Dauphine | PSL
UNIVERSITÉ PARIS

MIDO - L3 Math. Appliquées 2020–2021

Statistical modelling

Examen final du 4 janvier 2021
Durée 2h00 – Documents et calculatrice non-autorisés

Ex. 1	Ex. 2	Ex. 3	Total

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.

Formulaire

Loi	Notation	Densité
Beta	$\mathcal{B}eta(a,b)$	$f(x \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\mathbb{1}_{x\in]0,1[}$
Chi-deux	$\chi^2(n)$	correspond à la loi gamma $\mathcal{G}a(n/2, 1/2)$
Exponentielle	$\mathcal{E}(\lambda)$	$f(x \mid \lambda) = \lambda e^{-\lambda x}\mathbb{1}_{x>0}$
Gamma	$\mathcal{G}a(a,b)$	$f(x \mid a,b) = \frac{1}{\Gamma(a)} x^{a-1}e^{-bx}b^a\mathbb{1}_{x>0}$
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*
- fonction génératrice des moments : *moment-generating function*
- *i.i.d.* : *independent and identically distributed*
- statistique libre : *ancilary statistic*
- statistique exhaustive : *sufficient statistic*
- statistique complète : *complete statistic*
- vraisemblance : *likelihood*

Exercice 1

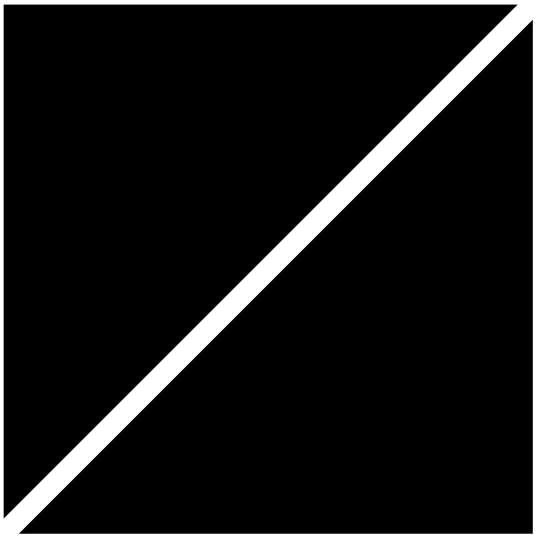
Dans cet exercice il vous est demandé de donner la ou les bonnes réponses. Seules les réponses justifiées seront validées. Il n'y a pas de points négatifs. Sauf mention contraire, n est un entier naturel supérieur à 2.

..... /**6**

1. Soient $X_1, \dots, X_n$ *i.i.d.* de loi $\mathcal{B}eta(1,\lambda)$, $\lambda \in \mathbb{R}_+^*$, et de réalisations $x_1, \dots, x_n$. L'estimateur du maximum de vraisemblance de λ associé à $(x_1, \dots, x_n)$ est donné par

(a) $\hat{\lambda}_n = 1/(1-\overline{x}_n)$,
(b) $\hat{\lambda}_n = -\sum_{i=1}^n \log(1-x_i)/n$,
(c) $\hat{\lambda}_n = -\log\left[\prod_{i=1}^n (1-x_i)\right]/n$,
(d) $\hat{\lambda}_n = -n/\sum_{i=1}^n \log(1-x_i)$.

..... /**0.5**



2. Soit un vecteur aléatoire $(X_1, \ldots, X_n)$ tel que pour $i, j \in [\![1, n]\!], i \neq j$, X_i indépendant de X_j et X_i suit la loi $\mathcal{N}(\mu, \sigma_i^2)$, où $\mu \in \mathbb{R}$ est inconnu et les $\sigma_i \in \mathbb{R}_+^*$ sont supposés connus, de réalisation $(x_1, \ldots, x_n)$. L'estimateur du maximum de vraisemblance de μ associé à $(x_1, \ldots, x_n)$ est donné par

(a) $\hat{\mu}_n = (\sigma_1^{-2} x_1 + \ldots + \sigma_n^{-2} x_n)/(\sigma_1^{-2} + \ldots + \sigma_n^{-2})$,
(b) $\hat{\mu}_n = (x_1 + \ldots + x_n)/n$,
(c) $\sum_{i=1}^n \sigma_i^2 (x_i - \hat{\mu}_n)^{-1} = 0$,
(d) $\hat{\mu}_n = (\sigma_1^2 x_1 + \ldots + \sigma_n^2 x_n)/(\sigma_1^2 + \ldots + \sigma_n^2)$.

