction

$$x \mapsto \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x} \mathbb{1}_{x>0}.$$

Exercice 1 (9 points)

Soient $a > 0$ et $n \geq 2$. On considère le cadre bayésien suivant :

$$\boldsymbol{\theta} \sim \Pi = \mathrm{IG}(a, 1)$$
$$\mathbf{X} = (X_1, \ldots, X_n) \,\big|\, \boldsymbol{\theta} \sim \mathcal{E}\left(\frac{1}{\boldsymbol{\theta}}\right)^{\otimes n},$$

et la fonction de perte $\ell : (\mathbb{R}_+^*)^2 \to \mathbb{R}_+$ définie par

$$\forall (\theta, t) \in (\mathbb{R}_+^*)^2,\ \ell(\theta, t) = \frac{1}{\theta^2}(t - \theta)^2.$$

1. Soit $Z \sim \mathrm{IG}(c, d)$, avec $c > 0$ et $d > 0$.
   (a) Montrer que si $c > 2$, alors $\mathbb{E}[Z] = \frac{d}{c-1}$ et $\mathrm{Var}(Z) = \frac{d^2}{(c-1)^2(c-2)}$.
   Correction. Soit $Z \sim \mathrm{IG}(c, d)$ avec $c > 2$ et $d > 0$. On a

$$\mathbb{E}[Z] = \frac{d^c}{\Gamma(c)} \int_0^{+\infty} x^{-c} e^{-d/x} dx = \frac{d^c}{\Gamma(c)} \frac{\Gamma(c-1)}{d^{c-1}} = \frac{d}{c-1},$$

   où l'on a reconnu le terme général de la densité d'une loi $\mathrm{IG}(c-1, d)$. De plus, en reconnaissant le terme général d'une loi $\mathrm{IG}(c-2, d)$, on a

$$\mathbb{E}[Z^2] = \frac{d^c}{\Gamma(c)} \int_0^{+\infty} x^{-c+1} e^{-d/x} dx = \frac{d^c}{\Gamma(c)} \frac{\Gamma(c-2)}{d^{c-2}} = \frac{d^2}{(c-1)(c-2)}.$$

   Ainsi

$$\mathrm{Var}(Z) = \frac{d^2}{(c-1)(c-2)} - \frac{d^2}{(c-1)^2} = \frac{d^2}{(c-1)^2(c-2)}.$$

(b) Quelle est la loi de $\frac{1}{Z}$? En déduire $\mathbb{E}\left[\frac{1}{Z}\right]$ et $\operatorname{Var}\left(\frac{1}{Z}\right)$.

Correction. Par définition, $Z \overset{\mathcal{L}}{=} \frac{1}{W}$ avec $W \sim \text{Gamma}(c,d)$ donc $\frac{1}{Z} \overset{\mathcal{L}}{=} W$, i.e. $\frac{1}{Z} \sim \text{Gamma}(c,d)$. Il vient $\mathbb{E}\left[\frac{1}{Z}\right] = \frac{c}{d}$ et $\operatorname{Var}\left(\frac{1}{Z}\right) \frac{c}{d^2}$.

Une autre méthode consiste en celle de la fonction muette : pour tout fonction $h \in \mathscr{C}_b^0(\mathbb{R})$,

$$\mathbb{E}\left[h\left(\frac{1}{Z}\right)\right] = \int_0^\infty h\left(\frac{1}{x}\right)\frac{d^c}{\Gamma(c)}x^{-c-1}e^{-d/x}\underset{\cdot}{\mathrm{x}} = \int_0^\infty h(y)\frac{d^c}{\Gamma(c)}y^{c-1}e^{-dy}\underset{\cdot}{\mathrm{y}},$$

avec le changement de variable $y = \frac{1}{x}$, $\underset{\cdot}{\mathrm{x}} = -\frac{\underset{\cdot}{\mathrm{y}}}{y^2}$. Ainsi, $\frac{1}{Z} \sim \text{Gamma}(c,d)$.

2. Donner la loi a posteriori $\Pi[\,\cdot \,|\, \mathbf{X}]$. Vérifier que l'espérance a posteriori $m_{\mathbf{X}}$ vaut $\frac{1+n\overline{X}_n}{a+n-1}$ et préciser la variance a posteriori $v_{\mathbf{X}}$.

