Université Paris-Dauphine Année 2017–2018
L3 - Statistique (1) : Estimation

Examen Final 10/01/2018 – Durée 2h00 - Aucun document autorisé

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 des copies et intercalaires doit avoir été faite au moment de la remise de chaque copie par les enseignants et surveillants. Il ne sera pas accordé de délai pour cette raison en fin d'épreuve.**

On rappelle que

— la loi normale $\mathcal{N}(\mu, \sigma^2)$ a pour densité $f_{\mu,\sigma^2}(x) = e^{-(x-\mu)^2/(2\sigma^2)}/\sqrt{2\pi}\sigma$

— la loi Exponentielle $\mathcal{E}(\lambda)$ a pour densité $f_\lambda(x) = \lambda e^{-\lambda x}\mathbf{1}_{x>0}$

— la loi $\Gamma(a, b)$ a pour densité $f_{a,b}(x) = \mathbb{I}_{x>0}x^{a-1}e^{-bx}b^a/\Gamma(a)$

— la loi de Pareto $\mathcal{P}(\alpha, \mu)$ a pour densité $f_{\alpha,\mu}(x) = \alpha\,\mu^\alpha\,x^{-\alpha-1}\,\mathbb{I}_{x\geq\mu}$

Exercice 1 (6 pts)

Dans cet exercice il vous est demandé de donner la bonne réponse en la justifiant, seules les réponses justifiées seront validées. Il n'y a pas de points négatifs.

1 Un modèle statistique est associé au vecteur $(X_{11}, X_{12}, \ldots, X_{n1}, X_{n2}))$ en supposant que $X_{i1}, X_{i2} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$, avec les μ_i et σ inconnus. Pour une observation $(x_{11}, x_{12}, \ldots, x_{n1}, x_{n2}))$, un estimateur $\hat\sigma^2$ du maximum de vraisemblance de σ^2 est donné par
(a) $^1/_{2n}\sum_{i=1}^n x_{i1}^2 + x_{i2}^2$ (b) $2(x_{11}+\ldots+x_{n1})^2/n - 2(x_{12}+\ldots+x_{n2})^2/n$ (c) $^1/_n\sum_{i=1}^n(x_{i1} - x_{i2})^2$ (d) $^1/_{2n}\sum_{i=1}^n(x_{i1} - x_{i2})^2$.

2 Si $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$ sont des échantillons iid aléatoires de lois exponentielles $\mathcal{E}(\lambda)$ et $\mathcal{E}(\lambda^2)$, respectivement, l'estimateur du maximum de vraisemblance de λ associé à l'ensemble des observations $(x_1, \ldots, x_n, y_1, \ldots, y_n)$, $\hat\lambda$ est donné par
(a) $\bar y/\bar x$ (b) $(\bar x + \sqrt{\bar y})/2$ (c) $\{\bar x\bar y\}^{1/3}$ (d) $\{\sqrt{\bar x^2 + 6\bar y} - \bar x\}/4\bar y$ (e) $\{\sqrt{\bar x^2 + 2\bar y} - \bar y\}/3\bar x$

3 Étant donnés $X_1, \ldots, X_n$ formant un échantillon iid de loi de Pareto $\mathcal{P}(\alpha, \mu)$, où α est fixe, l'estimateur $\hat\varrho$ du maximum de vraisemblance de la probabilité $\varrho = \mathbb{P}(X \in [2\mu, 4\mu])$ est donné par :
(a) $\log\{\frac{n+1}{n}\}/1 + \alpha$ (b) $\frac{(n-1)\alpha}{n(\alpha+1)}$ (c) $2^{-\alpha} - 4^{-\alpha}$ (d) $\frac{n+1}{4n}$ (e) $\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{[2,4]}(X_i/\bar X)$.

4 Étant donné $X_1, \ldots, X_n$ un échantillon iid de loi $\mathcal{E}(\lambda)$, on observe $(Z_1, \ldots, Z_n) = (\mathbb{I}(X_1 \leq \delta), \ldots, \mathbb{I}(X_n \leq \delta))$, avec δ connu. Si $z_1 = 1$ et $(z_2, \ldots, z_n) = (0, \ldots, 0)$, l'estimateur du maximum de vraisemblance de λ est donné par :
(a) $\hat\lambda = -\log\{\frac{n-1}{n}\}/\delta$ (b) $\hat\lambda = \frac{(n-1)\delta}{n}$ (c) $\hat\lambda = \frac{\log\{\delta\}}{n-1}$ (d) $\hat\lambda = \frac{n+1}{\delta}$ (e) $\hat\lambda = 0$.

