Université Paris I, Panthéon - Sorbonne

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

Statistiques I

Examen final, janvier 2018

Examen de 3h. Tout document ou calculatrice est interdit.

1. **(Sur 12 points)** Soit la variable X qui suit une loi normale $\mathcal{N}(m\,,\,\sigma^2)$, où $m \in \mathbf{R}$ et $\sigma^2 \in]0,\infty[$ sont deux paramètres inconnus. Soit également une suite $(\varepsilon_k)_k$ de variables i.i.d. de loi $\mathcal{N}(0\,,\,\sigma^2)$, indépendantes de X. Au final, on note $\theta = {}^t(m,\sigma^2)$ et on observe $(X_1,\cdots,X_n)$ défini par

$$X_k = X + \varepsilon_k \quad \text{pour tout } k \in \mathbf{N}.$$

   (a) Déterminer le modèle statistique paramétrique dominé induit par $(X_1,\cdots,X_n)$ après avoir montré que la loi de probabilité de $(X_1,\cdots,X_n)$ est gaussienne **(1.5pts)**.

   (b) Déterminer la loi de X_i pour tout i **(0.5pts)**.

   (c) Démontrer que Σ_n matrice de variance-covariance de $(X_1,\cdots,X_n)$ est une matrice avec $2\sigma^2$ sur la diagonale et σ^2 partout ailleurs **(0.5pts)**. Montrer que $\det(\Sigma_n) = (n+1)\sigma^{2n}$ **(1.5pts)**.

   (d) Démontrer que Σ_n^{-1} est la matrice avec $\frac{1}{\sigma^2}\,\frac{n}{n+1}$ sur la diagonale et $-\frac{1}{\sigma^2}\,\frac{1}{n+1}$ partout ailleurs **(1pt)**.

   (e) En déduire la vraisemblance de $(X_1,\cdots,X_n)$ **(1pt)**. Montrer que le modèle appartient à la famille exponentielle **(2.5pts)** et démontrer que $\big(\sum_{i=1}^n X_i\,,\,n\sum_{i=1}^n X_i^2 - 2\sum_{1\le i<j\le n} X_iX_j\big)$ est une statistique exhaustive complète pour ce modèle **(0.5pts)**.

   (f) Démontrer que $\widehat{m}_n = \frac{1}{n}\sum_{i=1}^n X_i$ est l'estimateur de m sans biais uniformément de variance minimale parmi les estimateurs sans biais de m **(1pt)**. Déterminer la loi de $\widehat{m}_n$ **(1pt)**. Est-ce un estimateur convergent **(1pt)**?

Proof. (a) Le vecteur $(X_1,\cdots,X_n)$ est la somme de deux vecteurs gaussiens indépendants: $(X,\dots,X)$ et $(\varepsilon_1,\dots,\varepsilon_n)$. C'est donc un vecteur gaussien. Le modèle statistique est donc $(\mathbf{R}^n,\mathcal{B}(\mathbf{R}^n),\mathbb{P}_\theta^{(n)})$, dominé par la mesure de Lebesgue sur $\mathbf{R}^n$.

(b) Il est clair que $X_i \overset{\mathcal{L}}{\simeq} \mathcal{N}(m,2\sigma^2)$

(c) On a $\Sigma_n = (\mathrm{cov}(X_i,X_j))_{1\le i,j\le n}$. D'après la question précédente, pour $i=j$, $\mathrm{cov}(X_i,X_j) = 2\sigma^2$. Pour $i\neq j$, $\mathrm{cov}(X_i,X_j) = \mathrm{cov}(X,X) = \sigma^2$ car les ε_k sont indépendants et indépendants de X.
Si on note $D_n = \det(\Sigma_n)$, on peut écrire la dernière ligne comme la somme $L_1+\cdots+L_n$. On obtient ainsi $(n+1)\sigma^2$ partout sur la dernière ligne. En conséquence:

$$D_n = (n+1)\sigma^{2n}\begin{vmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{vmatrix} = (n+1)\sigma^{2n}\begin{vmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{vmatrix} = (n+1)\sigma^{2n}.$$

(d) Il suffit de faire le produit matriciel pour le vérifier.

