Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2017 – 2018

Statistiques I

Contrôle continu n°1, novembre 2017

Examen de 1h30. Tout document ou calculatrice est interdit.

1. **(Sur 12 points)** Soit $(X_1, \dots, X_n)$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant une loi de Weibull, c'est-à-dire une loi dont la fonction de répartition est:

$$F(x) = 1 - \exp\Big(- \Big(\frac{x}{\lambda}\Big)^{\beta}\Big) \quad \text{pour } x \geq 0,$$

avec $\theta = (\lambda, \beta) \in]0, \infty[^2$ inconnu.

   (a) Déterminer $F(x)$ pour $x < 0$ **(0.5pts)**.

   (b) Déterminer le modèle statistique paramétrique induit par $(X_1, \dots, X_n)$ **(0.5pts)** et montrer qu'il est dominé **(1pt)**.

   (c) Préciser la vraisemblance de ce modèle **(1pt)**.

   (d) Le modèle fait-il partie de la famille exponentielle? (justifier...) **(3pts)**

   (e) Montrer que la statistique $\widehat{T} = (X_1, X_2, \dots, X_n)$ est exhaustive minimale pour ce modèle **(4pts)**.

   (f) On suppose que β est connu. Préciser alors le modèle statistique et déterminer une statistique exhaustive complète pour ce modèle **(2pts)**.

Proof. (a) $F(x) = 0$ pour $x < 0$ car $F(0) = 0$ et F croissante.

(b) Le modèle paramétrique induit est $\big([0, \infty[^n, \mathcal{B}([0, \infty[^n), \mathbb{P}_\theta^{\oplus n}\big)$ où $\theta = (\lambda, \beta) \in]0, \infty[^2$ car les variables sont iid, chacune de loi $\mathbb{P}_\theta$. Comme F est absolument continue, $F(x) = \int_0^x f_\theta(t)dt$ où $f_\theta(x) = \frac{\beta}{\lambda}\Big(\frac{x}{\lambda}\Big)^{\beta-1} \exp\Big(- \Big(\frac{x}{\lambda}\Big)^{\beta}\Big)$ pour $x \geq 0$. La mesure qui domine est la mesure de Lebesgue sur $[0, \infty[^n$.

(c) Pour $(x_1, \dots, x_n) \in [0, \infty[^n$, $L_\theta(x_1, \dots, x_n) = \Big(\frac{\beta}{\lambda}\Big)^n \exp\Big(- \sum_{i=1}^n \Big(\frac{x_i}{\lambda}\Big)^{\beta}\Big) \prod_{i=1}^n \Big(\frac{x_i}{\lambda}\Big)^{\beta-1}$ car vaiid.

(d) On peut encore écrire que $L_\theta(x_1, \dots, x_n) = \exp\Big(n \log(\beta/\lambda) + n(1-\beta)\log\lambda + (\beta-1)\sum_{i=1} \log x_i - \sum_{i=1}^n \Big(\frac{x_i}{\lambda}\Big)^{\beta}\Big)$. On peut donc noter $\beta(\theta) = n\log(\beta/\lambda) + n(1-\beta)\log\lambda$, $b(x_1, \dots, x_n) = 0$, $a_1(x_1, \dots, x_n) = \sum_{i=1} \log x_i$ et $\alpha_1(\theta) = \beta - 1$ qui correspondent à un modèle de la famille exponentielle, mais $\sum_{i=1}^n \Big(\frac{x_i}{\lambda}\Big)^{\beta}$ ne peut pas s'écrire sous la forme $a_2(x_1, \dots, x_n)\alpha_2(\theta)$. Cela peut se montrer pour $n = 1$: si $x^\beta = f(\beta)g(x)$ pour tout $\beta > 0$ et tout $x \geq 0$, on aurait $g(x) = f(2)x^2 = f(1)x$ ce qui n'est clairement pas possible pour tout $x \geq 0$.

(e) La statistique $\widehat{T}$ est exhaustive (voir le cours). Pour montrer qu'elle est minimale, on écrit:

$$\frac{L_\theta(x_1, \dots, x_n)}{L_\theta(y_1, \dots, y_n)} = \exp\Big((\beta-1)\sum_{i=1} \log(x_i/y_i) + \sum_{i=1}^n \Big(\frac{y_i}{\lambda}\Big)^{\beta} - \Big(\frac{x_i}{\lambda}\Big)^{\beta}\Big).$$

Pour que ce rapport ne dépende pas de θ, cela revient à ce que pour $\lambda = 1$, $(\beta-1)\sum_{i=1} \log(x_i/y_i) + \sum_{i=1}^n y_i^\beta - x_i^\beta$ ne dépende pas de β. On peut dériver deux fois par rapport à β, et on devra avoir $\sum_{i=1}^n \log^2(y_i) y_i^\beta - \log^2(x_i) x_i^\beta = 0$ pour tout $\beta > 0$. Comme ceci doit être vrai quelque soient les n-uplets $(x_1, \dots, x_n)$ et $(y_1, \dots, y_n)$ on en déduit (en prenant par exemple tous les x_i et y_i égaux à 1 sauf pour $i = j$), que $x_j = y_j$ nécessairement, et ceci pour tout j. Donc $\widehat{T}(x_1, \dots, x_j) = \widehat{T}(y_1, \dots, y_j)$: la statistique est bien minimale.