5 Pour simuler une variable aléatoire X suivant une loi exponentielle $\mathcal{E}(\lambda)$, avec $\lambda > 0$, à partir d'une seule simulation uniforme $\mathcal{U}(0,1)$, U, il faut utiliser la transformation
(a) $X = \lambda \log(U)$; (b) $X = -\log[(1-U)]^{\lambda}$; (c) $X = -\log(1-U)/\lambda$; (d) $X = \log[\lambda(1-U)]$; (e) $X = -\log(U/\lambda)$; (f) $X = \exp\{-\log(U)^{\lambda}\}$.

6 Soit X une variable aléatoire de loi normale $\mathcal{N}(0,1)$. Quelle commande R permet d'obtenir t tel que $\mathbb{P}(|X| \leq t) = 0.90$?

Réponse (a)

```
> pnorm(0.95)
```

Réponse (b)

```
> qnorm(0.90)
```

Réponse (c)

```
> qnorm(0.95)
```

Réponse (d)

```
> pnorm(0.90)
```

7 Un vecteur `x` de R contient n réalisations indépendantes de la loi uniforme sur $[-\theta, \theta]$ pour un paramètre θ inconnu. Parmi les codes suivants, lequel donne une estimation bootstrap du biais du maximum de vraisemblance de θ ?

Réponse (a)

```
> delta <- numeric(K)
> for(i in 1:K){
    y <- sample(x, replace = FALSE)
    delta[i] <- max(y)
  }
> mean(delta)-max(x)
```

Réponse (b)

```
> delta <- numeric(K)
> for(i in 1:K){
    y <- sample(x, replace = TRUE)
    delta[i] <- mean(abs(y))
  }
> max(abs(delta))-max(abs(x))
```

Réponse (c)

```
> delta <- numeric(K)
> for(i in 1:K){
    y <- sample(x, replace = TRUE)
    delta[i] <- mean(y)
  }
> max(delta)-max(x)
```

Réponse (d)

```
> delta <- numeric(K)
> for(i in 1:K){
    y <- sample(x, replace = FALSE)
    delta[i] <- max(abs(y))
  }
> mean(delta)-max(abs(x))
```

Réponse (e)

```
> delta <- numeric(K)
> for(i in 1:K){
    y <- sample(x, replace = TRUE)
    delta[i] <- max(abs(y))
  }
> mean(delta)-max(abs(x))
```

Réponse (f)

```
> delta <- numeric(K)
> for(i in 1:K){
    y <- sample(x, replace = TRUE)
    delta[i] <- max(y)
  }
> mean(delta)-max(x)
```

∞ Étant donné $X_1, \ldots, X_n$ un échantillon iid de loi $\mathcal{E}(\lambda)$, on observe $(Z_1, \ldots, Z_n) = (\mathbb{I}(X_1 \leq \delta), \ldots, \mathbb{I}(X_n \leq \delta))$, avec δ connu. Si $z_1 = 1$ et $(z_2, \ldots, z_n) = (0, \ldots, 0)$, l'estimateur du maximum de vraisemblance de λ est donné par :
(a) $\hat{\lambda} = -\log\{\frac{n-1}{n}\}/\delta$ (b) $\hat{\lambda} = \frac{(n-1)\delta}{n}$ (c) $\hat{\lambda} = \frac{\log\{\delta\}}{n-1}$ (d) $\hat{\lambda} = \frac{n+1}{\delta}$ (e) $\hat{\lambda} = 0$.

Exercice 2 (7 pts)

Soient $X_1, \dots, X_n$, i.i.d dont la loi admet la densité

$$f_{\mu,\lambda}(x) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right) \mathbb{I}_{x>0}$$

où $\lambda > 0$ et $\mu > 1$ sont deux paramètres. On a $\mathbb{E}(X_1) = \mu$ et $\mathbb{V}(X_1) = \frac{\mu^3}{\lambda}$.

1 (1,5 point) Montrer que ce modèle appartient à la famille exponentielle. On donnera la statistique exhaustive et le paramètre naturel. En déduire que le modèle est régulier.

2 (1 point) On suppose λ connu. Donner l'estimateur du maximum de vraisemblance de μ.

3 (1 point) On suppose λ connu. Calculer l'information de Fisher associée au modèle pour le paramètre μ et montrer que l'estimateur du maximum de vraisemblance de μ est fortement consistant et asymptotiquement efficace.

