INP ENSEEIHT TOULOUSE

Examen Statistique
Lundi 5 décembre 2011

Exercice 1

Enoncé
On considère n variables aléatoires $X_1, ..., X_n$ indépendantes suivant la même loi de densité

$$f(x;\theta) = \theta x^{\theta-1} I_{[0,1]}(x)$$

avec $\theta > 0$ et où $I_{[0,1]}(x)$ est la fonction indicatrice sur $[0,1]$ ($I_{[0,1]}(x) = 1$ si $x \in [0,1]$ et $I_{[0,1]}(x) = 0$ si $x \notin [0,1]$).
1) Montrer que la vraisemblance de $(x_1, ..., x_n)$ admet un unique maximum global pour une valeur de θ que l'on déterminera. En déduire l'estimateur du maximum de vraisemblance du paramètre θ noté $\widehat{\theta}_{\text{MV}}$.
2) Déterminer la loi de $Y_i = -\ln X_i$. En déduire la fonction caractéristique de Y_i et montrer que la variable aléatoire

$$Z = \sum_{i=1}^{n} Y_i = -\sum_{i=1}^{n} \ln X_i$$

suit une loi gamma de paramètres θ et n, ce que l'on notera $Z \sim \Gamma(\theta, n)$.
3) Déterminer l'estimateur du maximum du paramètre $a = \frac{1}{\theta}$ construit à partir des variables aléatoires X_i noté $\widehat{a}_{\text{MV}}$ (on pourra utiliser la propriété d'invariance fonctionnelle). Déterminer la moyenne et la variance de $\widehat{a}_{\text{MV}}$. En déduire que l'estimateur $\widehat{a}_{\text{MV}}$ est un estimateur sans biais et convergent du paramètre a.
4) Déterminer la borne de Cramer-Rao pour un estimateur non biaisé du paramètre a. L'estimateur $\widehat{a}_{\text{MV}}$ est-il l'estimateur efficace du paramètre a ?
5) Déterminer $E[X_i]$ en fonction de θ. En déduire un estimateur de θ noté $\widehat{\theta}_{\text{M}}$ en utilisant la méthode des moments.
6) On suppose désormais qu'on dispose d'une information a priori sur le paramètre θ résumée dans la loi gamma $\Gamma(\alpha, \beta)$ de densité

$$f(\theta|\,\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta)\, I_{\mathbb{R}^+}(\theta)$$

- Montrer que la loi a posteriori de $\theta|\, x_1, ..., x_n$ est aussi une loi inverse-gamma dont on précisera les paramètres.
- Déterminer l'estimateur du maximum a posteriori du paramètre θ noté $\widehat{\theta}_{\text{MAP}}$.
- Expliquer le comportement de $\widehat{\theta}_{\text{MAP}}$ lorsque $n \to \infty$.

Réponses

1) La vraisemblance de $(x_1, ..., x_n)$ est définie par

$$\begin{aligned} L(x_1, ..., x_n; \theta) &= \prod_{i=1}^{n} f(x_i; \theta), \\ &= \prod_{i=1}^{n} \left[\theta x_i^{\theta-1} I_{[0,1]}(x_i)\right], \\ &= \theta^n \prod_{i=1}^{n} x_i^{\theta-1} \prod_{i=1}^{n} I_{[0,1]}(x_i). \end{aligned}$$

On a alors

$$\ln L(x_1, ..., x_n; \theta) = n \ln \theta + \sum_{i=1}^{n} (\theta - 1) \ln x_i.$$

En étudiant le signe de $\frac{\partial \ln L(x_1,...,x_n;\theta)}{\partial \theta}$, on obtient :

$$\begin{aligned} \frac{\partial \ln L(x_1, ..., x_n; \theta)}{\partial \theta} &\geq 0 \Longleftrightarrow \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i \geq 0, \\ &\Longleftrightarrow \theta \leq \frac{n}{-\sum_{i=1}^{n} \ln x_i}. \end{aligned}$$

On remarquera que puisque x_i appartient à l'intervalle $]0, 1[$, on a $-\ln(x_i) \geq 0$. La vraisemblance possède donc un maximum global unique obtenu pour

$$\theta = \frac{n}{-\sum_{i=1}^{n} \ln x_i}$$

d'où

$$\widehat{\theta}_{\mathrm{MV}} = \frac{n}{-\sum_{i=1}^{n} \ln X_i}.$$

2) On effectue le changement de variables

$$Y_i = -\ln X_i \Leftrightarrow X_i = \exp(-Y_i)$$

qui est bijectif de $]0, 1[$ dans $]0, \infty[$. La densité de Y_i s'écrit

$$\begin{aligned} g(y_i) &= \theta \left[\exp(-y_i)\right]^{\theta-1} \left|\exp(-y_i)\right| I_{]0,\infty[}(y_i) \\ &= \theta \exp(-\theta y_i) I_{]0,\infty[}(y_i). \end{aligned}$$