(e) On a ainsi $L_\theta(X_1,\dots,X_n) = (2\pi)^{-n/2}\big((n+1)\sigma^{2n}\big)^{-1/2}\exp\Big(-\frac{1}{2}\,{}^t(X-m\,J)\sigma_n^{-1}(X-m\,J)\Big)$, avec $J={}^t(1,\dots,1)$. On peut simplifier et ainsi $L_\theta(X_1,\dots,X_n) = \sqrt{\dfrac{(2\pi)^{-n}}{(n+1)\sigma^{2n}}}\exp\Big(-\dfrac{1}{2(n+1)\sigma^2}\Big(n\sum_{i=1}^n(X_i-m)^2 - 2\sum_{1\le i<j\le n}(X_i-m)(X_j-m)\,J\Big)\Big)$.

On peut décomposer les sommes pour pouvoir ainsi écrire:

$$L_\theta(X_1,\dots,X_n) = \sqrt{\frac{(2\pi)^{-n}}{(n+1)\sigma^{2n}}}\exp\Big\{-\frac{1}{2(n+1)\sigma^2}\Big(\Big(n\sum_{i=1}^n X_i^2 - 2\sum_{1\le i<j\le n}X_iX_j\Big) - 2(n-1)m\sum_{i=1}^n X_i + nm^2\Big)\Big\}.$$

On en déduit que le modèle appartient bien à la famille exponentielle avec $\alpha_1(\theta) = \frac{(n-1)m}{(n+1)\sigma^2}$, $\alpha_2(\theta) = -\frac{1}{2(n+1)\sigma^2}$ et $a_1(X_1,\dots,X_) = \sum_{i=1}^n X_i$, $a_2(X_1,\dots,X_) = n\sum_{i=1}^n X_i^2 - 2\sum_{1\leq i<j\leq n} X_iX_j$ et $\beta(\theta) = -n\,\log(\sigma^2) - \frac{nm^2}{2(n+1)\sigma^2}$.
On trouve ainsi que $\alpha(\Theta) = \alpha(\mathbf{R}\times]0,\infty[) = \mathbf{R}\times]-\infty,0[$ donc d'intérieur non vide: la statistique est bien exhaustive complète.

(f) On sait d'après le Lemme de Sheffé que l'estimateur sans biais de variance minimale uniformément parmi tous les estimateurs sans biais sera une fonction d'une statistique exhaustive complète. Or $\mathbb{E}(\overline{X}_n) = m$, donc $\overline{X}_n$ est bien un estimateur sans biais de variance minimale uniformément parmi tous les estimateurs sans biais de m.
Comme $(X_1,\dots,X_n)$ est un vecteur gaussien, on en déduit que $\overline{X}_n$ est une variable gaussienne. Sa variance est $\mathrm{var}(\overline{X}_n) = \frac{1}{n^2}\left(\sum_{i=1}^n \mathrm{var}(X_i) + 2\sum_{1\leq i<j\leq n}\mathrm{cov}(X_i,X_j)\right) = \frac{1}{n^2}\left(2n\,\sigma^2 + (n^2-n)\sigma^2\right) = \left(1+\frac{1}{n}\right)\sigma^2$. Donc $\overline{X}_n \overset{\mathcal{L}}{\simeq} \mathcal{N}\left(m\,,\,\left(1+\frac{1}{n}\right)\sigma^2\right)$.
Comme $\lim_{n\to\infty}\mathrm{var}(\overline{X}_n) = \sigma^2$, on en déduit que $\overline{X}_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(m\,,\,\sigma^2\right)$: ce n'est pas un estimateur convergent de m.