Correction. On a

$$\begin{aligned}\pi(\theta \,|\, \mathbf{X}) &\propto \theta^{-a-1}e^{-1/\theta}\mathbb{1}_{\theta>0}\prod_{i=1}^n\left(\frac{1}{\theta}e^{-X_i/\theta}\right)\\ &\propto \theta^{-a-n-1}e^{-(1+n\overline{X}_n)/\theta}\mathbb{1}_{\theta>0}\,.\end{aligned}$$

On reconnaît le terme général de la densité d'une loi $\text{IG}(a+n, 1+n\overline{X}_n)$. Comme $n \geq 2$ et $a > 0$, on a $a+n > 2$, et, par la question 1, on a

$$m_{\mathbf{X}} = \frac{1+n\overline{X}_n}{a+n-1} \qquad \text{et} \qquad v_{\mathbf{X}} = \frac{(1+n\overline{X}_n)^2}{(a+n-1)^2(a+n-2)}\,.$$

3. Déterminer un estimateur de Bayes pour la fonction de perte ℓ et l'a priori Π. On le notera T^*.

Correction. On cherche à minimiser le risque a posteriori. Pour T un estimateur, on a

$$\begin{aligned}\mathbb{E}\left[\ell(\boldsymbol{\theta}, T) \,|\, \mathbf{X}\right] &= \frac{(1+n\overline{X}_n)^{a+n}}{\Gamma(a+n)}\int_0^{+\infty}(T-\theta)^2\theta^{-(a+n+2)-1}e^{-\frac{1+n\overline{X}_n}{\theta}}d\theta\\ &= \frac{(1+n\overline{X}_n)^{a+n}}{\Gamma(a+n)}\frac{\Gamma(a+n+2)}{(1+n\overline{X}_n)^{a+n+2}}\int_0^{+\infty}(T-\theta)^2 d\widetilde{\Pi}_{\mathbf{X}}(\theta)\,,\end{aligned}$$

où $\widetilde{\Pi}_{\mathbf{X}} = \text{IG}(a+n+2, 1+n\overline{X}_n)$. Cette quantité est minimale pour T donné par l'espérance de la loi $\widetilde{\Pi}_{\mathbf{X}}$, soit $\frac{1+n\overline{X}_n}{a+n+1}$.

Une autre façon de faire est de remarquer que le risque a posteriori d'un estimateur T s'écrit

$$\mathbb{E}\left[\ell(\boldsymbol{\theta}, T) \,|\, \mathbf{X}\right] = \mathbb{E}\left[\frac{1}{\boldsymbol{\theta}^2}(T-\boldsymbol{\theta})^2 \,\Big|\, \mathbf{X}\right] = T^2\mathbb{E}\left[\frac{1}{\boldsymbol{\theta}^2} \,\Big|\, \mathbf{X}\right] - 2T\mathbb{E}\left[\frac{1}{\boldsymbol{\theta}} \,\Big|\, \mathbf{X}\right] + 1\,.$$

Le risque a posteriori est donc minimal pour $T^*(\mathbf{X}) = \frac{\mathbb{E}\left[\frac{1}{\boldsymbol{\theta}} \,|\, \mathbf{X}\right]}{\mathbb{E}\left[\frac{1}{\boldsymbol{\theta}^2} \,|\, \mathbf{X}\right]}$. Comme $\frac{1}{\boldsymbol{\theta}} \,|\, \mathbf{X} \sim \text{Gamma}(a+n, 1+n\overline{X}_n)$, on a

$$\mathbb{E}\left[\frac{1}{\boldsymbol{\theta}} \,\Big|\, \mathbf{X}\right] = \frac{a+n}{1+n\overline{X}_n} \qquad \text{et} \qquad \mathbb{E}\left[\frac{1}{\boldsymbol{\theta}^2} \,\Big|\, \mathbf{X}\right] = \frac{(a+n)(a+n+1)}{(1+n\overline{X}_n)^2}\,,$$

et l'on obtient

$$T^*(\mathbf{X}) = \frac{1+n\overline{X}_n}{a+n+1}\,.$$

4. Montrer que pour tout $\theta > 0$, on a

$$\mathbf{R}(\theta, T^*) = \frac{1}{(a+n+1)^2}\left(n + \left(\frac{1}{\theta} - (a+1)\right)^2\right),$$