(f) Si β est connu, alors le modèle est le même que précédemment mais $\theta = \lambda \in]0, \infty[$.
Ce modèle appartient à la famille exponentielle avec la décomposition précédente. Comme β n'est plus inconnu, alors $b(x_1, \dots, x_n) = (\beta - 1) \sum_{i=1}^n \log x_i$ et $-\sum_{i=1}^n \left(\frac{x_i}{\lambda}\right)^\beta = a_1(x_1, \dots, x_n)\alpha_1(\theta)$, avec $\alpha_1(\lambda) = -\lambda^{-\beta}$ et $a_1(x_1, \dots, x_n) = \sum_{i=1}^n x_i^\beta$. On peut alors noter $\widehat{T} = \sum_{i=1}^n x_i^\beta$. Ce sera une statistique exhaustive complète si l'ensemble des $\alpha(\lambda) = -\lambda^{-\beta}$ pour $\lambda > 0$ est d'intérieur non nul, ce qui est vrai car c'est $]-\infty, 0[$.

□

2. **(Sur 16 points)** Soit une suite $(\varepsilon_i)_{n\geq 1}$ de variables aléatoires indépendantes et identiquement distribuées suivant une loi de Bernoulli de paramètre $p \in]0, 1[$ inconnu. On définit également la suite $(X_n)_{n\geq 1}$ par

$$X_{n+1} = \varepsilon_{n+1} X_n \quad \text{pour } n \in \mathbf{N}^*, \text{ et} \quad X_1 = \varepsilon_1.$$

On observe $(X_1, \dots, X_n)$ où $n \in \mathbf{N}^*$.

(a) Pour $k \in \mathbf{N}^*$ fixé, écrire X_k en fonction des ε_i **(0.5pts)**. Quelles sont les valeurs possibles pour X_k **(0.5pts)**? Quelle est sa loi **(0.5pts)**? Les variables (X_k) sont-elles indépendantes **(1.5pts)**? Identiquement distribuées **(0.5pts)**?

(b) Montrer que le vecteur $(X_1, \dots, X_n)$ ne peut prendre que $(n + 1)$ valeurs que l'on précisera **(1pt)**. En déduire le modèle statistique induit par ce vecteur (on n'explicitera pas sa mesure de probabilité) **(0.5pts)**.

(c) Démontrer que pour $(x_1, \dots, x_n)$ appartenant au modèle statistique **(1pt)**:

$$\mathbb{P}\Big((X_1, \dots, X_n) = (x_1, \dots, x_n)\Big) = \mathbb{P}(X_1 = x_1) \prod_{k=2}^n \mathbb{P}\Big(X_k = x_k \,|(X_1, \dots, X_{k-1}) = (x_1, \dots, x_{k-1})\Big).$$

(d) Montrer que $\mathbb{P}\Big(X_k = x_k \,|(X_1, \dots, X_{k-1}) = (x_1, \dots, x_{k-1})\Big) = \mathbb{P}\Big(X_k = x_k \,|X_{k-1} = x_{k-1}\Big)$ **(1pt)**, puis que $\mathbb{P}\Big(X_k = x_k \,|X_{k-1} = x_{k-1}\Big) = p^{x_k}(1-p)^{x_{k-1}-x_k}$ pour $k \geq 2$ **(2pts)**.

(e) En déduire l'expression de la vraisemblance du modèle **(1pt)**.

(f) Le modèle appartient-il à la famille exponentielle **(1pt)**?

(g) Peut-on en déduire que $\widehat{T} = (\sum_{i=1}^n X_i\,,\, X_n)$ est une statistique exhaustive complète pour le modèle **(3pts)**?

(h) Déterminer l'estimateur par maximum de vraisemblance de p (c'est-à-dire la valeur de $p \in]0, 1[$ qui maximise la vraisemblance du modèle prise en $(X_1, \dots, X_n)$) et vérifier que c'est une fonction de $\widehat{T}$ **(2pts)**.

Proof. (a) On a facilement $X_k = \varepsilon_1 \times \cdots \times \varepsilon_k$ pour tout $k \geq 1$.
Il est clair que X_k ne peut prendre pour valeurs que 0 ou 1. Elle suit donc une loi de Bernoulli. Par ailleurs, $\mathbb{P}(X_k = 1) = p^k$, donc X_k suit une loi de Bernoulli de paramètre p^k.
On a $\mathbb{P}(X_1 = 0 \cap X_2 = 1) = 0 \neq \mathbb{P}(X_1 = 0) \cap \mathbb{P}(X_2 = 0) = (1-p)p^2$, donc X_1 et X_2 sont indépendantes.
Comme $P(X_k = 1) = p^k$, dépend de k, la loi de X_k dépend de k: les variables ne sont pas identiquement distribuées.