□

2. **(Sur 18 points)** Soit une suite $(\varepsilon_{ij})_{i,j\in\mathbf{N}}$ de v.a.i.i.d. de loi Bernoulli de paramètre $p\in]0,1[$. Pour $m\in\mathbf{N}$, on définit

$$X_j = \sum_{i=1}^m \varepsilon_{ij} \quad \text{pour } j\in\mathbf{N}^*$$

(par convention $\sum_{i=1}^0 \varepsilon_{ij} = 0$). On observe $(X_1,\dots,X_n)$ où $n\in\mathbf{N}^*$.

(a) Pour $m\in\mathbf{N}$ fixé, démontrer que $(X_j)_{j\in\mathbf{N}}$ forme une suite de v.a.i.i.d. de loi à préciser **(1pt)**.

(b) Tout d'abord, on suppose que m est connu et p inconnu. Déterminer le modèle statistique induit par $(X_1,\dots,X_n)$ **(0.5pts)**, montrer qu'il est dominé **(0.5pts)** et appartient à la famille exponentielle **(0.5pts)**. Déterminer un estimateur de p sans biais et efficace **(1pt)** en précisant la borne de Cramèr-Rao **(0.5pts)**.

(c) On suppose maintenant que m est également inconnu et on note $\theta = {}^t(p,m)$. Déterminer le modèle statistique induit par $(X_1,\dots,X_n)$ **(0.5pts)**, montrer qu'il est dominé **(0.5pts)** et qu'il n'appartient pas à la famille exponentielle **(0.5pts)**. Soit l'estimateur $\widehat{\theta}_n = {}^t(\widehat{p}_n,\widehat{m}_n)$ avec $\widehat{m}_n = \max_{1\leq i\leq n} X_i$ et $\widehat{p}_n = \frac{1}{\widehat{m}_n n}\sum_{i=1}^n X_i$. Montrer que $\mathbb{P}(\max_{1\leq i\leq n} X_i = m) = 1-(1-p^m)^n$ **(1.5pts)**. En déduire que $\widehat{m}_n$ est un estimateur convergent de m **(0.5pts)**. En déduire que $\widehat{p}_n$ est également un estimateur convergent de p (utiliser par exemple la fonction de répartition de $\widehat{p}_n$...) **(2.5pts)**.

(d) On suppose enfin que

$$X_j = \sum_{i=1}^{T_j} \varepsilon_{ij} \quad \text{pour } j\in\mathbf{N}^*$$

où les (T_j) sont des v.a.i.i.d. de loi de Poisson de paramètre $\lambda>0$ inconnu, indépendantes des (ε_{ij}). Le vecteur de paramètre inconnu est maintenant $\theta = {}^t(p,\lambda)$. Déterminer le modèle statistique induit par $(X_1,\dots,X_n)$ **(0.5pts)**, montrer qu'il est dominé **(0.5pts)**. Montrer que pour $m\geq k$ alors $\mathbb{P}(X_j = k\cap T_j = m) = e^{-\lambda}\frac{\lambda^m}{k!(m-k)!}p^k(1-p)^{m-k}$ **(1.5pts)**. En déduire que $(X_j)_j$ est une suite de v.a.i.i.d. de loi de Poisson de paramètre $\lambda\,p$ (on pourra faire le changement d'indice $m' = m-k$...) **(1.5pts)**. En déduire que le modèle est exponentiel **(1pt)**. Montrer que $\sum_{i=1}^n X_i$ est une statistique exhaustive et complète pour le modèle **(1pt)**. Démontrer cependant qu'il n'y a pas unicité de l'estimateur du maximum de vraisemblance de $\theta = {}^t(p,\lambda)$ et non convergence de n'importe lequel d'entre eux **(2pts)**.

Proof. (a) On sait que la somme de m v.a.i.i.d. Bernoulli de paramètre p est bien une binomiale $\mathcal{B}(m,p)$, loi de chaque X_j. De plus les $(\varepsilon_{ij})_{i,j\in\mathbf{N}}$ étant indépendantes, on en déduit que les X_j le sont également.