..... / **0.5**

3. Soient $X_1, \ldots, X_n$ *i.i.d.* suivant une distribution discrète qui prend avec probabilité $1/3$ chacune des valeurs $\{\theta - 1, \theta, \theta + 1\}$, $\theta \in \mathbb{R}$. La statistique des observations extrêmes $S(X_1, \ldots, X_n) = (X_{(1)}, X_{(n)})$ est

(a) incomplète et non-exhaustive,
(b) complète et non-exhaustive,
(c) complète et exhaustive,
(d) incomplète et exhaustive.

..... / **0.5**

4. Soient $X_1, \dots, X_n$ *i.i.d.* de loi géométrique $\mathcal{G}(e^{-\theta_0})$, avec $\theta_0 \in \mathbb{R}_+^*$ inconnu, *i.e.*, telle que pour $k \in \mathbb{N}$, $\mathbb{P}(X_1 = k) = e^{-\theta_0}(1 - e^{-\theta_0})^{k-1}$. L'information de Fisher sur θ_0 fournie par $\sum_{i=1}^n X_i$ vaut

(a) $(1 - e^{-\theta_0})/n$, (b) $n/(e^{\theta_0} - 1)$, (c) $n/(1 - e^{-\theta_0})$, (d) $ne^{-\theta_0}/(1 - e^{-\theta_0})^2$.

..... /**1**

5. Soient $X_1, \dots, X_n$ *i.i.d.* de loi normale $\mathcal{N}(1, \sigma^2)$, $\sigma \in \mathbb{R}_+^*$. Si $\theta = \sigma^2$ est le paramètre naturel, quelle est la bonne fonction de score ?

(a) $\frac{1}{\theta^3}\sum_{i=1}^n X_i^2 - \frac{n}{2\theta^2}$, (b) $\frac{1}{\theta^{3/2}}\sum_{i=1}^n (X_i - 1)^2 - \frac{n}{2\theta}$, (c) $\frac{1}{2\theta^2}\sum_{i=1}^n (X_i - 1)^2 - \frac{n}{2\theta}$, (d) $\frac{n}{\theta^2}\overline{X_n}^2 - \frac{n}{2\theta}$.

..... /**0.5**

6. Si $X_1, \dots, X_n$ sont *i.i.d.* de loi $\mathcal{B}eta(\beta + 1, 1)$, $\beta \in \mathbb{R}_+^*$, un estimateur sans biais de $\beta/(1 + \beta)$ est donné par

(a) $1 + (n + 1)\left[\sum_{i=1}^n \log(X_i)\right]^{-1}$,
(b) $1 + n\left[\sum_{i=1}^n \log(X_i)\right]^{-1}$,
(c) $1 + (n - 1)\left[\sum_{i=1}^n \log(X_i)\right]^{-1}$,
(d) $1 + n^{-1}\left[\sum_{i=1}^n \log(X_i)\right]$.

..... /**1**

7. Soit $(x_1, \ldots, x_n)$ un échantillon *i.i.d.* de distribution F donnée par $X_1 \sim \mathcal{N}(\mu_0, 1/2)$. Si le modèle statistique imposé est $\{\mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\}$, quelle affirmation sur la vraisemblance $L(\mu \mid x_1, \ldots x_n)$ du modèle est correcte ?

(a) elle vaut $\prod_{i=1}^{n} \left[e^{-(x_i-\mu)^2}/\sqrt{2\pi}\right]$,

(b) $\int_{\mathbb{R}^n} L(\mu \mid x_1, \ldots, x_n)\mathrm{d}F(x_, \ldots, x_n) = 1$,

(c) $\int_{\mathbb{R}} L(\mu \mid x_1, \ldots, x_n)\mathrm{d}\mu = 1$,

(d) $\int_{\mathbb{R}^n} L(\mu \mid x_1, \ldots, x_n)\mathrm{d}(x_1, \ldots, x_n) = 1$.

