MIDO - L3 Math. Appliquées 2020–2021 **Examen partiel du 29 octobre 2020**
Statistical modelling **Durée 2h00 – Documents interdits**

Exercise 1 (8.5 points, les questions peuvent être résolues indépendamment)

On réalise une étude de qualité sur une chaîne de production de composants électroniques utilisés dans le secteur de l'aéronautique. Pour cela, on compte le nombre de composants défectueux dans des lots pris au hasard. Les résultats de l'expérience sont sauvegardés dans un fichier `dataset.csv` dont les premières lignes sont données ci-dessous (la colonne `Defective` correspond aux nombres de composants défectueux du lot indiqué dans la colonne `Reference`) :

```
Reference,Defective
QcP-s-1120,2
QcP-s-9120,1
QcP-s-2220,3
```

1. (0.5 pt) Écrire le code R qui permet de charger, à partir de ce fichier, les données concernant le nombre de pièces défectueuses dans un vecteur `x`.

```r
# Option n°1
x <- read.table("dataset.csv", header = TRUE, sep = ",")$Defective
# Option n°2
x <- read.csv("dataset.csv", header = TRUE)$Defective
```

2. (0.5 pt) Parmi les codes ci-dessous, donner, en justifiant, celui qui permet de tracer la distribution de l'échantillon contenu dans `x`.

(a) `hist(table(x) / length(x), freq = FALSE)`
(b) `barplot(x)`
(c) `hist(x, freq = FALSE)`
(d) `barplot(table(x) / length(x))`

(d) La fonction `barplot` est utilisée pour représenter la distribution de variables aléatoires discrètes. Pour un échantillon observé, elle prend en entrée la fréquence de chaque issue observée que l'on obtient avec `table` qui renvoie le nombre d'occurences de chaque issue.

On considère le modèle statistique à support $\{1, 2, 3, 4\}$

$$\mathcal{P} = \left\{ \mathbb{P}_\theta = \theta^2\delta_1 + (1-\theta)^2\delta_2 + \theta(1-\theta)(\delta_3 + \delta_4) \;\middle|\; \theta \in]0;1[\right\}.$$

3. (0.5 pt) Écrire le code R d'une fonction `rgen(`n`, `θ`)` qui simule n réalisations indépendantes suivant $\mathbb{P}_\theta$.

```r
rgen <- function(n, theta) {
  p <- c(theta^2, (1 - theta)^2, rep(theta * (1 - theta), 2))
  return(sample(1:4, n, TRUE, p))
}
```

On note $X_1, \dots, X_n$ des variables aléatoires indépendantes et identiquements distribuée de loi $\mathbb{P}_\theta$. On considère les estimateurs

$$\hat{\delta}_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k=1\}}} \quad \text{et} \quad \hat{\theta}_n = g(X_1, \dots, X_n), \quad \text{où la fonction } g \text{ est donnée par le code :}$$

```
g <- function(x) { return(0.5 * (1 + mean(x == 1) - mean(x == 2))) }
```

4. (0.5 pt) Déduire du code ci-dessus l'expression de $\hat{\theta}_n$.

$$\hat{\theta}_n = \frac{1}{2}\left(1 + \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k=1\}} - \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k=2\}}\right).$$

5. (1.5 pt) Montrer que les estimateurs $\hat{\delta}_n$ et $\hat{\theta}_n$ convergent vers θ en probabilité.

Pour $j \in \{1, 2\}$, la suite $\left(\mathbb{1}_{\{X_n=j\}}\right)_{n\geq 1}$ est une suite de variables aléatoires *i.i.d.* comme transformation mesurable de variables aléatoires *i.i.d.*. De plus, $\mathbb{E}_\theta\left[\mathbb{1}_{\{X_1=j\}}\right] = \mathbb{P}_\theta\left[X_1 = j\right] < \infty$. La loi faible des grands nombres conduit alors à

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k=j\}} \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{P}_\theta[X_1 = j].$$

La fonction $x \mapsto \sqrt{x}$ est continue sur $\mathbb{R}_+$, donc

$$\hat{\delta}_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k=1\}}} \xrightarrow[n\to+\infty]{\mathbb{P}} \sqrt{\mathbb{P}_\theta[X_1 = 1]} = \sqrt{\theta^2} = \theta \quad (\text{car } \theta > 0).$$

