MIDO - L3 Math. Appliquées 2021–2022

Statistical modelling

Examen partiel du 29 octobre 2021

Durée 2h00 – Documents et Calculatrice Non Autorisés

PARTIE À RABATTRE

NOM : ....

PRÉNOM : .... (lisiblement)

Ex. 1	Ex. 2	Ex. 3	Total
...... / 5.5	...... / 7	...... / 8	...... / 20.5

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 de la copie de composition doit se faire au moment de la remise de la copie par les enseignants et surveillants. Il ne sera pas accordé de délai pour cette raison en fin d'épreuve.

Les exercices sont indépendants. Toutes les réponses sont à fournir sur la copie d'énoncé. L'espace blanc alloué à chaque question est amplement suffisant pour apporter une réponse correcte.

Sauf mention contraire, la mesure de référence est la mesure de Lebesgue.

Formulaire

Loi	Notation	Densité
Exponentielle	$\mathcal{E}(\lambda)$	$f(x \mid \lambda) = \lambda e^{-\lambda x}\mathbb{1}_{x>0}$
Gamma	$\mathcal{G}\mathrm{a}(a,b)$	$f(x \mid a,b) = \frac{1}{\Gamma(a)}x^{a-1}e^{-bx}b^a\mathbb{1}_{x>0}$
Normale	$\mathcal{N}(\mu,\sigma^2)$	$f(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}\mathbb{1}_{x\in\mathbb{R}}$

French – English Lexicon

- échantillon : *sample*
- famille exponentielle : *exponential family*
- espace naturel des paramètres : *natural parameter space*
- fonction génératrice des moments : *moment-generating function*
- *i.i.d.* : *independent and identically distributed*

Exercice 1

...... / **5.5**

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, mais toute réponse fausse conduit à une note nulle.

1. On considère f définie, pour $x \in \mathbb{R}$, par $f(x) = (x+0.5)\mathbb{1}_{\{x\in]-1,1[\}}$. Pour la mesure de Lebesgue

(a) $x \mapsto 2f(x)$ est une densité

(b) f est une densité

(c) $x \mapsto f(x)/2$ est une densité

(d) Aucune de ces propositions

...... / **0.5**

PARTIE À RABATTRE

2. Soit $n, p \in \mathbb{N}^*$. On considère la variable aléatoire $X : \omega \mapsto 2^n \mathbf{1}_{\omega \in [0,1/n]}$, pour $\omega \in [0,1]$. Alors,

(a) $\mathbb{E}[X^p]$ est non définie (b) $\mathbb{E}[X^p] = 2^{np}/n$ (c) $\mathbb{E}[X^p] = 2^{np}n$ (d) $\mathbb{E}[X^p] = 2^{np-1}/n^2$

...... /**0.5**

3. Soit un jeu de données `x` qui ne contient que les valeurs possibles suivantes : `aa, aA, AA`. Pour représenter la distribution de `x`, on utilise la commande

(a) `barplot(table(x)/length(x), names.arg = c(aa, aA, AA))`

(b) `hist(x, xlab = c(aa, aA, AA))`

(c) `hist(table(x)/length(x), xlab = c(aa, aA, AA))`

(d) `barplot(x, names.arg = c(aa, aA, AA))`

...... /**0.5**

4. Soit X une variable aléatoire réelle de densité définie pour $x \in [0,1]$ par $f(x) = (\pi\sqrt{x(1-x)})^{-1}$.

(a) $\mathbb{E}[X] = 2$ (b) $\mathbb{E}[X]$ est non définie (c) $\mathbb{E}[X] = 1/2$ (d) $\mathbb{E}[X] = 0$

...... /**0.5**

5. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires telles que pour tout $\varepsilon > 0$

$$\sum_{n\in\mathbb{N}} \mathbb{P}\left[|X_n - X| > \varepsilon\right] < +\infty.$$

(a) X_n converge en loi vers X.

(b) X_n converge dans L^p vers X.

(c) X_n converge en probabilité vers X.

(d) X_n converge presque sûrement vers X.

...... / **0.5**

6. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires *i.i.d.* telle que $\mathbb{E}[X_1^2] < +\infty$. On pose

$$Y_n = \frac{1}{2n}\sum_{k=1}^{n}(X_{2k-1} - X_{2k})^2.$$

(a) $\mathbb{E}[Y_n] = \mathbb{V}ar[X_1]$

(b) Y_n converge presque sûrement vers $\mathbb{V}\mathrm{ar}[X_1]$

(c) Y_n converge presque sûrement vers $\mathbb{E}[X_1^2]$

(d) $\sqrt{n}(Y_n - \mathbb{E}[X_1^2])$ converge en loi vers $\mathcal{N}(0, \mathbb{V}ar[X_1])$

...... / **0.5**

7. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables *i.i.d.* de loi uniforme sur $[0,\theta]$, avec $\theta > 0$. Parmi les variables suivantes, donner celle(s) qui sui(ven)t asymptotiquement la loi $\mathcal{N}(0,1)$.