où $\mathbf{R}$ est la fonction de risque associée à la fonction de perte ℓ.
Correction. Pour $\theta > 0$, on a

$$\begin{aligned}
\mathbf{R}(\theta, T^*) &= \frac{1}{\theta^2}\mathbb{E}_\theta\left[\left(\frac{1+n\overline{X}_n}{a+n+1} - \theta\right)^2\right] \\
&= \frac{1}{\theta^2(a+n+1)^2}\mathbb{E}_\theta\left[\left(n(\overline{X}_n - \theta) + 1 - (a+1)\theta\right)^2\right] \\
&= \frac{1}{\theta^2(a+n+1)^2}\left(n\theta^2 + (1-(a+1)\theta)^2\right) \\
&= \frac{1}{(a+n+1)^2}\left(n + \left(\frac{1}{\theta} - (a+1)\right)^2\right),
\end{aligned}$$

où l'on a utilisé que $\mathbb{E}_\theta[(\overline{X}_n - \theta)^2] = \mathrm{Var}_\theta(\overline{X}_n) = \frac{\theta^2}{n}$.

5. En déduire le risque de Bayes $\mathbf{R}_B(\Pi)$ pour l'a priori Π et la perte ℓ.
Correction. Pour $\boldsymbol{\theta} \sim \mathrm{IG}(a,1)$, soit $\frac{1}{\boldsymbol{\theta}} \sim \mathrm{Gamma}(a,1)$, on a $\mathbb{E}\left[\frac{1}{\boldsymbol{\theta}}\right] = a$ et $\mathrm{Var}\left(\frac{1}{\boldsymbol{\theta}}\right) = a$. Ainsi

$$\mathbb{E}\left[\left(\frac{1}{\boldsymbol{\theta}} - (a+1)\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\boldsymbol{\theta}} - a\right)^2\right] + 1 = \mathrm{Var}\left(\frac{1}{\boldsymbol{\theta}}\right) + 1 = a+1\,.$$

On obtient donc

$$\mathbf{R}_B(\Pi) = \mathbb{E}\left[\mathbf{R}(\boldsymbol{\theta}, T^*)\right] = \frac{1}{a+n+1}\,.$$

6. Pour $\gamma > 0$, on pose $T_\gamma(\mathbf{X}) = \gamma\overline{X}_n$.
   (a) Calculer la fonction de risque de T_γ (pour la perte ℓ). Que vaut $\mathbf{R}_{\max}(T_\gamma)$?
   Correction. Pour $\theta > 0$, on a

$$\begin{aligned}
\mathbf{R}(\theta, T_\gamma) &= \frac{1}{\theta^2}\mathbb{E}_\theta\left[(\gamma\overline{X}_n - \theta)^2\right] \\
&= \frac{1}{\theta^2}\mathbb{E}_\theta\left[\left(\gamma(\overline{X}_n - \theta) + \theta(\gamma-1)\right)^2\right] \\
&= \frac{1}{\theta^2}\left(\gamma^2\mathrm{Var}_\theta(\overline{X}_n) + \theta^2(\gamma-1)^2\right) \\
&= \frac{1}{\theta^2}\left(\frac{\gamma^2\theta^2}{n} + \theta^2(\gamma-1)^2\right) \\
&= \frac{\gamma^2}{n} + (\gamma-1)^2 \\
&= \frac{n+1}{n}\gamma^2 - 2\gamma + 1\,.
\end{aligned}$$

   On obtient une fonction de risque constante. On a donc $\mathbf{R}_{\max}(T_\gamma) = \frac{n+1}{n}\gamma^2 - 2\gamma + 1$
   (b) Déterminer $\gamma^* = \arg\min_{\gamma > 0} \mathbf{R}_{\max}(T_\gamma)$.
   Correction. On trouve $\gamma^* = \frac{n}{n+1}$.

(c) Montrer que T_{γ^*} est minimax.

Correction. On trouve que $\mathbf{R}_{\max}(T_{\gamma^*}) = \frac{1}{n+1}$, qui correspond à la limite du risque de Bayes pour l'a priori $\mathrm{IG}(a,1)$ quand $a \to 0$. Par un résultat du cours, l'estimateur T_{γ^*} est minimax.