On reconnaît une loi gamma $\Gamma(\theta, 1)$ dont la fonction caractéristique est (voir tables)

$$\varphi_{Y_i}(t) = \frac{1}{1 - i\frac{t}{\theta}}$$

On en déduit la fonction caractéristique de Z (puisque les variables aléatoires Y_i sont indépendantes)

$$\varphi_Z(t) = \frac{1}{\left(1 - i\frac{t}{\theta}\right)^n}$$

qui est la fonction caractéristique d'une loi gamma $\Gamma(\theta, n)$, d'où

$$Z \sim \Gamma(\theta, n).$$

3) En utilisant la propriété d'invariance fonctionnelle, on a directement

$$\begin{aligned}
\widehat{a}_{\mathrm{MV}} &= \frac{1}{\widehat{\theta}_{\mathrm{MV}}} \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i \\
&= -\frac{1}{n}\sum_{i=1}^{n} \ln X_i \\
&= \frac{Z}{n}
\end{aligned}$$

La moyenne de l'estimateur $\widehat{a}_{\mathrm{MV}}$ est

$$E\left[\widehat{a}_{\mathrm{MV}}\right] = \frac{1}{n}E\left[Z\right] = \frac{n}{n\theta} = \frac{1}{\theta}.$$

On en déduit que $\widehat{a}_{\mathrm{MV}}$ est donc un estimateur non biaisé de $a = \frac{1}{\theta}$. La variance de l'estimateur $\widehat{a}_{\mathrm{MV}}$ est

$$\begin{aligned}
\mathrm{var}\left[\widehat{a}_{\mathrm{MV}}\right] &= \frac{1}{n^2}\mathrm{var}\left[Z\right] \\
&= \frac{1}{n^2}\frac{n}{\theta^2} \\
&= \frac{1}{n\theta^2} = \frac{a^2}{n}
\end{aligned}$$

L'estimateur $\widehat{a}_{\mathrm{MV}}$ est un estimateur non biaisé de $a = \frac{1}{\theta}$ tel que

$$\lim_{n\to\infty} \mathrm{var}\left[\widehat{a}_{\mathrm{MV}}\right] = 0$$

donc l'estimateur $\widehat{a}_{\mathrm{MV}}$ est convergent.
4) La borne de Cramer-Rao pour un estimateur non biaisé du paramètre $a = \frac{1}{\theta}$ est définie par

$$\mathrm{BCR}\left(a\right) = \frac{-1}{E\left[\frac{\partial \ln L(X_1,\ldots,X_n;a)}{\partial a^2}\right]}$$

Mais

$$\begin{aligned}
\ln L(x_1, ..., x_n; \theta) &= n\ln\theta + \sum_{i=1}^{n}(\theta - 1)\ln x_i \\
&= -n\ln a + + \sum_{i=1}^{n}\left(\frac{1}{a} - 1\right)\ln x_i
\end{aligned}$$

donc

$$\begin{aligned}
\mathrm{BCR}(a) &= -E\left[\frac{\partial}{\partial a}\left[\frac{-n}{a}-\frac{1}{a^2}\sum_{i=1}^{n}\ln X_i\right]\right]^{-1} \\
&= -E\left[\frac{n}{a^2}+\frac{2}{a^3}\sum_{i=1}^{n}\ln X_i\right]^{-1} \\
&= -\left(\frac{n}{a^2}+\frac{2}{a^3}\left[-nE\left(Y_i\right)\right]\right)^{-1} \\
&= -\left(\frac{n}{a^2}-\frac{2na}{a^3}\right)^{-1} \\
&= \frac{a^2}{n}.
\end{aligned}$$