(a) $\sqrt{3n}\left(1-\theta/(2\overline{X}_n)\right)$

(b) $2\sqrt{3n}\left(1-\theta/\overline{X}_n\right)$

(c) $\sqrt{n}\left(1-(2\overline{X}_n)/\theta\right)$

(d) $\frac{\sqrt{n}}{3}\left(1-(2\overline{X}_n)/\theta\right)$

(e) Aucune

...... /**1**

8. Soit une famille exponentielle définie par $f(x \mid \theta) = c(\theta)h(x)\exp[\eta(\theta)T(x)]$, pour $x \in \mathbb{R}$ et $\theta \in \mathbb{R}$. On suppose que η est une bijection de $\mathbb{R}$ dans $\mathbb{R}$, de bijection réciproque $\tau \mapsto \eta^{-1}(\tau)$.

(a) $\mathbb{E}[T(X)] = -\frac{\mathrm{d}}{\mathrm{d}\theta}\log c(\theta)$

(b) $\mathbb{E}[X] = -\frac{\mathrm{d}}{\mathrm{d}\theta}\log c(\theta)$

(c) $\mathbb{E}[T(X)] = -\frac{\mathrm{d}}{\mathrm{d}\tau}[\log(c \circ \eta^{-1})](\eta(\theta))$

(d) $\mathbb{E}[T(X)] = -\frac{\mathrm{d}}{\mathrm{d}\theta}c(\theta)$

(e) $\mathbb{E}[X] = -\frac{\mathrm{d}}{\mathrm{d}\theta}c(\theta)$

(f) $\mathbb{E}[T(X)] = -\frac{\mathrm{d}}{\mathrm{d}\tau}[c \circ \eta^{-1}](\eta(\theta))$

...... / **0.5**

9. Soit $k \in \mathbb{N}^*$. On considère $f(\cdot \mid v)$ la densité paramétrée par $v \in \mathbb{R}$ et définie sur $\mathbb{R}^n$ par

$$f(x \mid v) = C(v) \exp\left[-\sum_{i=1}^{n}(x_i - v)^k\right].$$

(a) $f(\cdot \mid v)$ ne définit pas une famille exponentielle.

(b) $f(\cdot \mid v)$ forme une famille exponentielle de dimension k.

(c) $f(\cdot \mid v)$ forme une famille exponentielle de dimension n.

(d) $f(\cdot \mid v)$ forme une famille exponentielle de dimension $k-2$.

(e) $f(\cdot \mid v)$ forme une famille exponentielle de dimension $k-1$.

...... /**0.5**

10. On enregistre dans un vecteur `x` le temps passé par n consommateurs sur un site de vente en ligne. On considère que le modèle statistique est une loi gamma de paramètre $(1, 2)$. Si dans le contexte de cette expérience je souhaite obtenir la valeur de la fonction de répartition empirique au point t, j'utilise la commande

(a) `mean(x) <= t`

(b) `pgamma(t, 1, 2)`

(c) `qgamma(t, 1, 2)`

(d) `mean(x <= t)`

(e) `sum(x <= t)`

(f) `mean(pgamma(x, 1, 2) <= t)`

...... /**0.5**

Exercice 2

...... /**7**

Soit X une variable aléatoire de loi de Kumaraswamy $K(a, b)$ de paramètre $(a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$, dont la densité est donnée pour $x \in \mathbb{R}$ par

$$f(x \mid a, b) = abx^{a-1}(1 - x^a)^{b-1}\mathbb{1}_{[0,1]}(x).$$

1. Calculer la fonction de répartition de X, notée F.

...... /**0.5**

2.a Montrer que si U est une variable aléatoire de loi uniforme sur $[0,1]$ alors $Y = (1 - U^{1/b})^{1/a}$ est une variable aléatoire qui suit la loi $K(a,b)$.

...... / **0.5**

2.b Écrire le code R d'une fonction `rkum(n, a, b)` permettant de simuler n réalisations indépendantes de la loi $K(a,b)$. Le code n'utilisera ni boucle `for`, ni boucle `while`.

...... / **0.5**

3. Dans quel(s) cas la densité de f définit une famille exponentielle ? Lorsque cela a un sens, écrire la forme canonique et l'espace naturel du ou des paramètres.

(a) Lorsque a et b sont inconnus.

(b) Lorsque a est supposé connu et b inconnu.

(c) Lorsque a est supposé inconnu et b connu.

(d) Jamais.