..... /**0.5**

8. Soit `x` un vecteur de n observations supposées être des réalisations de la loi de Cauchy de paramètre $(\theta, 1)$, $\theta \in \mathbb{R}$. La médiane de l'échantillon est un estimateur de θ dont on souhaite estimer le biais par une approche bootstrap non-paramétrique. Quel code permet d'obtenir une réalisation bootstrap de la médiane ?

(a) `median(rcauchy(n, median(x)))`

(b) `sample(x, replace = TRUE)`

(c) `median(sample(x, replace = FALSE))`

(d) `median(pcauchy(n, median(x)))`

(e) `rcauchy(n, median(x))`

(f) `median(sample(x, replace = TRUE))`

..... /**0.5**

9. Soit `x` un vecteur de n observations. Supposons que l'on travaille avec le modèle statistique associé à la loi Gamma $\mathcal{G}a(1, 2)$, quelle ligne de commande permet d'obtenir le premier quartile de cette loi ?

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

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

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

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

(e) `quantile(x, 0.25)`

..... /**0.5**

10. Pour $X_1, \ldots, X_n$ *i.i.d.* de loi $\mathcal{N}(0, \lambda)$, une statistique libre de λ est

(a) X_2X_3, (b) $\mathbb{I}_{\overline{X}_n \geq 0}$, (c) $(\overline{X}_n)^2$, (d) X_2/X_3, (e) $-\overline{X}_n$.

..... /**0.5**

Exercice 2

..... /**9.5**

La loi binomiale négative est une loi discrète dont la densité par rapport à la mesure de comptage sur $\mathbb{N}$ est

$$f(x \mid p) = \binom{x+r-1}{x} p^r (1-p)^x, \quad \text{avec} \quad r \in \mathbb{R}_+^* \quad \text{et} \quad p \in]0,1[.$$

Dans cet exercice, on suppose que le paramètre r est connu et fixé, et le paramètre p est inconnu. On note $X_1, \ldots, X_n$ des variables aléatoires *i.i.d.* suivant la loi binomiale négative de paramètre (r, p) et $x_1, \ldots, x_n$ un n-échantillon pour cette loi.

1. Montrer que cette distribution constitue une famille exponentielle. Donner un paramètre naturel, une statistique naturelle, et démontrer que la représentation est minimale et régulière.

..... /**1**

2. Montrer que la fonction génératrice des moments de cette loi vaut

$$M_X(t) = \left[\frac{p}{1-e^t(1-p)}\right]^r, \qquad t < -\log(1-p).$$

..... /**0.5**

3. Montrer que l'estimateur suivant converge en probabilité vers p

$$\hat{p}_n = \frac{nr}{nr + \sum_{k=1}^{n} X_k}\,.$$

..... /**1**

4. Trouver une suite c_n indépendante de p telle que

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

..... /**1.5**

5. Donner la vraisemblance de p associée à un n-échantillon $(x_1, \ldots, x_n)$ et montrer que l'estimateur du maximum de vraisemblance de p est égal à $\hat{p}_n$.

..... /**1**

6. Montrer que l'erreur quadratique moyenne pour $\hat{p}_n$ est **strictement** minorée par $\mathbb{V}\text{ar}[\hat{p}_n]$, *i.e.*, $\mathbb{E}_p\left[(\hat{p}_n - p)^2\right] > \mathbb{V}\text{ar}[\hat{p}_n]$. Sachant que l'on dispose d'un vecteur d'observations `x`, écrire un code R donnant une estimation de cette erreur quadratique moyenne utilisant `k` itérations d'un bootstrap non-paramétrique.

..... /**2**

7. Montrer que $S(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$ suit une loi binomiale négative de paramètre (nr, p).

..... /**0.5**

8. Montrer que $\hat{\theta}_n = (nr-1)/[S(X_1, \ldots, X_n) + nr - 1]$ est un estimateur sans biais de p.

..... /**1**

9. Montrer que $\hat{\theta}_n$ est de variance minimale parmi les estimateurs sans biais de p.

Indication. *Si une famille exponentielle paramétrée par $\theta \in \Theta$ et de statistique naturelle $T(X)$ est régulière alors, pour un échantillon i.i.d. $(X_1, \ldots, X_n)$ de cette famille, $\sum_{k=1}^{n} T(X_i)$ est une statistique complète de θ.*

..... /**1**

Exercice 3

..... /**11**

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon *i.i.d.* de loi $\mathcal{N}(\mu, \sigma^2)$, avec $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+^*$. On rappelle que la loi de Student $\mathcal{T}(n-1)$ est historiquement (*c.f.* Student, 1908) définie comme la loi du rapport

$$T(\mathbf{X}) \overset{\text{def}}{=} \sqrt{n}\frac{(\bar{X}_n - \mu)}{S_n}, \quad \text{où} \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

sont respectivement la moyenne et la variance empiriques associées à l'échantillon.