La fonction $(x, y) \mapsto 0.5(1 + x - y)$ étant continue sur $\mathbb{R}^2$, on a

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{2}\left(1 + \mathbb{P}_\theta[X_1 = 1] - \mathbb{P}_\theta[X_k = 2]\right) = \frac{1}{2}\left[1 + \theta^2 - (1-\theta)^2\right] = \theta.$$

6. (1 pt) Rappeler la loi de $\sum_{k=1}^{n} \mathbb{1}_{\{X_k=1\}}$ et donner le code `R` correspondant à une estimation bootstrap paramétrique du biais de $\hat{\delta}_n$. On notera k le nombre d'échantillons bootstrap utilisés.
Bonus (0.5 pt) : code sans boucle `for`.

$\sum_{k=1}^{n} \mathbb{1}_{\{X_k=1\}}$ correspond à la somme de variables aléatoires indépendantes de loi de Bernoulli de paramètre $\mathbb{P}_\theta[X_1 = 1] = \theta^2$. Cette somme suit donc une loi binomiale de paramètre (n, θ^2). On déduit que dans une version paramétrique, on peut obtenir des réalisations de $\hat{\delta}_n$ à l'aide du générateur de la loi binomiale.

```
k <- 100
delta_n <- sqrt(mean(x == 1))
bias <- mean(sqrt(rbinom(k, length(x), delta_n^2) / length(x))) - delta_n
```

7. (2 pts) Trouver une suite c_n indépendante de θ telle que

$$c_n \left(\hat{\delta}_n - \theta\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

En déduire un intervalle de confiance asymptotique pour θ au niveau $1-\alpha$.

La suite $\left(\mathbb{1}_{\{X_n=j\}}\right)_{n\geq 1}$ est une suite de variables aléatoires *i.i.d.* de variance finie

$$\mathbb{V}\mathrm{ar}_\theta\left[\mathbb{1}_{\{X_n=j\}}\right] = \mathbb{P}_\theta[X_1=1]\left(1-\mathbb{P}_\theta[X_1=1]\right) = \theta^2(1-\theta^2).$$

Le théorème central limite s'écrit

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{\{X_k=1\}} - \mathbb{E}_\theta\left[\mathbb{1}_{\{X_1=1\}}\right]\right) = \sqrt{n}\left(\hat{\delta}_n^2-\theta^2\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0,\theta^2(1-\theta^2)\right).$$

La fonction $g : x \mapsto \sqrt{x}$ est dérivable sur $]0,1[$ et $g'(\theta^2) = (2\theta)^{-1} \neq 0$. La méthode delta conduit à

$$\sqrt{n}\left[\hat{\delta}_n - \theta\right] \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0,\theta^2(1-\theta^2)g'(\theta^2)^2\right) \equiv \mathcal{N}\left(0,\frac{1-\theta^2}{4}\right).$$

La fonction $x \mapsto 2(1-x^2)^{-\frac{1}{2}}$ est continue sur $]0,1[$. D'après la question 5, on a donc

$$\frac{2}{\sqrt{1-\hat{\delta}_n^2}} \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{2}{\sqrt{1-\theta^2}}.$$

Le théorème de Slutsky permet alors de conclure que

$$\frac{2\sqrt{n}}{\sqrt{1-\hat{\delta}_n^2}}\left[\hat{\delta}_n-\theta\right] \xrightarrow[n\to+\infty]{d} \frac{2}{\sqrt{1-\theta^2}}\mathcal{N}\left(0,\frac{1-\theta^2}{4}\right) \equiv \mathcal{N}(0,1).$$

On en déduit c_n et l'intervalle de confiance asymptotique pour θ au niveau $1-\alpha$

$$c_n = \frac{2\sqrt{n}}{\sqrt{1-\hat{\delta}_n^2}} \quad \text{et} \quad \left[\hat{\delta}_n - q_{1-\alpha/2}\frac{1}{c_n}, \hat{\delta}_n + q_{1-\alpha/2}\frac{1}{c_n}\right].$$

8. (2 pts) Trouver une fonction h inversible $]0;1[$ telle que

$$\sqrt{n}\left[h\left(\hat{\theta}_n\right) - h\left(\theta\right)\right] \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

En déduire un intervalle de confiance asymptotique pour θ au niveau $1-\alpha$.