Puisque $\widehat{a}_{\mathrm{MV}}$ est un estimateur non biaisé de a et que $\mathrm{var}[\widehat{a}_{\mathrm{MV}}] = \mathrm{BCR}(a)$, l'estimateur $\widehat{a}_{\mathrm{MV}}$ est l'estimateur efficace de a.
5) La moyenne de X_i est

$$E\left[X_i\right] = \int_0^1 \theta x^{\theta} dx = \frac{\theta}{\theta+1}\left[x^{\theta+1}\right]_{x=0}^{x=1} = \frac{\theta}{\theta+1}$$

d'où

$$\theta = \frac{E\left[X_i\right]}{1-E\left[X_i\right]}$$

On en déduit l'estimateur des moments noté $\widehat{\theta}_{\mathrm{M}}$ défini par

$$\widehat{\theta}_{\mathrm{M}} = \frac{\overline{X}}{1-\overline{X}} \text{ avec } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

5) On suppose désormais qu'on dispose d'une information a priori sur le paramètre θ résumée dans la loi gamma $\Gamma(\alpha,\beta)$ de densité

$$f\left(\theta|\,\alpha,\beta\right) = \frac{\alpha^{\beta}}{\Gamma\left(\beta\right)}\theta^{\beta-1}\exp\left(-\alpha\theta\right) I_{\mathbb{R}^+}\left(\theta\right)$$

- La loi a posteriori de $\theta|\,x_1,...,x_n$ s'écrit

$$\begin{aligned}
f\left(\theta|\,x_1,...,x_n\right) &\propto f\left(x_1,...,x_n|\,\theta\right) f(\theta) \\
&\propto \theta^n\left(\prod_{i=1}^{n} x_i^{\theta-1}\right)\theta^{\beta-1}\exp\left(-\alpha\theta\right) I_{\mathbb{R}^+}\left(\theta\right) \\
&\propto \theta^{n+\beta-1}\exp\left\{-\theta\left[\alpha-\sum_{i=1}^{n}\ln X_i\right]\right\} I_{\mathbb{R}^+}\left(\theta\right) \\
&\propto \theta^{n+\beta-1}\exp\left\{-\theta\left[\alpha+\sum_{i=1}^{n} Y_i\right]\right\} I_{\mathbb{R}^+}\left(\theta\right)
\end{aligned}$$

Cette densité est la densité d'une loi gamma

$$\theta|\, x_1, ..., x_n \sim \Gamma\left(\alpha + \sum_{i=1}^{n} Y_i, n + \beta\right)$$

- L'estimateur du maximum a posteriori du paramètre θ noté $\widehat{\theta}_{\text{MAP}}$ est obtenu en maximisant le logarithme de la loi a posteriori $f(\theta|\, x_1, ..., x_n)$. Mais

$$\ln\left[f(\theta|\, x_1, ..., x_n)\right] = C + (n + \beta - 1)\ln\theta - \theta\left(\alpha + \sum_{i=1}^{n} Y_i\right)$$

d'où

$$\begin{aligned}\frac{\partial \ln\left[f(\theta|\, x_1, ..., x_n)\right]}{\partial\theta} \quad &\geq \quad 0 \Longleftrightarrow \frac{n + \beta - 1}{\theta} - \left(\alpha + \sum_{i=1}^{n} Y_i\right) \geq 0 \\ &\Longleftrightarrow \quad \theta \leq \frac{n + \beta - 1}{\alpha + \sum_{i=1}^{n} Y_i}\end{aligned}$$

La loi a posteriori $f(\theta|\, x_1, ..., x_n)$ possède donc un unique maximum global, d'où

$$\widehat{\theta}_{\text{MAP}} = \frac{n + \beta - 1}{\alpha + \sum_{i=1}^{n} Y_i}$$

- On peut exprimer $\widehat{\theta}_{\text{MAP}}$ en fonction de $\widehat{\theta}_{\text{MV}}$ puisque

$$\begin{aligned}\widehat{\theta}_{\text{MAP}} &= \frac{1 + \frac{\beta - 1}{n}}{\frac{\alpha}{n} + \frac{1}{n}\sum_{i=1}^{n} Y_i} \\ &= \frac{1 + \frac{\beta - 1}{n}}{\frac{\alpha}{n} + \frac{1}{\widehat{\theta}_{\text{MV}}}}\end{aligned}$$

L'estimateur $\widehat{\theta}_{\text{MAP}}$ se comporte donc comme $\widehat{\theta}_{\text{MV}}$ lorsque $n \to \infty$. Lorsqu'on a beaucoup d'observations, on fait confiance à ces observations et donc l'effet de la loi a priori est négligeable.