1. En utilisant (obligatoirement) le théorème de Basu, montrer que si $(X_1, \ldots, X_n)$ est un échantillon *i.i.d.* de loi $\mathcal{N}(\mu, 1)$, alors $\overline{X}_n$ et S_n sont indépendantes. En déduire que ce résultat reste vrai pour un échantillon *i.i.d.* de loi $\mathcal{N}(\mu, \sigma)$.

Indication. *On pourra utiliser l'indication de la question 9. de l'Exercice 2.*

..... /**1**

2. Montrer, en utilisant une transformation affine des X_i, $i \in [\![1, n]\!]$, que quelque soit $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*$, $T(\mathbf{X})$ a la même loi de probabilité. Peut-on en déduire que $T(\mathbf{X})$ est une statistique libre ?

..... /**0.5**

3. Montrer (ou admettre) que la loi de $T(\mathbf{X})$ est celle du rapport A_1/A_2 où A_1 et A_2 sont deux variables aléatoires indépendantes telles que A_1 soit de loi $\mathcal{N}(0,1)$ et $A_2 = \sqrt{B_2/(n-1)}$ avec B_2 de loi $\chi^2(n-1)$. En déduire que, pour $\nu \in \mathbb{N}^*$, la densité de la loi de Student $\mathcal{T}(\nu)$ s'exprime, pour $t \in \mathbb{R}$, comme

$$f(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\,(1+{}^{t^2}/_{\nu})^{-\frac{\nu+1}{2}}\,, \quad \text{où } \Gamma(\cdot) \text{ désigne la fonction Gamma.}$$

..... /**1**

On considère alors l'extension de la loi ci-dessus en la famille de lois indexée par (μ,τ), de densité

$$g(t;\nu,\mu,\tau) = \frac{\Gamma(\frac{\nu+1}{2})}{\tau\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{(t-\mu)^2}{\nu\tau^2}\right)^{-\frac{\nu+1}{2}} \qquad t \in \mathbb{R} \tag{1}$$

Cette nouvelle famille, notée $\mathcal{T}(\nu,\mu,\tau^2)$, est toujours appelée loi de Student. **Dans toute la suite de l'exercice, le paramètre ν, appelé degré de liberté de la loi, est supposé fixe et connu.**

4. Montrer que, si T est de loi $\mathcal{T}(\nu,\mu,\tau^2)$, alors $(T-\mu)/\tau$ est de loi $\mathcal{T}(\nu)$.

..... /**0.5**

5. En déduire que si l'échantillon *i.i.d.* $\mathbf{X} = (X_1, \ldots, X_n)$ est de loi $\mathcal{N}(\mu, \tau^2)$, alors le rapport $\sqrt{n}\,\overline{X}_n / S_n$ **ne suit pas** une loi $\mathcal{T}(n-1, \mu, \tau^2)$ *(mais une autre loi appelée Student décentrée).*

..... /**0.5**

6. Montrer que la famille associée aux densités (1), de paramètre (μ, τ) n'est pas une famille exponentielle. Est-ce que l'estimateur du maximum de vraisemblance de (μ, τ) peut être une statistique exhaustive pour n'importe quelle taille $m \in \mathbb{N}^*$ d'un échantillon $(T_1, \ldots, T_m)$ de $\mathcal{T}(\nu, \mu, \tau^2)$?

Indication. *On ne calculera pas l'estimateur du maximum de vraisemblance de* (μ, τ)*.*

..... /**1**

7. En admettant que l'information de Fisher sur (μ, τ) apportée par T de loi $\mathcal{T}(\nu, \mu, \tau^2)$ vaut

$$\mathfrak{I}(\mu, \tau) = \begin{bmatrix} \frac{\nu+1}{(\nu+3)\tau^2} & 0 \\ 0 & \frac{\nu}{2(\nu+3)\tau^4} \end{bmatrix},$$

montrer que X de loi $\mathcal{N}(\mu, \tau^2)$ est plus informatif que T sur (μ, τ).

