Dauphine | PSL
UNIVERSITÉ PARIS

MIDO - L3 Math. Appliquées 2022–2023

Statistical modelling

Examen final du 11 Janvier 2023

DURÉE 2H00 – DOCUMENTS ET CALCULATRICE NON AUTORISÉS

Ex. 1	Ex. 2	Total
…… / 9	…… / 12	…… / 21

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. Si besoin, vous pouvez utiliser l'espace complémentaire disponible en fin de copie.

Formulaire

Loi	Notation	Densité
Bernoulli	$\mathscr{B}(p)$	$f(x \mid p) = p^x(1-p)^{1-x}\mathbb{1}_{x\in\{0,1\}}$
Binomiale	$\mathscr{B}(n,p)$	$f(x \mid n,p) = \binom{n}{x}p^x(1-p)^{n-x}\mathbb{1}_{x\in[\![0,n]\!]}$
Gamma	$\mathscr{G}\mathrm{a}(a,b)$	$f(x \mid a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}\mathbb{1}_{x>0}$
Normale	$\mathscr{N}(\mu,\sigma^2)$	$f(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}\mathbb{1}_{x\in\mathbb{R}}$

French – English Lexicon

- *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*
- vraisemblance : *likelihood*
- statistique exhaustive : *sufficient statistic*
- statistique libre : *ancillary statistic*
- statistique complète : *complete statistic*

PARTIE À RABATTRE

Exercise 1

...... /9

For the following statements, give the correct answer(s). Incorrect answers and missing justification return zero point while incomplete answers gain partial points.

1. Let $(X_n)_{n\in\mathbb{N}^*}$ be a sequence of independent 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, $\mathbb{E}[g(X_n)]$ converges to 0,

(d) the sequence converges in distribution,

(e) the sequence does not converge at all.

...... /1

2. Consider the exponential family associated to the Bernoulli distribution with unknwon parameter $p \in (0,1)$. The moment generating function of natural statistic $T(x) = x$ of the family is defined for $t \in \Theta \subseteq \mathbb{R}$ by

(a) $(1-p)/(1-p-t)$ (b) $1+p(\exp(t)-1)$ (c) $(1-p-t)/(1-p)$ (d) $1/[1+p(\exp(t)-1)]$

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

3. 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

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

(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 neither minimal nor canonical.

(d) None of the other answers.

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

4. 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 R command lines does return the first quartile of the sample?

(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(0.25, 1, 2)`

(f) `quantile(x, 0.25)`

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

5. Let X be a random variable with density, parametrised by $\lambda \in \mathbb{R}_+^*$, with respect to the composition of the counting measure on $\mathbb{N}$ and the Lebesgue measure on $\mathbb{R}_+^*$:

$$f_X(x) = \begin{cases} \frac{\lambda^x \exp(-\lambda)}{2(x!)}, & \text{if} \quad x \in \mathbb{N}, \\ \frac{\lambda}{2} \exp(-\lambda x), & \text{otherwise.} \end{cases}$$

The likelihood for the sample $(1, 1, 2, 2, 2, x_1, \ldots, x_n)$, with $x_1, \ldots, x_n \notin \mathbb{N}$ is

(a) $\dfrac{\lambda^n}{2^{n+5}} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$

(b) $\dfrac{\lambda^{n+8}}{2^{n+8}} \exp\left[-\lambda \left(\sum_{i=1}^{n} x_i + 5\right)\right]$

(c) $\dfrac{\lambda^8 \exp(-5\lambda)}{2^{n+8}}$

(d) $\dfrac{\lambda^{n+3}}{2^{n+6}} \exp\left[-\lambda \left(\sum_{i=1}^{n} x_i + 2\right)\right]$

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

6. Consider X distributed according to the Binomial distribution $\mathscr{B}(n, p)$, $n \in \mathbb{N}^*$ **known** and $p \in (0, 1)$ **unknown**. If we denote θ the parameter of the canonical form of this exponential family, $I(p)$ and $I(\theta)$ the Fisher information contained in X for p and θ respectively, we have

(a) $I(p) = n/[p(1-p)]$ (b) $I(p) = 1/[p(1-p)]$ (c) $I(\theta) = ne^{-\theta}\left(1+e^{\theta}\right)^2$ (d) $I(\theta) = ne^{\theta}/\left(1+e^{\theta}\right)^2$

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

7. Consider a regular and minimal exponential family with natural statistic $T(\cdot)$ and density $f(\cdot \mid \theta)$, $\theta \in \Theta \subseteq \mathbb{R}$. For $X_1, \ldots, X_n$ *i.i.d.* random variables distributed according to $f(\cdot \mid \theta)$, we set $S = \sum_{i=1}^{n} T(X_i)$.