7. Soit $\theta_0 > 0$.

(a) Montrer que, sous $P_{\theta_0}^{\otimes n}$, on a

$$n v_{\mathbf{X}} \xrightarrow{\mathbb{P}} \theta_0^2 \,.$$

Correction. Par la loi des grands nombres, $\overline{X}_n \xrightarrow{\mathbb{P}} \theta_0$ sous $P_{\theta_0}^{\otimes n}$. Ainsi, en divisant par n^3 le numérateur et le dénominateur de $n v_{\mathbf{X}}$ et en utilisant le théorème de continuité, on a

$$n v_{\mathbf{X}} = \frac{\left(\frac{1}{n} + \overline{X}_n\right)^2}{\left(1 + \frac{a-1}{n}\right)^2 \left(1 + \frac{a-2}{n}\right)} \xrightarrow{\mathbb{P}} \theta_0^2 \,.$$

(b) Montrer que, sous $P_{\theta_0}^{\otimes n}$, on a

$$\sqrt{n}\,(m_{\mathbf{X}} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta_0^2) \,.$$

Correction. On a

$$\sqrt{n}(m_{\mathbf{X}} - \theta_0) = \sqrt{n}\left(\frac{1 + n\overline{X}_n}{n + a - 1} - \theta_0\right) = \sqrt{n}\frac{n}{n+a-1}\left(\overline{X}_n - \theta_0\right) + \sqrt{n}\frac{1 - (a-1)\theta_0}{n} \,.$$

Or sous $P_{\theta_0}^{\otimes n}$, on a $\sqrt{n}(\overline{X}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta_0^2)$. Comme $\frac{n}{n+a-1} \underset{n\to\infty}{\longrightarrow} 1$, et $\sqrt{n}\frac{1-(a-1)\theta_0}{n} \underset{n\to\infty}{\longrightarrow} 0$, le lemme de Slutsky donne $\sqrt{n}(m_{\mathbf{X}} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta_0^2)$.

(c) Montrer que pour toute suite $(M_n)_{n\geq 1}$ avec $M_n \to +\infty$, on a, sous $P_{\theta_0}^{\otimes n}$:

$$\mathbb{P}\left(|\boldsymbol{\theta} - \theta_0| > \frac{M_n}{\sqrt{n}} \,\middle|\, \mathbf{X}\right) \xrightarrow{\mathbb{P}} 0 \,.$$

Correction. Soit (M_n) une suite tendant vers $+\infty$ quand $n \to +\infty$. Par l'inégalité triangulaire et une borne union, on a

$$\mathbb{P}\left(|\boldsymbol{\theta} - \theta_0| > \frac{M_n}{\sqrt{n}} \,\middle|\, \mathbf{X}\right) \leq \mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| > \frac{M_n}{2\sqrt{n}} \,\middle|\, \mathbf{X}\right) + \mathbb{1}_{\left\{|m_{\mathbf{X}} - \theta_0| > \frac{M_n}{2\sqrt{n}}\right\}} \,.$$

Pour le premier terme, on utilise l'inégalité de Tchebytchev :

$$\mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| > \frac{M_n}{2\sqrt{n}} \,\middle|\, \mathbf{X}\right) \leq \frac{4 n v_{\mathbf{X}}}{M_n^2} \xrightarrow{\mathbb{P}} 0 \,,$$

car $n v_{\mathbf{X}} \xrightarrow{\mathbb{P}} \theta_0^2$ et $M_n \to +\infty$. Pour le deuxième terme, on sait que $\sqrt{n}(m_{\mathbf{X}} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta_0^2)$, donc $\frac{\sqrt{n}|m_{\mathbf{X}} - \theta_0|}{M_n} \xrightarrow{\mathbb{P}} 0$. Ainsi

$$\mathbb{P}_{\theta_0}\left(\frac{\sqrt{n}|m_{\mathbf{X}} - \theta_0|}{M_n} > \frac{1}{2}\right) \underset{n\to\infty}{\longrightarrow} 0 \,,$$

ce qui équivaut à $\mathbb{1}_{\left\{|m_{\mathbf{X}} - \theta_0| > \frac{M_n}{2\sqrt{n}}\right\}} \xrightarrow{\mathbb{P}} 0$. On obtient bien le résultat voulu.