(b) Le modèle est alors $\left(\{0,\dots,m\}^n,\mathcal{P}(\{0,\dots,m\}^n),\mathcal{B}(m,p)^{\otimes n}\right)_p$, dominé par la mesure de comptage sur $\{0,\dots,m\}^n$.
Du fait de l'indépendance, sa vraisemblance est $L_p(x_1,\dots,x_n) = \prod_{j=1}^n \begin{pmatrix} m \\ x_j\end{pmatrix} p^{x_j}(1-p)^{m-x_j} = \left(\prod_{j=1}^n \begin{pmatrix} m \\ x_j\end{pmatrix}\right) p^{\sum_{j=1}^n x_j}(1-p)^{nm-\sum_{j=1}^n x_j}$. En passant à l'exponentielle, on a $L_p(x_1,\dots,x_n) = \exp\left\{\sum_{j=1}^n \log\begin{pmatrix} m \\ x_j\end{pmatrix} + nm\log(1-p) + \left(\log(p)-\log(1-p)\right)\sum_{j=1}^n x_j\right\}$: le modèle appartient à la famille exponentielle avec $a_1(X_1,\dots,X_n) = \sum_{j=1}^n x_j$, $\alpha_1(p) = \left(\log(p)-\log(1-p)\right)$ et $\beta(p) = nm\log(1-p)$.
La fonction de p que l'on peut estimer efficacement à une transformation affine près est $g(p) = \beta'(p)/\alpha_1'(p) = -p/nm$. Ainsi p peut être estimé efficacement. Or $\mathbb{E}\left(\frac{1}{nm}\sum_{j=1}^n X_j\right) = p$, donc $\widehat{p}_n = \frac{1}{nm}\sum_{j=1}^n X_j$ est un estimateur efficace de p.
Sa variance est $\frac{m}{n}\,p(1-p)$ qui est donc également la borne de Cramèr-Rao atteinte par $\widehat{p}_n$.

(c) Le modèle est alors $\left(\mathbf{N}^n,\mathcal{P}(\mathbf{N}^n),\mathcal{B}(m,p)^{\otimes n}\right)_\theta$.
Il est dominé par la mesure de comptage sur $\mathcal{P}(\mathbf{N}^n)$.
La vraisemblance du modèle est pour $(x_1,\dots,x_n)\in\mathbf{N}^n$, $L_\theta(x_1,\dots,x_n) = \left(\prod_{j=1}^n \begin{pmatrix} m \\ x_j\end{pmatrix}\right) p^{\sum_{j=1}^n x_j}(1-p)^{nm-\sum_{j=1}^n x_j}\,\mathbb{I}_{\max_{1\leq j\leq n}(x}$

ce n'est pas un modèle appartenant à la famille exponentielle car cette vraisemblance peut s'annuler (le support de la loi dépend de θ).
On a $\mathbb{P}(\max_{1\leq i\leq n} X_i = m) = 1 - \mathbb{P}(\max_{1\leq i\leq n} X_i \leq m-1) = 1 - \prod_{j=1}^n \mathbb{P}(X_j \leq m-1) = 1 - \big(1 - \mathbb{P}(X_j = m)\big)^n$. Puisque $\mathbb{P}(X_j = m) = p^m$, on déduit $\mathbb{P}(\max_{1\leq i\leq n} X_i = m) = 1 - (1-p^m)^n$.
Comme $0 < p^m < 1$, on a $\mathbb{P}(\max_{1\leq i\leq n} X_i = m) \underset{n\to+\infty}{\longrightarrow} 1$, donc $\max_{1\leq i\leq n} X_i \overset{\mathcal{P}}{\underset{n\to+\infty}{\longrightarrow}} m$: l'estimateur est bien convergent.
Enfin, à partir de la loi de grands nombres, on sait que pour $\widehat{m}_n = m$, alors $\widehat{p}_n \overset{\mathcal{P}}{\underset{n\to+\infty}{\longrightarrow}} p$. Donc on peut écrire que $\mathbb{P}(\widehat{p}_n \leq x \mid \widehat{m}_n = m) \underset{n\to+\infty}{\longrightarrow} F_{\delta_p}(x)$. Mais $\mathbb{P}(\widehat{p}_n \leq x) = \mathbb{P}(\widehat{p}_n \leq x \mid \widehat{m}_n = m)\mathbb{P}(\widehat{m}_n = m) + \mathbb{P}(\widehat{p}_n \leq x \mid \widehat{m}_n \neq m)(1 - \mathbb{P}(\widehat{m}_n = m))$.
Comme $\mathbb{P}(\widehat{m}_n = m) \underset{n\to+\infty}{\longrightarrow} 1$, on en déduit que $\mathbb{P}(\widehat{p}_n \leq x) \underset{n\to+\infty}{\longrightarrow} F_{\delta_p}(x)$: donc $\widehat{p}_n \overset{\mathcal{P}}{\underset{n\to+\infty}{\longrightarrow}} p$.

