Dauphine | PSL
UNIVERSITÉ PARIS

MIDO - L3 Math. Appliquées 2022–2023

Statistical modelling

Examen partiel du 2 Novembre 2022

DURÉE 2H00 – DOCUMENTS ET CALCULATRICE NON AUTORISÉS

Ex. 1	Ex. 2	Ex. 3	Total
..... / 7	..... / 7.5	..... / 7.5	..... / 22

PARTIE À RABATTRE

NOM : .....

PRÉNOM : .....

(lisiblement)

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.

French – English Lexicon

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

Exercise 1

..... / **7**

For the following statements, give the correct answer(s). There is no negative point but any incorrect/missing answer or incorrect/missing justification give no point.

1. Consider the function f defined, for $x \in \mathbb{R}$, by $f(x) = (x^2 - x - 2)\mathbb{1}_{\{x \in]-1,2[\}}$. For the Lebesgue measure

(a) $x \mapsto 9f(x)/2$ is a density,

(b) $x \mapsto 2f(x)/9$ is a density,

(c) $x \mapsto -2f(x)/9$ is a density,

(d) None of the other answers.

..... / **0.5**

PARTIE
À
RABATTRE

2. Given X a random variable with density on $\mathbb{R}$ (with respect to the Lebesgue measure) proportional to

$$g(x) = \exp(-b|x|), \quad b \in \mathbb{R}_+^*.$$

The characteristic function of X at point $t \in \mathbb{R}$ is given by

(a) $2b/(b^2+t^2)$,
(b) $4/(b^2+t^2)$,
(c) $b^2/(b^2+t^2)$,
(d) None of the other answers.

Hint. for any $t \in \mathbb{R}$, $\exp(-bx \pm itx) \underset{x\to+\infty}{\longrightarrow} 0$.

..... /**1**

3. A sequence $(Y_n)_{n\in\mathbb{N}^*}$ of random variables is said to c-converge to a random variable Y if, for any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\{|Y_n - Y| > \varepsilon\} < \infty.$$

(a) c-convergence implies convergence in distribution.

(b) almost sure convergence implies c-convergence.

(c) convergence in probability implies c-convergence

(d) c-convergence implies convergence in probability.

(e) c-convergence implies almost sure convergence.

(f) none of the other answers holds in full generality

..... / **0.5**

4. Let $(X_n)_{n\in\mathbb{N}^*}$ be a sequence of *i.i.d.* discrete random variables such that

$$\mathbb{P}[X_n = 0] = \frac{n-1}{n} \quad \text{and} \quad \mathbb{P}[X_n = \sqrt{n}] = \frac{1}{n}.$$

Then, when n goes to $+\infty$,

(a) the sequence converges in L^1 (convergence in mean),

(b) the sequence converges in L^2 (convergence in quadratic mean),

(c) for any continuous function g, we have $\mathbb{E}[g(X_n)]$ converges to 0,

(d) converges in distribution,

(e) does not converge at all.

..... / **1.5**

5. Given a vector `x` = $(x_1, \ldots, x_n)$, which of the following codes compute a vector `u` = $(u_1, \ldots, u_n)$ such that for $i \in \{1, \ldots, n\}$,

$$u_i = \sum_{k=1}^{i} \frac{x_k}{i} \quad ?$$

(a) `mean(x)`

(b) `cumsum(x)/cumsum(rep(1, length(x)))`

(c) `cumsum(x)/seq_len(length(x))`

(d) `cumsum(x)/(1:length(x))`

(e) `sum(x)/length(x)`

(f) `sum(x)/sum(length(x))`

..... /**0.5**

6. Consider an exponential family defined by $f(x \mid \theta) = c(\theta)h(x)\exp[\eta(\theta)T(x)]$, for $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$. Assume that η is a bijective function from $\mathbb{R}$ to $\mathbb{R}$, and denote $\tau \mapsto \eta^{-1}(\tau)$ its inverse. If X has density $f(\cdot \mid \theta)$, the moment generating function of $T(X)$ is defined for $t \in \mathbb{R}$ by