Exercice 2 (7,5 points)

On considère une variable aléatoire X de densité f_θ définie par

$$f_\theta(x) = \frac{3x^2}{\theta^3}\mathbf{1}_{[0,\theta]}(x),$$

avec $\theta > 0$ un paramètre inconnu que l'on souhaite estimer. Pour ce faire, on dispose de variables aléatoires i.i.d. $X_1, \dots, X_n$ de même loi que X.

1. Le modèle est-il régulier ?
   Correction. Pour tout $x \geq 0$, la fonction $\theta \mapsto f_\theta(x)$ est discontinue en $\theta = x$, donc le modèle n'est pas régulier.
2. Calculer et représenter la fonction de répartition F de la variable aléatoire X.
   Correction. La fonction de répartition est
   $$F(x) = \left(\frac{x}{\theta}\right)^3 \mathbf{1}_{[0,\theta[}(x) + \mathbf{1}_{[\theta,\infty[}(x).$$
3. Déterminer la médiane de la loi de X et en déduire un estimateur $\widehat{\theta}_n$ de θ. Préciser sa normalité asymptotique. Pour $0 < \alpha < 1$, donner un intervalle de confiance de niveau asymptotique $(1 - \alpha)$.
   Correction. La médiane est donc $x_{1/2} = 2^{-1/3}\theta$ d'où l'estimateur $\widehat{\theta}_n = 2^{1/3}x_{1/2}(n)$ où $x_{1/2}(n)$ désigne comme d'habitude la médiane empirique. Puisque $f_\theta(x_{1/2}) = 3/(2^{2/3}\theta) > 0$, on sait que
   $$\sqrt{n}\left(x_{1/2}(n) - x_{1/2}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 2^{-2/3}\theta^2/9)$$
   donc
   $$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \theta^2/9).$$
   On en déduit que
   $$3\sqrt{n}\left(\frac{\widehat{\theta}_n}{\theta} - 1\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$
   d'où l'intervalle de confiance pour θ de niveau asymptotique $(1 - \alpha)$:
   $$\left[\frac{\widehat{\theta}_n}{1 + \Phi^{-1}(1 - \alpha/2)/(3\sqrt{n})} ; \frac{\widehat{\theta}_n}{1 - \Phi^{-1}(1 - \alpha/2)/(3\sqrt{n})}\right],$$
   avec Φ la fonction de répartition de $\mathcal{N}(0, 1)$.
4. Déterminer l'estimateur du maximum de vraisemblance.
   Correction. Pour tout θ, la vraisemblance s'écrit
   $$L_n(\theta) = \frac{3^n \prod_{i=1}^n X_i^2}{\theta^{3n}}\mathbf{1}_{[X_{(n)},\infty[}(\theta),$$
   fonction nulle à gauche de $X_{(n)}$ puis strictement décroissante donc l'estimateur du maximum de vraisemblance est $X_{(n)}$.
5. Soit $X_{(n)} = \max_{1\leq i\leq n} X_i$. Construire un intervalle de confiance de niveau $(1 - \alpha)$ de la forme $[X_{(n)}; c_{\alpha,n}X_{(n)}]$, où $c_{\alpha,n} \geq 1$ est une constante que l'on précisera.
   Correction. Puisque $X_{(n)} \leq \theta$, ceci revient à trouver $c_{\alpha,n} \geq 1$ tel que $\mathbb{P}(c_{\alpha,n}X_{(n)} < \theta) = \alpha$, or
   $$\mathbb{P}(X_{(n)} < \theta/c_{\alpha,n}) = F(\theta/c_{\alpha,n})^n = \left(\frac{1}{c_{\alpha,n}}\right)^{3n},$$
   ce qui donne
   $$c_{\alpha,n} = \alpha^{-1/3n}.$$
   On vérifie au passage qu'on a bien $c_{\alpha,n} > 1$ pour tout $0 < \alpha < 1$.