...... /**1**

4. Donner la fonction de répartition et la densité de la variable aléatoire $Y = -\ln(X)$.

...... /**1**

5.a Soient $a > 0$ et X_n de densité $f(\cdot \mid a, 1 + n^{-1})$, $n \in \mathbb{N}^*$. Montrer que $Y_n = -\ln(X_n)$ converge en loi, lorsque n tend vers $+\infty$, vers une loi exponentielle dont on donnera la valeur du paramètre.

...... $/\mathbf{1}$

5.b Écrire un code R pour illustrer numériquement la convergence en loi de Y_n.

...... $/\mathbf{1}$

6. Soit $b > 0$, donner la fonction de répartition de la variable $Z = -\ln(1 - X)$. En déduire une suite de variables aléatoires Z_n, $n \in \mathbb{N}^*$, qui sont une transformation de variables X_n de densité f dont on précisera les paramètres et qui converge en loi, lorsque n tend vers $+\infty$, vers la loi exponentielle de paramètre b.

...... / **1.5**

Exercice 3

...... / **8**

Lorsque Y suit une loi gamma $\mathcal{G}\mathrm{a}(\alpha, \beta)$ (*c.f.* Formulaire), $(\alpha, \beta) \in \Theta \subset \mathbb{R}^2$, on dit que $X = 1/Y$ suit une loi inverse gamma de paramètres α, β, notée $\mathrm{Inv} - \mathcal{G}\mathrm{a}(\alpha, \beta)$. Dans cet exercice, on suppose que **les deux paramètres α and β sont inconnus**.

1. Montrer que la densité de X est

$$f_X(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(\frac{-\beta}{x}\right) \mathbb{1}_{\{x>0\}}.$$

...... / **0.5**

2. Montrer que la famille ainsi définie est une famille exponentielle et calculer l'espace naturel des paramètres. La famille ainsi définie est-elle sous forme canonique ? Minimale ? Régulière ?

...... /**1.5**

3.a Calculer $\mathbb{E}[1/X]$ en fonction de α, β.

...... $/$ **0.5**

3.b Calculer $\mathbb{E}[\log(X)]$ en fonction de α, β et de la fonction digamma définie pour $x > 0$ par

$$\Psi(x) = \frac{1}{\Gamma(x)} \frac{\mathrm{d}}{\mathrm{d}x} \Gamma(x).$$

...... $/$ **0.5**

4.a Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables *i.i.d.* suivant la loi $\mathrm{Inv}-\mathcal{G}\mathrm{a}(\alpha,1)$. Montrer qu'il existe une suite $(a_n)_{n\in\mathbb{N}^*}$ telle que

$$a_n \underset{n\to\infty}{\longrightarrow} 0 \quad \text{et} \quad a_n \log\left(\prod_{i=1}^{n} X_i\right) \underset{n\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} -\Psi(\alpha).$$

...... /**1**

4.b On admet que $\mathbb{V}\mathrm{ar}(\log(X_1)) = \sigma^2 < \infty$. Trouver une suite $(c_n)_{n\in\mathbb{N}^*}$ dont l'expression ne dépend ni de σ^2 ni de $\Psi(\alpha)$ telle que

$$c_n\left(\frac{1}{n}\sum_{i=1}^{n}\log(X_i) + \Psi(\alpha)\right) \underset{n\to\infty}{\overset{d}{\longrightarrow}} \mathcal{N}(0,1).$$

...... /**1.5**

4.c Pour $n \in \mathbb{N}^*$ et $\varepsilon \in]0,1[$, en déduire la valeur de t_n en fonction de c_n et ε telle que

$$\mathbb{P}\left[-t_n - \frac{1}{n}\sum_{i=1}^{n}\log(X_i) \leq \Psi(\alpha) \leq t_n - \frac{1}{n}\sum_{i=1}^{n}\log(X_i)\right] \underset{n\to\infty}{\longrightarrow} \varepsilon.$$

...... /**1**

5.a On souhaite trouver une approximation de $\Psi(\alpha)$ pour $\alpha = 63$. Donner un code R qui,

1. pour $n = 1000$, permet de simuler un échantillon $\mathtt{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n})$ suivant la loi $\mathrm{Inv}-\mathcal{G}\mathrm{a}(63, 1)$ à partir du générateur de la loi gamma ;
2. pour cet échantillon `x` renvoie un intervalle de confiance asymptotique pour $\Psi(63)$ au niveau $\varepsilon = 90\%$.

Le code n'utilisera ni boucle `for`, ni boucle `while`.

...... /**1**

5.b (Bonus) Donner un code `R` d'une fonction donnant l'ensemble des valeurs de la suite $(a_k \log \left(\prod_{i=1}^{k} x_i\right))_{k\in\mathbb{N}^*}$ pour un vecteur `x` donné. Le code n'utilisera ni boucle `for`, ni boucle `while`.

...... /**0.5**

Espace Complémentaire