Posons pour $n \in \mathbb{N}^*$,

$$Y_n = \frac{1}{2}\left(1+\mathbb{1}_{\{X_n=1\}} - \mathbb{1}_{\{X_n=2\}}\right),$$

et remarquons que $\hat{\theta}_n = \bar{Y}_n$.
$(Y_n)_{n\geq 1}$ est une suite de variables aléatoires *i.i.d* comme transformée mesurables de variables aléatoires *i.i.d* avec pour tout $n \in \mathbb{N}^*$

$$Y_n = \begin{cases} 0 & \text{avec probabilité } (1-\theta)^2 \\ 0.5 & \text{avec probabilité } 2\theta(1-\theta) \\ 1 & \text{avec probabilité } \theta^2. \end{cases}$$

On en déduit $\mathbb{E}_\theta[Y_1] = \theta$ et $\mathbb{V}_\theta\text{ar}[Y_1] = 0.5\theta(1-\theta)$. Il en résulte d'après le théorème central limite

$$\sqrt{n}\left(\bar{Y}_n - \mathbb{E}_\theta[Y_1]\right) = \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0, 0.5\theta(1-\theta)\right).$$

Supposons que $h'(\theta) \neq 0$. La méthode delta s'écrit

$$\sqrt{n}\left[h(\hat{\theta}_n) - h(\theta)\right] \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0, 0.5\theta(1-\theta)h'(\theta)^2\right).$$

La fonction h cherchée est donc solution de

$$0.5x(1-x)h'(x)^2 = 1 \quad \Rightarrow \quad h'(x) = \frac{\sqrt{2}}{\sqrt{x(1-x)}}.$$

Pour $a \in]0,1[$ et pour tout $t \in]0,1[$

$$h(t) - h(a) = \int_a^t \frac{\sqrt{2}}{\sqrt{x(1-x)}}\mathrm{d}x = \int_a^t \frac{2\sqrt{2}}{\sqrt{1-(2x-1)^2}}\mathrm{d}x = \sqrt{2}\arcsin(2t-1) - \sqrt{2}\arcsin(2a-1).$$

On en déduit $h : x \mapsto \sqrt{2}\arcsin(2x-1)$, qui est bien inversible sur $]0,1[$ et telle que pour $x \in]0,1[$, $h'(x) \neq 0$ (donc la méthode delta appliquée à h est valide).
On a $h^{-1} : x \mapsto 0.5\sin\left(x/\sqrt{2}\right) + 0.5$ et donc un intervalle de confiance asymptotique pour θ au niveau $1-\alpha$

$$h^{-1}\left(\left[h\left(\hat{\theta}_n\right) - q_{1-\alpha/2}\frac{1}{\sqrt{n}}, h\left(\hat{\delta}_n\right) + q_{1-\alpha/2}\frac{1}{\sqrt{n}}\right]\right).$$

Exercice 2 (4 points) *Dans cet exercice il vous est demandé de donner la ou les bonnes réponses, seules les réponses justifiées seront validées. Il n'y a pas de points négatifs.*

1. Si `x` est un échantillon continu de taille `n = 13` et si on étudie la variabilité de la médiane de l'échantillon par bootstrap, le nombre de valeurs possibles d'une réalisation bootstrap de cette médiane vaut

(a) `8.9161e+12` (b) `479001600` (c) `1352078` (d) `1` (e) `13`

(e). L'échantillon initial a 13 valeurs distinctes (puisqu'issues d'une variable aléatoire à valeurs continues). L'échantillon bootstrap ne contient que des valeurs parmis ces 13 valeurs initiales. La médiane ne peut donc prendre que l'une de ces 13 valeurs initiales.

2. Si X dénote une variable aléatoire de la loi de Poisson de paramètre 2, quelle commande R retourne $\mathbb{P}[X=4]$?

(a) `ppois(2, 4)` (c) `dpois(2,4)` (e) `ppois(4,2)`
(b) `dpois(4,2)` (d) `rpois(4,2)` (f) `qpois(4,2)`

(b). $\mathbb{P}[X=4]$ est la valeur de la densité de la loi de Poisson (densité définie par rapport à la mesure de comptage), d'où le prefixe `d`. Cette fonction prend en premier argument le point où l'on souhaite évaluer la densité et en second argument le paramètre de la loi.