6. Déduire de la question précédente un test au niveau α pour les hypothèses :

$$H_0 : \theta = 1 \quad \text{contre} \quad H_1 : \theta \neq 1.$$

Soit $x_{(n)}$ une réalisation de $X_{(n)}$. Calculer la p-valeur associée.
Correction. Le lien entre intervalle de confiance et test assure que

$$T(\mathbf{X}) := \mathbf{1}_{1\notin[X_{(n)};c_{\alpha,n}X_{(n)}]} = \mathbf{1}_{1\notin[X_{(n)};\alpha^{-1/3n}X_{(n)}]}$$

est un test de niveau α. Maintenant, si $x_{(n)}$ est une réalisation de $X_{(n)}$, la p-value associée est donc définie par

$$\alpha_0(\mathbf{x}) = \inf\left\{\alpha \in [0,1], 1 \notin [x_{(n)}; \alpha^{-1/3n}x_{(n)}]\right\}.$$

Si $1 \leq x_{(n)}$ alors clairement $\alpha_0(\mathbf{x}) = 0$. Si $x_{(n)} < 1$, alors $\alpha^{-1/3n}x_{(n)} < 1$ si et seulement si $\alpha \geq x_{(n)}^{3n}$, donc $\alpha_0(\mathbf{x}) = x_{(n)}^{3n}$.

7. Montrer que $n(\theta - X_{(n)})$ tend en loi vers une loi exponentielle dont on précisera le paramètre.
Correction. Les variables $n(\theta - X_{(n)})$ sont positives. Pour tout réel $t \geq 0$ fixé, il suffit alors d'écrire

$$\mathbb{P}\left(n(\theta - X_{(n)}) \leq t\right) = 1 - F(\theta - t/n)^n$$

avec

$$F(\theta - t/n)^n = \left(1 - \frac{t}{n\theta}\right)^{3n} \mathbf{1}_{[0,\theta[}(\theta - t/n) \xrightarrow[n\to\infty]{} e^{-3t/\theta}$$

ce qui montre que

$$n(\theta - X_{(n)}) \xrightarrow[n\to\infty]{d} \mathcal{E}(3/\theta).$$

8. Soient $\lambda > 0$ et $T \sim \mathcal{E}\left(\frac{\lambda}{\theta}\right)$. Déterminer $q_{\alpha,\theta}$ tel que $\mathbb{P}(0 \leq T \leq q_{\alpha,\theta}) = 1 - \alpha$. Déduire de ceci et de la question précédente un intervalle de confiance de niveau asymptotique $(1-\alpha)$ pour θ.
On prendra pour convention $[A, B] = \emptyset$ lorsque $B < A$ et on vérifiera que l'intervalle de confiance obtenu est de la forme $[X_{(n)}; d_{\alpha,n}X_{(n)}]$, où $d_{\alpha,n}$ est une constante (supérieure à 1 à partir d'un certain rang) que l'on précisera.
Correction. Par un calcul classique, on obtient $q_{\alpha,\theta} = -\frac{\theta}{\lambda}\log\alpha$. Or on sait que

$$n\left(\theta - X_{(n)}\right) \xrightarrow[n\to\infty]{d} \mathcal{E}(3/\theta)$$

donc

$$\mathbb{P}\left(0 \leq n\left(\theta - X_{(n)}\right) \leq -\frac{\theta}{3}\log\alpha\right) \xrightarrow[n\to\infty]{} 1 - \alpha$$

ou encore

$$\mathbb{P}\left(X_{(n)} \leq \theta \leq \frac{X_{(n)}}{1 + (\log\alpha)/(3n)}\right) \xrightarrow[n\to\infty]{} 1 - \alpha$$

ce qui est le résultat voulu en posant

$$d_{\alpha,n} := \frac{1}{1 + (\log\alpha)/(3n)}.$$

9. Pour $\alpha \in]0,1[$ fixé et lorsque n tend vers l'infini, donner des développements asymptotiques de la forme $c_{\alpha,n} \sim 1 + c_\alpha/n$ et $d_{\alpha,n} \sim 1 + d_\alpha/n$. Commenter.
Correction. On a d'une part

$$c_{\alpha,n} = \alpha^{-1/3n} = e^{-\frac{1}{3n}\log\alpha} \sim 1 - \frac{\log\alpha}{3n}$$

et d'autre part

$$d_{\alpha,n} = \frac{1}{1 + (\log\alpha)/(3n)} \sim 1 - \frac{\log\alpha}{3n}.$$