(a) Any bijective transform of S is sufficient for θ.

(b) S is minimal sufficient for θ.

(c) Any sufficient statistic for θ that is a function of S is minimal sufficient for θ.

(d) None of the other answers.

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

8. Let $X_1, \ldots, X_n$ be *i.i.d.* random variables distributed according to the normal distribution $\mathcal{N}(\mu, 1)$ and denote

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad X_{(1)} = \min(X_1, \ldots, X_n) \quad \text{and} \quad X_{(n)} = \max(X_1, \ldots, X_n).$$

(a) $X_{(n)}$ - $X_{(1)}$ is independent of $\overline{X}_n$.

(b) $\left(X_{(1)}, X_{(n)}\right)$ is not a complete statistic.

(c) $\left(X_{(1)}, X_{(n)}\right)$ is a sufficient statistic for μ.

(d) $\left(X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n\right)$ is independent of $\overline{X}_n$.

(e) None of the other answers.

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

Exercise 2

...... /12

Given $a \in \mathbb{R}$, the Lévy distribution of scale parameter $b \in \Theta_0 \subseteq \mathbb{R}_+^*$ admits a density with respect to the Lebesgue measure on $(a, +\infty)$

$$f(x \mid b) = \frac{1}{x-a}\sqrt{\frac{b}{2\pi(x-a)}}\exp\left(-\frac{b}{2x-2a}\right)\mathbb{1}_{x>a}.$$

In this exercise we consider $X_1, \ldots, X_n$, $n > 2$, *i.i.d.* random variables distributed according to $f(\cdot \mid b)$ for a **known** parameter a and an **unknown** scale parameter b.

1. Show that $f(\cdot \mid b)$ can define an exponential family with a natural statistic $T(\cdot)$. Precise its canonical form and its natural parameter space. Is the family regular and minimal?

...... /1

2. Show that there is a **unique** maximum likelihood estimator of b defined as

$$\widehat{b}_n = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i - a}\right)^{-1}$$

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

3. Show that $\widehat{b}_n \xrightarrow[n\to+\infty]{\mathbb{P}} b$.

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

4. Show that

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

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

5. Prove that the bias of the estimator $\widehat{b}_n$ is strictly positive.

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

6. Given a vector of observations `x`, write a R code that computes a non-parametric bootstrap estimation of the bias for k bootstrap samples.

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

7. Denote $b_1^\star, \dots, b_k^\star$ the estimations of b we got with the previous bootstrap procedure. Give the definition of the empirical boostrap confidence interval on b with $1-\alpha$ confidence level, $\alpha \in (0,1)$. Choose the R code that computes this interval for a 90% confidence level, where `b_ref` $= \widehat{b}_n$ and `b_star` $= (b_1^\star, \dots, b_k^\star)$.

(a) `quantile(b_star, c(.05, .95))`

(b) `b_ref - quantile(b_star - b_ref, c(.95, .05))`

(c) `quantile(b_star, c(.025, .975))`

(d) `b_ref - quantile(b_star - b_ref, c(.975, .025))`

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

8. Show that if X_1 is distributed according to $f(\cdot \mid b)$ then $Y_1 = (2X_1 - 2a)^{-1}$ admits a density $g(\cdot \mid b)$ with respect to the Lebesgue measure on $\mathbb{R}_+^*$ given by

$$g(y \mid b) = \frac{\sqrt{b}}{\sqrt{\pi y}} \exp(-by) \mathbb{1}_{y>0}.$$

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

9. Admit the following result : if $Y_1, \ldots, Y_n$ are *i.i.d.* random variables distributed according $g(\cdot \mid b)$ then $n/(Y_1 + \ldots + Y_n)$ has a density $d(\cdot \mid b)$ with respect to the Lebesgue measure on $\mathbb{R}_+^*$ given by

$$d(x \mid b) = \frac{(nb)^{n/2}}{\Gamma(n/2)} x^{-\frac{n}{2}-1} \exp\left(-\frac{nb}{x}\right) \mathbb{1}_{x>0}.$$

Show that the following estimator is an unbiased estimator of b :

$$\widehat{\beta}_n = \frac{n-2}{n} \widehat{b}_n.$$

Hint. You can use without justification that for all $\alpha > 1$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ and

$$\int_0^{+\infty} \frac{(nb)^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{nb}{x}\right) \mathrm{d}x = 1.$$

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

10. Show that

$$\mathbb{V}\mathrm{ar}_b\left[\widehat{b}_n\right] \geq 2\left(\frac{nb}{n-2}\right)^2.$$

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

11. Show that $\widehat{\beta}_n$ is the unique uniformly minimum variance unbiased estimator.

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