4. Pour construire une variable aléatoire X suivant une loi exponentielle $\mathcal{E}xp(\lambda)$, avec $\lambda > 0$, à partir d'une variable uniforme $\mathcal{U}(0,1)$, U, on peut utiliser la transformation

(a) $X = \lambda\log(U)$ (c) $X = -\log(1-U)/\lambda$ (e) $X = -\log(U/\lambda)$
(b) $X = -\log[(1-U)]^\lambda$ (d) $X = \log[\lambda(1-U)]$ (f) $X = \exp\{-\log(U)^\lambda\}$

(c). Pour une variable absolument continue, si U suit une loi uniforme sur $]0,1[$ alors $F^{-1}(U)$ suit la loi de fonction de répartition F. Dans le cas de l'exponentielle, la fonction de répartition s'écrit $1-\exp(-\lambda x)$, que l'on peut facilement inverser en $-\log(1-u)/\lambda$.

5. La loi de Laplace a pour densité $f(x|\mu,\sigma) = \frac{1}{2\sigma}\exp(-|x-\mu|/\sigma)$, $\sigma > 0$. Quelle affirmation est exacte ? La loi de Laplace

(a) à μ connu forme une famille exponentielle de paramètre naturel $\tau(\sigma) = -\frac{1}{\sigma}$.

(b) forme une famille exponentielle de paramètre naturel $\tau(\mu,\sigma) = (-\mu, \sigma^2)$.

(c) forme une famille exponentielle de paramètre naturel $\tau(\mu,\sigma) = (-\mu/\sigma, \sigma^2)$.

(d) ne forme pas une famille exponentielle

(a). À μ connu, on a une famille exponentielle de dimension 1 et de statistique naturelle $|x-\mu|$.
(d). Lorsque μ n'est pas connu, $|x-\mu|$ ne permet pas de séparer la dépendance entre μ et x.

Exercice 3 (10 points) La loi inverse gaussienne $IG(\mu,\lambda)$ est définie par la densité

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right]\mathbb{1}_{\{x\in\mathbb{R}_+^*\}}, \quad \text{avec} \quad \mu > 0 \quad \text{et} \quad \lambda > 0.$$

Dans le cas particulier où $\lambda = \mu^2$, on parle de variable inverse gaussienne à un seul paramètre $IG(\mu,\mu^2)$.

Formulaire : dans cet exercice, on pourra utiliser, sans la démontrer, l'identité suivante

$$\int_0^{+\infty} x^{-3/2}\exp\left[-\frac{1}{2}\left(ax+\frac{b}{x}\right)\right]\mathrm{d}x = 2\left(\frac{a}{b}\right)^{\frac{1}{4}} K(\sqrt{ab}), \quad \textit{avec, pour } x\in R_+^*, \quad K(x) = \sqrt{\frac{\pi}{2x}}\exp(-x).$$

1. (2 pts) Montrer que la loi $IG(\mu,\lambda)$ constitue une famille exponentielle de dimension deux. Donner un paramètre naturel, une statistique naturelle, et démontrer que la représentation est minimale et régulière.

On peut réécrire la densité sous la forme :

$$f(x;\mu,\lambda) = \sqrt{\lambda}\frac{1}{\sqrt{2\pi x^3}}\exp\left(-\frac{\lambda x}{2\mu^2}+\frac{\lambda}{\mu}-\frac{\lambda}{2x}\right)\mathbb{1}_{\{x\in\mathbb{R}_+^*\}} = \sqrt{\frac{\lambda}{2\pi}}\exp\left(\frac{\lambda}{\mu}\right)\frac{\mathbb{1}_{\{x\in\mathbb{R}_+^*\}}}{\sqrt{x^3}}\exp\left(\langle\begin{pmatrix}x\\ \frac{1}{x}\end{pmatrix},\begin{pmatrix}-\frac{\lambda}{2\mu^2}\\ -\frac{\lambda}{2}\end{pmatrix}\rangle\right)$$

C'est à dire que cette distribution constitue une famille exponentielle avec pour statistique naturelle $T(x) = (x, 1/x)^T$ et pour paramètre naturel $\eta(\mu,\lambda) = (-\lambda/(2\mu^2), -\lambda/2)^T$.