Sans surprise, on constate que $c_\alpha = d_\alpha = (-\log\alpha)/3$, qui n'est rien d'autre que le quantile d'ordre $(1-\alpha)$ d'une loi $\mathcal{E}(3)$. En effet, notons

$$c_{\alpha,n} = 1 + \frac{c_\alpha}{n} + \frac{\varepsilon_n}{n}$$

avec (ε_n) qui tend vers 0 lorsque n tend vers l'infini. Puisque $\theta \geq X_{(n)}$, on a pour tout n

$$1 - \alpha = \mathbb{P}(X_{(n)} \leq \theta \leq c_{\alpha,n} X_{(n)}) = \mathbb{P}(\theta \leq c_{\alpha,n} X_{(n)}) = \mathbb{P}\left(\frac{n(\theta - X_{(n)})}{X_{(n)}} - \varepsilon_n \leq c_\alpha\right)$$

Or on a vu ci-dessus que

$$n\left(\theta - X_{(n)}\right) \xrightarrow[n\to\infty]{d} \mathcal{E}(3/\theta)$$

ce qui implique en particulier que $X_{(n)}$ tend en probabilité vers θ, donc le Lemme de Slutsky et le fait que (ε_n) tende vers 0 impliquent

$$\frac{n(\theta - X_{(n)})}{X_{(n)}} - \varepsilon_n \xrightarrow[n\to\infty]{d} \mathcal{E}(3).$$

La fonction de répartition de cette loi limite est continue sur $\mathbb{R}$ donc

$$1 - \alpha = \mathbb{P}\left(\frac{n(\theta - X_{(n)})}{X_{(n)}} - \varepsilon_n \leq c_\alpha\right) \xrightarrow[n\to\infty]{} \mathbb{P}(\mathcal{E}(3) \leq c_\alpha).$$

Mais si l'on applique la même démarche pour

$$d_{\alpha,n} = 1 + \frac{d_\alpha}{n} + \frac{\delta_n}{n}$$

il vient d'une part

$$\mathbb{P}\left(\frac{n(\theta - X_{(n)})}{X_{(n)}} - \delta_n \leq d_\alpha\right) \xrightarrow[n\to\infty]{} \mathbb{P}(\mathcal{E}(3) \leq d_\alpha)$$

sachant d'autre part que, par la question précédente,

$$\mathbb{P}\left(\frac{n(\theta - X_{(n)})}{X_{(n)}} - \delta_n \leq d_\alpha\right) \xrightarrow[n\to\infty]{} 1 - \alpha$$

ce qui prouve bien que $c_\alpha = d_\alpha = (-\log\alpha)/3$.

10. Entre $X_{(n)}$ et $\widehat{\theta}_n$, quel estimateur choisissez-vous ?
Correction. On choisit bien entendu $X_{(n)}$ puisqu'il converge à vitesse $1/n$ vers θ tandis que $\widehat{\theta}_n$ ne converge qu'à vitesse $1/\sqrt{n}$.

Exercice 3 (3,5 points)

Les questions 1. et 2. de cet exercice sont indépendantes.

1. On dispose du tableau de données suivant, représentant des réalisations $x_1, \dots, x_n$ de variables aléatoires réelles $X_1, \dots, X_n$ i.i.d., et on souhaite tester (au niveau $\alpha \in]0, 1[$)

$$H_0 : \theta \leq 1 \qquad \text{contre} \qquad H_1 : \theta > 1$$

dans le modèle $\mathcal{P} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$.

```
print(x.shape)
print(x[:10])
```

```
(200,)
[ 3.25617103 -1.24290386  0.21805856 -0.72901134  2.81435003  2.29702943
  1.68190239  1.17001241 -0.76642552  1.7031285 ]
```

(a) On note Φ la fonction de répartition de $\mathcal{N}(0, 1)$. Par quoi faut-il remplacer `???` et `$$$` pour calculer la réalisation du test $T(x) = \mathbb{1}_{\overline{x}_n > 1 + \frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha)}$ (avec $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$) et la p-valeur associée $\alpha_0(x) = 1 - \Phi(\sqrt{n}(\overline{x}_n - 1))$?

```
alpha = 0.05
threshold = 1 + stats.norm.ppf(1-alpha) / np.sqrt(n)
test = ??? > threshold
alpha0 = 1 - $$$
print(test, alpha0)
```

```
False 0.4135433062246249
```

Correction. `???` : `x.mean()` et `$$$` : `stats.norm.cdf(np.sqrt(n) * (x.mean() - 1))`

(b) Au vu du résultat, rejette-t-on H_0 ?
Correction. On ne rejette pas H_0 au niveau α, d'une part parce que `test` vaut `False`, d'autre part parce que la p-valeur est plus grande que α.