(b) Il est clair que si $X_k = 0$ alors $X_{k+i} = 0$ pour tout $i \geq 0$. Donc $(X_1, \dots, X_n)$ ne peut prendre que pour valeurs que les n-uplets $(0, \dots, 0)$, $(1, \dots, 1)$ ou tous ceux de type $(1, \dots, 1, 0, \dots, 0)$. Cela en fait bien $(n + 1)$.
Le modèle statistique induit est donc $(\Omega_n, \mathcal{P}(\Omega_n), P_\theta^{(n)})$ où $\theta = p$, Ω_n l'ensemble des $n+1$ n-uplet défini au-dessus, $\mathcal{P}(\Omega_n)$ l'ensemble des parties de cet ensemble et $P_\theta^{(n)}$ la loi de probabilité de $(X_1, \dots, X_n)$.

(c) On a $\mathbb{P}(A \cap B) = \mathbb{P}(B)\,\mathbb{P}(A \mid B)$ pour tous évènements A et B. Donc $\mathbb{P}\Big(X_n = x_n \cap \{X_{n-1} = x_{n-1}, \dots, X_1 = x_1\}\Big) = \mathbb{P}\Big(X_{n-1} = x_{n-1}, \dots, X_1 = x_1\Big)\,\mathbb{P}\Big(X_n = x_n \mid \{X_{n-1} = x_{n-1}, \dots, X_1 = x_1\}\Big)$ et on itère l'écriture.

(d) On a $\mathbb{P}\Big(X_k = x_k \,|(X_1, \dots, X_{k-1}) = (x_1, \dots, x_{k-1})\Big) = \mathbb{P}\Big(\varepsilon_k X_{k-1} = x_k \,|X_{k-1} = x_{k-1}\Big)$. Or ε_k est indépendant de $(X_1, \dots, X_{k-1})$ donc $\varepsilon_k X_{k-1}$ ne dépend que de X_{k-1}, d'où le résultat.
On a $\mathbb{P}\Big(X_k = x_k \,|X_{k-1} = x_{k-1}\Big) = 1$ pour $x_{k-1} = 0$ et tout $x_k \in \{0, 1\}$, $\mathbb{P}\Big(X_k = x_k \,|X_{k-1} = x_{k-1}\Big) = p$

pour $x_{k-1} = 1$ et $x_k = 1$, et $\mathbb{P}\Big(X_k = x_k \,|X_{k-1} = x_{k-1}\Big) = 1-p$ pour $x_{k-1} = 1$ et $x_k = 0$. La formule $\mathbb{P}\Big(X_k = x_k \,|X_{k-1} = x_{k-1}\Big) = p^{x_k}(1-p)^{x_{k-1}-x_k}$ permet de retrouver ces 3 cas.

(e) En utilisant la formule établie plus haut, on en déduit que la vraisemblance du modèle est

$$L_\theta(x_1,\dots,x_n) = p^{x_1}(1-p)^{1-x_1}\prod_{k=2}^{n} p^{x_k}(1-p)^{x_{k-1}-x_k} = p^{\sum_{k=1}^n x_k}(1-p)^{1-x_n}.$$

(f) On peut également écrire que

$$L_\theta(x_1,\dots,x_n) = \exp\Big(\log(1-p) + \log p \sum_{k=1}^{n} x_k - x_n\log(1-p)\Big).$$

Donc le modèle appartient à la famille exponentielle.

(g) D'après cette écriture on sait que $\widehat{T} = \big(\sum_{i=1}^n X_i\,,\, X_n\big)$ sera une statistique exhaustive complète si $\alpha(p) = (\log p, -\log(1-p))$ décrit un ensemble d'intérieur non nul quand p varie dans $]0,1[$, ce qui n'est pas le cas! (c'est une courbe dans un espace de dimension 2 donc on ne peut pas trouver une boule ouverte contenu dans cette courbe). On ne peut donc pas en déduire que la statistique est complète.

(h) On peut dériver $p \in]0,1[\mapsto \log(1-p) + \log p \sum_{k=1}^n x_k - x_n\log(1-p)$ et on obtient

$$\frac{X_n - 1}{1-\widehat{p}} + \frac{1}{\widehat{p}}\sum_{k=1}^{n} X_k = 0 \quad\Longleftrightarrow\quad \widehat{p} = \frac{\sum_{k=1}^n X_k}{1 - X_n + \sum_{k=1}^n X_k},$$

qui est bien une fonction de $\widehat{T}$. C'est bien un maximum car la dérivée seconde est $\frac{X_n-1}{(1-\widehat{p})^2} - \frac{1}{\widehat{p}^2}\sum_{k=1}^n X_k$ qui est négative strictement.

□