Soit $(\alpha_1,\alpha_2) \in \mathbb{R}^2 \setminus \{(0,0)\}$ et b une constante. $\{x \in \mathbb{R}_+^* \mid \alpha_1 x + \alpha_2/x = b\}$ contient au plus deux points (nombres de solutions maximales du polynôme de degré 2). Cet ensemble est donc de mesure nulle par rapport à la mesure de Lebesgue (mesure par rapport laquelle la densité est définie). La famille est donc minimale.

L'espace naturel des paramètres est

$$\Theta = \left\{(\theta_1,\theta_2)\in\mathbb{R}^2 \,\middle|\, \int_{\mathbb{R}_+^*} x^{-3/2}\exp\left(\theta_1 x + \frac{\theta_2}{x}\right) < \infty.\right\}$$

L'intégrande est localement intégrable en 0 si $\theta_2 < 0$ et localement intégrable en $+\infty$ pour $\theta_1 < 0$ (croissances comparées). D'où $\Theta = \mathbb{R}_-^* \times \mathbb{R}_-^*$. L'espace est donc ouvert et la famille est régulière.

2. (0.5 pt) Dans le cas d'une loi inverse gaussienne à un seul paramètre, donner la dimension de la famille exponentielle correspondante et une statistique naturelle.

Dans le cas d'une famille à un paramètre, $\lambda = \mu^2$, et on en déduit

$$f(x;\mu,\mu^2) = \frac{\mu}{\sqrt{2\pi}}\exp(\mu)\frac{\mathbb{1}_{\{x\in\mathbb{R}_+^*\}}}{\sqrt{x^3}}\exp\left(\langle\begin{pmatrix}x\\ \frac{1}{x}\end{pmatrix},\begin{pmatrix}-\frac{1}{2}\\ -\frac{\mu^2}{2}\end{pmatrix}\rangle\right) = \mu\exp(\mu)\frac{\exp\left(-\frac{x}{2}\right)\mathbb{1}_{\{x\in\mathbb{R}_+^*\}}}{\sqrt{2\pi x^3}}\exp\left(-\frac{\mu^2}{2x}\right).$$

On a donc une famille exponentielle de dimension 1 de paramètre naturel $-\mu^2/2$ et de statistique $1/x$.

3. (1 pt) Pour $t < \lambda/(2\mu^2)$, montrer que la fonction génératrice des moments $M_X(t)$ de la loi inverse gaussienne $IG(\mu, \lambda)$ vaut

$$\exp\left[\frac{\lambda}{\mu}\left\{1 - \left(1 - {}^{2t\mu^2}/_{\lambda}\right)^{1/2}\right\}\right].$$

Par définition, la fonction génératrice des moments est donnée par

$$\begin{aligned} M_X(t) = \mathbb{E}_\theta\left[e^{tX}\right] &= \int_{\mathbb{R}_+^*} \sqrt{\lambda}\frac{1}{\sqrt{2\pi x^3}} \exp\left(tx - \frac{\lambda x}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2x}\right) \mathrm{d}x \\ &= \sqrt{\frac{\lambda}{2\pi}} \exp\left(\frac{\lambda}{\mu}\right) \int_{\mathbb{R}_+^*} x^{-3/2} \exp\left[-\frac{1}{2}\left(x\left\{\frac{\lambda}{\mu^2} - 2t\right\} + \frac{\lambda}{x}\right)\right] \mathrm{d}x. \end{aligned}$$

On utilise alors l'indication de l'énoncé avec

$$a = \frac{\lambda}{\mu^2} - 2t \quad \text{et} \quad b = \lambda$$

D'où

$$\begin{aligned} M_X(t) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(\frac{\lambda}{\mu}\right) 2\left(\frac{a}{b}\right)^{\frac{1}{4}} \sqrt{\frac{\pi}{2\sqrt{ab}}} \exp(-\sqrt{ab}) = \sqrt{\frac{\lambda}{\sqrt{b}}} \exp\left(\frac{\lambda}{\mu} - \sqrt{ab}\right) &= \exp\left(\frac{\lambda}{\mu} - \sqrt{\frac{\lambda^2}{\mu^2} - 2t\lambda}\right) \\ &= \exp\left[\frac{\lambda}{\mu}\left(1 - \sqrt{1 - 2\frac{t\mu^2}{\lambda}}\right)\right]. \end{aligned}$$

4. (1 pt) Montrer que pour X de loi inverse gaussienne $IG(\mu, \lambda)$, on a $\mathbb{E}[X] = \mu$ et $\mathbb{E}(1/X) = {}^1/_\mu + {}^1/_\lambda$.