(a) $c(\theta)/c(\theta + t)$

(b) $c\left(\eta^{-1}(\tau)\right)/c\left(\eta^{-1}(\tau + t)\right)$

(c) $c\left(\eta^{-1}(\tau)\right)/c\left(\eta^{-1}(\tau) + t\right)$

(d) $c(\theta)/c\left(\eta^{-1}(\eta(\theta) + t)\right)$

..... /**0.5**

7. Consider the density (with respect to the Lebesgue measure on $\mathbb{R}$) parametrised by an **unknown** $(k, \lambda) \in \mathbb{N}^* \times \mathbb{R}_+^*$ and defined by

$$f(x \mid k, \lambda) = \frac{\lambda^k x^{k-1} \exp(-\lambda x)}{(k-1)!} \mathbb{1}_{x \geq 0}.$$

(a) It constitutes a minimal and canonical exponential family.

(b) It constitutes a minimal exponential family but is not in a canonical form.

(c) It constitutes an exponential family that is weither minimal nor canonical.

(d) None of the other answers.

..... /**1**

8. We run an experiment where we measure how much time n different customers spend on a specific page of a website. Our observations $x_1, \ldots, x_n$ are stored in a vector `x`. We assume that the underlying statistical model is a Gamma distribution with parameter (α, β). Which one among the following command lines does return the first quartile of the Gamma model with parameter $(1,2)$?

(a) `rgamma(0.25, 1, 2)`

(b) `pgamma(0.25, 1, 2)`

(c) `dgamma(0.25, 1, 2)`

(d) `qgamma(0.25, 1, 2)`

(e) `quantile(x, 0.25, 1, 2)`

(f) `quantile(0.25, 1, 2)`

..... /**0.5**

9. When given a sample `x` of size n from F and considering the median $\text{med}(X)$ as the quantity of interest, a bootstrap approximation of a 95% interval of variability of the empirical median is given by

(a) `quantile(median(matrix(sample(x, n * m, rep = TRUE), m)), c(.025, .975))`

(b) `quantile(apply(matrix(sample(x, n * m, rep = TRUE), m), 1, median), c(.025, .975))`

(c) `quantile(matrix(sample(median(x), n * m, rep = TRUE), m), c(.035, .985))`

(d) `median(apply(matrix(sample(x, n * m, rep=TRUE), m), 1, sum), prob = .95)`

(e) `quantile(0.25, 1, 2)`

..... /**1**

Exercise 2

..... /**7.5**

Suppose that Z_1 and Z_2 are *i.i.d.* random variables distributed according to the standard normal distribution $\mathcal{N}(0,1)$. We define the general Rayleigh distribution with **unknown** scale parameter $b \in \mathbb{R}_+^*$ as the density of the random variable $X = b\sqrt{Z_1^2 + Z_2^2}$.

Reminder. The probability density function of the normal $\mathcal{N}(0,1)$ distribution is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

1. Show that the cumulative distribution function of X is defined for all $t \in \mathbb{R}_+^*$ by

$$F(t) = 1 - \exp\left(-\frac{t^2}{2b^2}\right).$$

..... /**0.5**

2. Compute the density $f(\cdot \mid b)$ of X and show that it can define an exponential family with a natural statistic $T(\cdot)$ that is positive. Precise its canonical form and its natural parameter space. Is the family regular and minimal?

..... /**1**

3. Show that the moment generating function of X^2 is defined by

$$M_{X^2}(t) = \frac{1}{1-2b^2t}, \quad \text{for} \quad t \in \left(-\infty, \frac{1}{2b^2}\right).$$

..... /**0.5**

4. Given a sequence $(X_n)_{n\in\mathbb{N}^*}$ of *i.i.d.* random variables distributed according to $f(\cdot \mid b)$, show that

$$\widehat{b}_n^2 = \frac{1}{2n}\sum_{k=1}^{n} X_k^2 \xrightarrow[n\to+\infty]{\mathbb{P}} b^2.$$

..... /**1**

5. Show that

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

..... / **1.5**

6. Show that there exists a sequence c_n, whose expression depends on known constants and $\widehat{b}_n^2$, such that

$$c_n\left(\sqrt{\widehat{b}_n^2} - b\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

..... / **1**

7. Find a function g such that

$$\sqrt{n}\left[g\left(\sqrt{\frac{2}{\pi}}\overline{X}_n\right) - g(b)\right] \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1), \quad \text{where} \quad \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

..... / **1.5**

8. In order to approximate b, is it better to use an approximation based on $\overline{X}_n$ or an approximation based on $\widehat{b}_n^2$?