Rappel : un estimateur $\hat{\mu}_n$ de μ est asymptotiquement efficace si $\sqrt{n}\left(\hat{\mu}_n - \mu\right)$ converge en loi vers une loi normale de moyenne nulle et de variance l'inverse de l'information de Fisher associée à X_1.

4 (2 points) **On ne suppose plus λ connu** dans la suite. Donner les estimateurs du maximum de vraisemblance de μ et λ.

5 (1,5 point) Calculer la matrice d'information de Fisher.

Exercice 3 (6 pts)

Soient $X_1, \dots, X_n$ n variables aléatoires i.i.d de densité

$$f_\theta(x) = \theta^2 x e^{-\theta x} \mathbb{I}_{x \geq 0}$$

où $\theta > 0$.

1 (1 point) Calculer $\mathbb{E}(X)$ et en déduire un estimateur fortement consistant de θ que l'on notera $\hat{\theta}_n^{MM}$.

2 (0.5 point) Montrer que $\hat{\theta}_n^{MM}$ est une statistique exhaustive.

3 (1 point) Calculer l'estimateur du maximum de vraisemblance de θ que l'on notera $\hat{\theta}_n^{MV}$ et montrer qu'il est fortement consistant.

4 (0.5 point) Montrer que $\mathbb{V}(X_1) = \frac{2}{\theta^2}$.

5 (1.5 point) Montrer que $\sqrt{n}\left(\hat{\theta}_n^{MV} - \theta\right)$ converge en loi vers une loi normale de moyenne 0 et de variance δ que l'on précisera.

6 (1.5 point) Montrer que le modèle est régulier et calculer l'information de Fisher du modèle d'échantillonnage. Comparer l'information de Fisher à δ.

Exercice 4 (8 pts)

Pour $X \sim f(x|\theta)$, on rappelle (i) la notion de statistique *complète* : *Une statistique $T(X)$ est complète si les seules fonctions de $T(X)$, $\Psi(T(X))$, à espérance constante $\mathbb{E}_\theta[\Psi(T(X))] = c$, sont les fonctions constantes, $\Psi(t) = c$* et (ii) le théorème de Basu : *Toute statistique exhaustive et complète $T(X)$ est indépendente de toute statistique libre $A(X)$.*

1 (1 point) En appliquant ce théorème de Basu au cas d'un échantillon Normal $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ lorsque σ^2 est connue, montrer que $\bar{X}_n = {}^1\!/\!_n \sum_{i=1}^n X_i$ est indépendante de $S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2$. (On admettra que $\bar{X}_n$ est une statistique complète.)

2 (0.5 point) En déduire que $\bar{X}_n$ et S_n^2 sont aussi indépendants quand σ^2 est inconnu.

3 (2 points) Démontrer que, s'il existe une statistique exhaustive minimale pour un échantillon $X_1, \ldots, X_n$, $T(X_1, \ldots, X_n)$, alors toute statistique complète est aussi exhaustive minimale. *(Note : Cette question n'a pas d'impact sur les questions suivantes.)*

4 (1.5 point) On approche la variance de la médiane empirique $\hat{\text{med}}_n$ de l'échantillon Normal $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ par Monte Carlo, c'est à dire, en générant N échantillons Normaux $\mathcal{N}(0, 1)$ de taille n, calculant la médiane $\hat{\text{med}}_n$ pour chacun des N échantillons et en calculant la variance empirique de cette collection de médianes.

1. Expliquer pourquoi on peut se ramener à $\mu = 0$ et $\sigma = 1$.
2. Ecrire un code R permettant de réaliser cette expérience de Monte Carlo.

5 (2 point) On répète l'expérience précédente en estimant la variance de la différence entre la médiane empirique et $\bar{X}_n$ par les mêmes étapes de Monte Carlo puis en rajoutant la variance connue de $\bar{X}_n$.

1. Rappeler la variance de $\bar{X}_n$.
2. Expliquer pourquoi
$$\text{Var}(\hat{\text{med}}_n) = \text{Var}(\hat{\text{med}}_n - \bar{X}_n) + \text{Var}(\bar{X}_n)$$
en utilisant le théorème de Basu.
3. Ecrire un code R permettant de réaliser cette expérience de Monte Carlo.

French – English Lexicon

échantillon – sample
famille exponentielle – exponential family
statistique complète – complete statistic
statistique exhaustive – sufficient statistic
statistique libre – ancilary statistic
vraisemblance – likelihood