D'après la question 1, la forme canonique est obtenue pour la paramétrisation $(\theta_1, \theta_2) = (-\lambda/(2\mu^2), -\lambda/2)$. Autrement dit, on a la correspondance $(\mu, \lambda) = (\sqrt{\theta_2/\theta_1}, -2\theta_2)$ dans l'expression de la densité. On a vu que la fonction de partition pour la paramétrisation (μ, λ) est

$$c(\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(\frac{\lambda}{\mu}\right).$$

On en déduit que la fonction de partition de la forme canonique est

$$a(\theta_1, \theta_2) = \sqrt{-\frac{\theta_2}{\pi}} \exp\left(-2\sqrt{\theta_1\theta_2}\right).$$

On en déduit

$$\begin{aligned} \mathbb{E}[X] &= -\frac{\partial}{\partial\theta_1} \log a(\theta_1, \theta_2) = \frac{\partial}{\partial\theta_1}\left[2\sqrt{\theta_1\theta_2} - \frac{1}{2}\log\left(-\frac{\theta_2}{\pi}\right)\right] = \frac{\sqrt{\theta_2}}{\sqrt{\theta_1}} = \mu \\ \mathbb{E}\left[\frac{1}{X}\right] &= -\frac{\partial}{\partial\theta_2} \log a(\theta_1, \theta_2) = \frac{\partial}{\partial\theta_2}\left[2\sqrt{\theta_1\theta_2} - \frac{1}{2}\log\left(-\frac{\theta_2}{\pi}\right)\right] = \frac{\sqrt{\theta_1}}{\sqrt{\theta_2}} - \frac{1}{2\theta_2} = \frac{1}{\mu} + \frac{1}{\lambda}. \end{aligned}$$

5. (1 pt) Montrer que si X suit la loi $IG(\mu, \lambda)$, alors pour $\alpha > 0$, αX est de loi $IG(\alpha\mu, \alpha\lambda)$. En déduire qu'une variable inverse gaussienne de paramètres μ et λ peut être transformée en variable inverse gaussienne Z à un seul paramètre μ_0 à préciser en fonction de μ et λ.

Pour $\alpha > 0$, la fonction $\phi : x \mapsto \alpha x$ réalise un $\mathcal{C}^1$-difféomorphisme de $\mathbb{R}_+^*$ dans $\mathbb{R}_+^*$. La densité de αX est donnée pour $z \in \mathbb{R}_+^*$ par

$$f_{\alpha X}(z) = \left|\left(\phi^{-1}\right)'(z)\right| f\left(\phi^{-1}(z); \mu, \lambda\right) = \frac{1}{\alpha}\sqrt{\frac{\lambda\alpha^3}{2\pi z^3}} \exp\left[-\frac{\lambda\left(\frac{z}{\alpha} - \mu\right)^2}{2\mu^2\frac{z}{\alpha}}\right] = \sqrt{\frac{\lambda\alpha}{2\pi z^3}} \exp\left[-\frac{\lambda\alpha\left(z - \alpha\mu\right)^2}{2(\alpha\mu)^2 z}\right].$$

Il s'agit de la densité la loi inverse gaussienne $IG(\alpha\mu, \alpha\lambda)$.

On cherche à construire une variable de loi inverse gaussienne $IG(\mu_0, \mu_0^2)$. Il s'agit donc de trouver la valeur de α pour laquelle $\alpha^2\mu^2 = \alpha\lambda$. Il suffit donc de prendre $\alpha = \lambda/\mu^2$ et on a alors $\mu_0 = \lambda/\mu$.

6. (1 pt) Montrer que si $X_1, \dots, X_n$ sont des variables aléatoires *i.i.d.* suivant $IG(\mu, \lambda)$, alors

$$S_n = \textstyle\sum_{i=1}^n X_i \sim IG(n\mu, n^2\lambda).$$

Donner le comportement asymptotique de la moyenne empirique $\bar{X}_n$ quand n tend vers l'infini.