Hint. $4/\pi \approx 1.27$.

..... /**0.5**

Exercise 3

..... /**7.5**

Let f be a probability density function on $\mathbb{R}$ such that

$$f(x) = c\tilde{f}(x), \quad \forall x \in \mathbb{R},$$

where the positive function $\tilde{f}$ is known and computable (for instance, by an R function `df(x)`), and the constant c is unknown. This setting is called a *missing normalising constant* problem.

1. Show that that the constant c is uniquely defined by $\tilde{f}$.

..... /**0.5**

Until further notice, consider the special case of an interval $]a,b[$ such that

$$\forall x \notin]a,b[, \quad \tilde{f}(x) = 0 \quad \text{and} \quad \sup_{x\in\mathbb{R}} \tilde{f}(x) = M < \infty.$$

2.a. Defining the rectangle $\mathscr{R} =]a,b[\times]0,M[$ and the subset of $\mathscr{R}$

$$\mathscr{S} = \{x = (x_1, x_2) \in \mathscr{R}\,;\, x_2 \leq \tilde{f}(x_1)\},$$

give the ratio of the surfaces of $\mathscr{S}$ and of $\mathscr{R}$.

..... /**0.5**

2.b. Considering $U = (U_1, U_2)$ a Uniform random point on $\mathscr{R}$, compute the probability $\mathbb{P}[U_2 \leq \tilde{f}(U_1)]$.

..... /**0.5**

2.c. Given an *i.i.d.* sequence $U^1, \ldots, U^n$ of Uniform random points on $\mathscr{R}$, and exploiting the Law of Large Numbers, construct a converging (with n) approximation of c^{-1}.

..... /**0.5**

2.d. Derive a converging estimator of c.

..... /**0.5**

2.e. Write an R code calling df() that produces this estimator of c.

..... /**1.5**

From now and till the end of the Exercise, consider two simultaneous missing normalising constant densities

$$f_1(x) = c_1 \tilde{f}_1(x) \quad \text{and} \quad f_2(x) = c_2 \tilde{f}_2(x),$$

where c_1, c_2 are unknown and $\tilde{f}_1$, $\tilde{f}_2$ are known (with associated R functions `df1` and `df2`).

3.a. Let X_1 be a random variable with density $f_1(x)$. Show that

$$\mathbb{E}\left[\frac{\tilde{f}_2(X_1)}{\tilde{f}_1(X_1)}\right] = \frac{c_1}{c_2}.$$

..... /**0.5**

3.b. Given an *i.i.d.* sequence $X_{11}, \ldots, X_{1n}$ of random variable with density f_1, deduce from question 3.a. an approximation of c_1/c_2 that is converging with n.

..... /**0.5**

3.c. Let $\alpha(\cdot)$ be a positive function such that

$$\int \alpha(x)\tilde{f}_1(x)\tilde{f}_2(x)\mathrm{d}x < +\infty.$$

Show that, if X_1 is a random variable with density f_1 and X_2 is a random variable with density f_2,

$$\frac{\mathbb{E}\left[\alpha(X_1)\tilde{f}_2(X_1)\right]}{\mathbb{E}\left[\alpha(X_2)\tilde{f}_1(X_2)\right]} = \frac{c_1}{c_2}.$$

..... /**1**

3.d. Deduce from question 3.c. a converging approximation of c_1/c_2 based on two sequences $X_{11},\ldots,X_{1n}$ and $X_{21},\ldots,X_{2n}$ of *i.i.d.* random variables with density f_1 and f_2 respectively.

..... /**0.5**

3.e. Assuming there exist simulation functions `rf1(n)` and `rf2(n)` to simulate from f_1 and f_2, write an R code that produces this estimator of c_1/c_2.

..... /**1**