2. (a) Expliquer en quoi le script suivant échantillonne approximativement suivant la loi de densité proportionnelle à la fonction $f : (x, y) \mapsto y^2\, \mathrm{e}^{-(1+x)y} \mathbb{1}_{x \geq 0,\, y \geq 0}$. On détaillera la réponse.

```
n = 5000

x, y = 0, 1
sample = [(x, y)]
for i in range(n):
    x = stats.expon(scale=1/y).rvs()
    y = stats.gamma(a=3, scale=1/(1+x)).rvs()
    sample.append((x, y))
sample = np.asarray(sample)
```

Correction. Dans l'itération, x est une réalisation d'une loi qui dépend de y, et y est une réalisation d'une loi qui dépend de x. Cela conduit naturellement à penser à un échantillonneur de Gibbs, qui produit bien une trajectoire $((X_1, Y_1), \dots, (X_n, Y_n))$ pouvant être vue approximativement comme un échantillon de loi F, correspondant à la loi stationnaire de la chaîne de Markov créée.

Dans ce context, en prenant $(X, Y) \sim F$ un couple aléatoire de loi cible voulue, l'algorithme de Gibbs correspond (étant donné une initialisation (x_0, y_0)) à répéter :

— construire une réalisation x_{n+1} de la loi de $X \mid Y = y_n$;

— construire une réalisation y_{n+1} de la loi de $Y \mid X = y_{n+1}$.

Pour répondre à la question, il suffit donc de vérifier que les lois conditionnelles de l'itération et de la loi de densité proportionnelle à f coïncident.

D'un côté, l'itération suggère que $X \mid Y = y \sim \mathcal{E}(y)$ et $Y \mid X = x \sim \text{Gamma}(3, 1 + x)$ (on a bien fait attention au fait que, pour les deux lois, le paramètre de forme correspond à l'inverse du paramètre d'échelle).

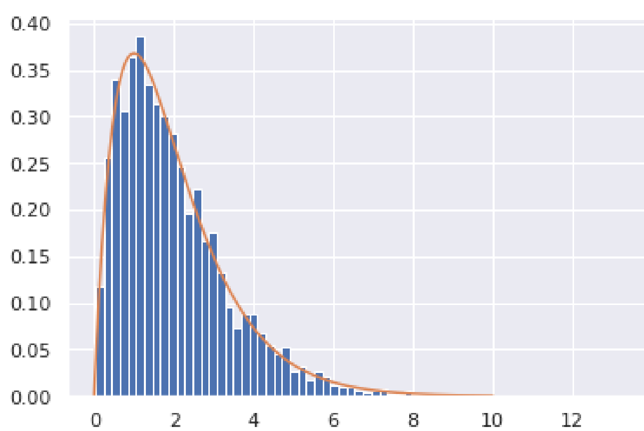
De l'autre côté, $f(x, y) \propto_x \mathrm{e}^{-yx} \mathbb{1}_{x \geq 0}$, qui correspond au terme général de la densité de $\mathcal{E}(y)$, et $f(x, y) \propto_y y^{3-1} \mathrm{e}^{-(1+x)y} \mathbb{1}_{y \geq 0}$, qui correspond au terme général de la densité de $\text{Gamma}(3, 1+x)$. Ces deux lois conditionnelles correspondent à celles utilisées dans le script présenté.

On en conclut que la loi de densité proportionnelle à f est bien stationnaire pour la chaîne de Markov ainsi créée et `sample` peut être vu approximativement comme un tableau de réalisations i.i.d. de densité proportionnelle à f.

(b) Justifier l'illustration ci-dessous.

```
u = np.linspace(0, 10, num=200)

plt.hist(sample[:, 1], density=True, bins='auto')
plt.plot(u, stats.gamma(a=2, scale=1).pdf(u))
```



Correction. On vérifie par intégration que si (X, Y) suit la loi de densité proportionnel à f, alors $Y \sim \text{Gamma}(2, 1)$. Le script donné compare donc bien un estimateur de la densité marginale des Y_i à la loi attendue, $\text{Gamma}(2, 1)$.