Les variables étant indépendantes et identiquement distribuées, on a

$$\begin{aligned} M_{S_n}(t) &= M_{X_1}(t)^n \\ &= \exp\left[n\frac{\lambda}{\mu}\left(1 - \left(1 - \frac{2\mu^2 t}{\lambda}\right)^{1/2}\right)\right] \\ &= \exp\left[\frac{n^2\lambda}{n\mu}\left(1 - \left(1 - \frac{2(n\mu)^2 t}{n^2\lambda}\right)^{1/2}\right)\right] \end{aligned}$$

Il s'agit de la fonction génératrice des moments de $IG(n\mu, n^2\lambda)$, d'où le résultat.

En utilisant le résultat ci-dessus et celui de la question précédente, on obtient que $\bar{X}_n$ suit la loi $IG(\mu, n\lambda)$. Il s'agit donc de comprendre l'influence du paramètre $n\lambda$ lorsque n tend vers l'infini. La variance de la loi $IG(\mu, \lambda)$ est donnée par

$$\mathbb{V}\mathrm{ar}[X] = -\frac{\partial^2}{\partial\theta_1^2}a(\theta_1, \theta_2) = -\frac{\sqrt{\theta_2}}{2\theta_1\sqrt{\theta_1}} = -\frac{1}{2\theta^2}\left(\frac{\theta_2}{\theta_1}\right)^{\frac{3}{2}} = \frac{\mu^3}{\lambda};$$

On en déduit que la variance de $\bar{X}_n$ est $\mu^3/(n\lambda)$. Ainsi lorsque n tends vers l'infini, $\bar{X}_n$ suit une dirac en μ.

7.a. (1.5 pt) Si $X \sim IG(\mu, \lambda)$, soit $Y = \min\{X, \mu^2/X\}$. Montrer que $Y \leq \mu$ avec probabilité 1 et donner la densité de la loi de Y. Quand μ est fixé, est-ce que cette loi constitue une famille exponentielle ? Si oui, donner sa dimension et une statistique naturelle.

On a

$$\mathbb{P}(\min(X, \mu^2/X) \leq \mu) = 1 - \mathbb{P}(\min(X, \mu^2/X) \geq \mu) = 1 - \mathbb{P}(X \geq \mu, \mu^2/X \geq \mu) = 1 - \mathbb{P}(X \geq \mu, X \leq \mu) = 1.$$

Donc $\mathbb{P}(Y \leq \mu) = 1$.

Pour calculer la densité de Y, on peut calculer et dériver sa fonction de répartition. Y est à valeurs dans $]0, \mu[$, pour $y \in]0, \mu[$ on a :

$$\begin{aligned}\mathbb{P}(Y<y)=P(X<y,X<\mu^2/X)+\mathbb{P}(X>\mu^2/y,X>\mu)&=\mathbb{P}(X<y,X^2<\mu^2)+\mathbb{P}(X>\mu,\mu^2/X<y)\\&=\mathbb{P}(X<y)+\mathbb{P}(X>\mu^2/y)\qquad\qquad(\text{car } y<\mu)\\&=\int_0^y f_X(x)\mathrm{d}x+\int_{\mu^2/y}^\infty f_X(x)\mathrm{d}x.\end{aligned}$$

En dérivant cette expression en y, on obtient :

$$f_Y(y)=f_X(y)+\frac{\mu^2}{y^2}f\left(\frac{\mu^2}{y}\right).$$

Or $f_X(\mu^2/y)=f_X(y)$ (*c.f.* question suivante et terme dans l'exponentielle) et donc :

$$f_Y(y)=\sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda(\mu-y)^2}{2\mu^2y}\right)\left[y^{-3/2}+\mu^{-1}y^{-1/2}\right]=\sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda(\mu-y)^2}{2\mu^2y}\right)y^{-3/2}\mu^{-1}\left[\mu+y\right].$$

μ étant fixé, cette distribution constitue une famille exponentielle de dimension 1 avec pour statistique naturelle $(\mu-y)^2/y$.