..... /**0.5**

8. Donner un argument calculatoire pour caractériser la complexité du calcul de l'estimateur du maximum de vraisemblance de (μ, τ) fondé sur un m-échantillon $(T_1, \ldots, T_m)$ de $\mathcal{T}(\nu, \mu, \tau^2)$, $m \in \mathbb{N}^*$.

..... /**0.5**

9. Soit $T = \mu + \tau\sqrt{\nu}A_1/\sqrt{A_2}$, avec A_1 de loi $\mathcal{N}(0,1)$, indépendant de A_2, et A_2 de loi $\chi^2(\nu)$. Donner

(a) la densité de la loi conditionnelle de T conditionnellement à A_2 comme une densité normale ;

(b) la densité de la loi jointe de (T, A_2), $f^c(t, a_2; \nu, \mu, \tau)$;

(c) la densité de la loi conditionnelle de A_2 conditionnellement à T, $f^\ell(a_2; t, \nu, \mu, \tau)$ comme une densité Gamma.

Indication. *On ne redémontrera pas que T est de loi $\mathcal{T}(\nu, \mu, \tau^2)$.*

..... /**1.5**

Admis. Si A_2 est de loi $\mathcal{G}a(a,b)$, alors $\mathbb{E}_{a,b}[A_2] = a/b$ et $\mathbb{E}_{a,b}[\log A_2] = \psi(a) - \log b$, où $\psi(\cdot)$ est une fonction spéciale, dite *digamma*.

10. En vue d'implémenter l'algorithme EM pour trouver l'estimateur du maximum de vraisemblance, associé à un m-échantillon $(T_1, \ldots, T_m)$ de $\mathcal{T}(\nu, \mu, \tau^2)$, $m \in \mathbb{N}^*$, montrer que, pour $i \in [\![1, m]\!]$,

$$\begin{aligned} Q_i(\mu, \tau; t_i, \mu^0, \tau^0) &\stackrel{\text{def}}{=} \mathbb{E}_{\mu^0,\tau^0}\left[\log f^c(t, A_{2i}; \nu, \mu, \tau) \mid T_i = t_i\right] \\ &= -\frac{1}{2}\log[2\pi\Gamma(\nu/2)] - \frac{\nu}{2}\log(2) - \log(\tau) + \frac{\nu+1}{2}\mathbb{E}_{\mu^0,\tau^0}[\log(A_{2i}) \mid T_i = t_i] \\ &\quad - \left\{\frac{1}{2} + \frac{(t-\mu)^2}{2\nu\tau^2}\right\}\mathbb{E}_{\mu^0,\tau^0}[A_{2i} \mid T_i = t_i] \end{aligned}$$

où l'espérance conditionnelle est calculée sous la densité $f^\ell(a_{2i}; t_i, \nu, \mu^0, \tau^0)$ *(ce qui correspond à l'étape E).*

..... /**1**

11. Montrer que la solution de $(\mu^1, \tau^1) \stackrel{\text{def}}{=} \arg\max_{\mu,\tau} \sum_{i=1}^m Q_i(\mu, \tau; t_i, \mu^0, \tau^0)$ est

$$\mu^1 = \frac{\sum_{i=1}^m t_i \mathbb{E}_{\mu^0,\tau^0}[A_{2i} \mid T_i = t_i]}{\sum_{i=1}^m \mathbb{E}_{\mu^0,\tau^0}[A_{2i} | T_i = t_i]} \quad \text{et} \quad {\tau^1}^2 = \frac{1}{m\nu}\sum_{i=1}^m (t_i - \mu)^2 \mathbb{E}_{\mu^0,\tau^0}[A_{2i} \mid T_i = t_i]$$

(ce qui correspond à l'étape M).

..... /**1**

12. Écrire un code R implémentant l'algorithme EM, c'est-à-dire l'iteration des étapes E et M ci-dessus jusqu'à congergence à partir d'un m-échantillon `x` et une valeur `nu` fixée de ν.

..... /**2**