(d) Le modèle est alors $\big(\mathbf{N}^n, \mathcal{P}(\mathbf{N}^n), \mathbb{P}_\theta^{\otimes n}\big)_\theta$ car les variables X_j sont indépendantes.
Il est dominé par la mesure de comptage sur $\mathcal{P}(\mathbf{N}^n)$.
On a $\mathbb{P}(X_j = k \cap T_j = m) = \mathbb{P}(X_j = k \mid T_j = m)\,\mathbb{P}(T_j = m) = \frac{m!}{k!(m-k)!}p^k(1-p)^{m-k} \times e^{-\lambda}\frac{\lambda^m}{m!}p^k(1-p)^{m-k} = e^{-\lambda}\frac{\lambda^m}{k!(m-k)!}p^k(1-p)^{m-k}$.
On peut écrire par la formule des probabilités totales que $\mathbb{P}(X_j = k) = \sum_{m=k}^\infty \mathbb{P}(X_j = k \cap T_j = m) = e^{-\lambda}\sum_{m=k}^\infty \frac{\lambda^m}{k!(m-k)!}p^k(1-p)^{m-k}$. With $m' = m-k$ and therefore $m = m'+k$, we finally obtain $\mathbb{P}(X_j = k) = e^{-\lambda}\sum_{m'=0}^\infty \frac{\lambda^{m'+k}}{k!m'!}p^k(1-p)^{m'} = e^{-\lambda}\frac{(\lambda p)^k}{k!}\sum_{m'=0}^\infty \frac{((1-p)\lambda)^{m'}}{m'!} = e^{-\lambda}\frac{(\lambda p)^k}{k!}e^{(1-p)\lambda} = e^{-p\lambda}\frac{(\lambda p)^k}{k!}$. C'est bien une loi de Poisson de paramètre λp.
Du fait de l'indépendance des variables, la vraisemblance est $L_\theta(X_1,\ldots,X_n) = \prod_{j=1}^n e^{-p\lambda}\frac{(\lambda p)^{X_j}}{X_j!} = \frac{e^{-np\lambda}}{\prod_{j=1}^n X_j!}(\lambda p)^{\sum_{j=1}^n X_j}$.
On en déduit que $L_\theta(X_1,\ldots,X_n) = \exp\Big\{-np\lambda - \sum_{j=1}^n \log(X_j!) + \log(\lambda p)\sum_{j=1}^n X_j\Big\}$. Le modèle appartient à la famille exponentielle, avec $\alpha_1(\theta) = \log(\lambda p)$, $a_1(X_1,\ldots,X_n) = \sum_{j=1}^n X_j$ et $\beta(\theta) = -np\lambda$.
Comme $\alpha_1(\Theta) =]-\infty, 0[$ est d'intérieur non vide, alors $\sum_{j=1}^n X_j$ est une statistique exhaustive complète du modèle.
Le modèle repose sur le produit des paramètres λp, donc il est sur-paramétré et on pourrait remplacer λp par $\lambda')\lambda p$. Il n'est donc pas possible d'estimer de manière différenciée λ ou p.

$\square$