7.b. (0.5 pt) Si $X\sim IG(\mu,\lambda)$, $Y=\min\{X,\mu^2/X\}$ et

$$Z=\frac{\lambda(X-\mu)^2}{\mu^2X}$$

montrer que Z a la même loi que

$$W=\frac{\lambda(Y-\mu)^2}{\mu^2Y}$$

(On ne demande pas de trouver cette loi.)

Si $X<\mu^2/X$, on a de toute évidence $W=Y$; Si $X>\mu^2/X$:

$$W=\frac{\lambda(\mu^2/X-\mu)^2}{1/X}=\frac{\lambda(X-\mu^2)^2}{\mu^2X}.$$

7.c. (1.5 pt) Admettant que, si $X\sim IG(\mu,\lambda)$, la variable aléatoire

$$Z=\frac{\lambda(X-\mu)^2}{\mu^2X}$$

suit une loi $\chi^2(1)$, montrer la validité de la fonction de géneration de $IG(\mu,\lambda)$ suivante

```
z <- rnorm(1)**2
x <- mu + (mu**2 * z - mu * sqrt(4 * mu * lambda * z + (mu * z)**2)) / (2 * lambda)
if (runif(1) > mu / (mu + x))  x <- mu**2 / x
```

Bonus (2 pts) : Démontrer que la loi de Z est bien une loi $\chi^2(1)$.

Il s'agit de résoudre le polynôme du second degré associé à la relation entre X et Z :

$$0=\lambda X^2+X(-\mu^2Z-2\lambda\mu)+\mu^2\lambda,$$

qui a pour discriminant :

$$(-\mu^2Z-2\lambda\mu)^2-4\mu^2\lambda^2=\mu^4Z^2+4\lambda\mu^3Z.$$

Les deux solutions sont alors

$$x_1 = \mu + \frac{\mu^2 Z - \mu\sqrt{4\mu\lambda Z - \mu^2 Z^2}}{2\lambda} \quad \text{et} \quad x_2 = \mu^2/x_1 \quad (\text{car } x_1 x_2 = \mu^2).$$

Reste alors à choisir entre les deux racines, i.e. attribuer une probabilité au fait d'en choisir une plutôt que l'autre. Ce calcul nous donne qu'il faut choisir x_1 avec probabilité $\mu/(\mu + x_1)$, et x_2 sinon.

Voir Michael, John R. ; Schucany, William R. ; Haas, Roy W. (1976), *Generating Random Variates Using Transformations with Multiple Roots*, The American Statistician, 30 (2) : 88–90 pour le détail des calculs.

Bonus. On va se servir du résultat de la question 7.b., à savoir que Z a même loi que W. Rappelons que la densité de Y est donnée par :

$$f_Y(y) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda(\mu - y)^2}{2\mu^2 y}\right) y^{-3/2} \mu^{-1} [\mu + y].$$

Le changement de variable

$$\psi : y \mapsto \frac{\lambda(\mu - y)^2}{\mu^2 y},$$

réalise un $\mathcal{C}^1$ difféormorphisme entre les ouverts $]0, \mu[$ et $\mathbb{R}_+^*$. On a alors pour

$$f_Z(z) = \left| \left(\psi^{-1}\right)'(z) \right| f_Y\left(\psi^{-1}(z)\right) = \frac{1}{|\psi'(\psi^{-1}(z))|} f_Y\left(\psi^{-1}(z)\right).$$

Or pour $y \in]0, \mu[$

$$\psi'(y) = -\frac{\lambda(\mu - y)(\mu + y)}{\mu^2 y^2} = -\frac{\sqrt{\lambda\psi(y)}(\mu + y)}{\mu y^{3/2}}.$$

On en déduit pour $z \in \mathbb{R}_+^*$

$$\begin{aligned} f_Z(z) = \frac{1}{|\psi'(\psi^{-1}(z))|} f_Y\left(\psi^{-1}(z)\right) &= \frac{\mu\left(\psi^{-1}(z)\right)^{3/2}}{\sqrt{\lambda z}\left(\mu + \psi^{-1}(z)\right)} \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{z}{2}\right) \left(\psi^{-1}(z)\right)^{-3/2} \mu^{-1} \left[\mu + \psi^{-1}(z)\right] \\ &= \frac{1}{\sqrt{2\pi z}} \exp\left(-\frac{z}{2}\right). \end{aligned}$$

Il s'agit de la densité de la loi du $\chi^2(1)$.