Exercice 2

Enoncé

Comme dans l'exercice précédent, on considère n variables aléatoires $X_1, ..., X_n$ indépendantes suivant la même loi de densité

$$f(x;\theta) = \theta x^{\theta-1} I_{[0,1]}(x)$$

avec $\theta > 0$ et on désire effectuer le test d'hypothèses

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

avec $\theta_1 > \theta_0$.

1) Déterminer la statistique du test de Neyman-Pearson notée T_n (qui ne dépend que des variables aléatoires $X_1, ..., X_n$) et donner la région critique de ce test. Ce résultat est-il cohérent avec le fait que la moyenne de $Y_i = -\ln X_i$ est

$$E[Y_i] = \frac{1}{\theta}$$

2) On rappelle que

$$\sum_{i=1}^{n} Y_i = -\sum_{i=1}^{n} \ln X_i$$

suit une loi gamma de paramètres θ et n.

- Déterminer la densité de la variable aléatoire $U = \frac{1}{\theta T_n}$ et montrer que cette densité dépend de n mais est indépendante de θ. On notera $H_n(u)$ la fonction de répartition associée à la variable aléatoire U_n.
- Exprimer les risques de première et de seconde espèce α et β en fonction du seuil du test de Neyman Pearson noté S_α, des paramètres θ_0 et θ_1 et de la fonction H_n.
- En déduire l'expression analytique des courbes COR de ce détecteur. Représenter la forme approximative de ces courbes COR pour différentes valeurs des paramètres θ_0 et θ_1.

3) Avant de tester la valeur du paramètre θ, on se propose de vérifier si les données x_i suivant la loi de densité $f(x;\theta)$ (donnée ci-dessus) à l'aide du test de Kolmogorov.

- Déterminer la fonction de répartition associée à la loi de densité $f(x;\theta)$ et représenter la graphiquement.
- Donner le principe du test de Kolmogorov en précisant a) la statistique de test (fonction des observations x_i), b) la règle de décision, c) comment calculer le seuil du test à l'aide de la loi (asymptotique) de Kolmogorov.

Réponses

1) Le test de Neyman-Pearson est défini par

$$\text{Rejet de } H_0 \text{ si} \quad \frac{L(x_1,...,x_n;\theta_1)}{L(x_1,...,x_n;\theta_0)} > k_\alpha.$$

Mais

$$\begin{aligned}\frac{L(x_1,...,x_n;\theta_1)}{L(x_1,...,x_n;\theta_0)} &> k_\alpha \Leftrightarrow \frac{\theta_1^n \prod_{i=1}^n x_i^{\theta_1-1} \prod_{i=1}^n I_{[0,1]}(x_i)}{\theta_0^n \prod_{i=1}^n x_i^{\theta_0-1} \prod_{i=1}^n I_{[0,1]}(x_i)} > k_\alpha, \\ &\Leftrightarrow (\theta_1 - 1 - \theta_0 + 1) \sum_{i=1}^n \ln x_i > k_\alpha \\ &\Leftrightarrow (\theta_0 - \theta_1) \sum_{i=1}^n y_i > k_\alpha\end{aligned}$$

Puisque $\theta_1 > \theta_0$, on en déduit le test équivalent

$$\text{Rejet de } H_0 \text{ si} \quad \sum_{i=1}^n y_i < s_\alpha.$$

La statistique du test de Neyman-Pearson est donc

$$T_n = \sum_{i=1}^n Y_i.$$

La région critique de ce test est définie par

$$\left\{ y \in \mathbb{R}^n \left| \sum_{i=1}^n y_i < s_\alpha \right. \right\}.$$

Puisque $E[Y_i] = \frac{1}{\theta}$, la moyenne de Y_i sous l'hypothèse H_0 est plus grande que la moyenne de Y_i sous l'hypothèse H_1. C'est ce que dit le test de Neyman-Pearson : "lorsque la moyenne $\frac{1}{n}\sum_{i=1}^n y_i$ est petite (i.e., inférieure à s_α), on accepte l'hypothèse H_1.

2) Comme rappelé dans l'énoncé

$$T_n = \sum_{i=1}^n Y_i \sim \Gamma(\theta, n)$$

- La loi de $U_n = \frac{1}{\theta T_n}$ s'obtient à l'aide d'un simple changement de variables. Le Jacobien de la transformation est donc

$$J = \frac{dt_n}{du_n} = \frac{-1}{\theta u_n^2}$$

Comme la densité d'une loi gamma $\Gamma(\theta, n)$ s'écrit

$$g_n(t) = \frac{\theta^n}{\Gamma(n)} \exp(-\theta t)\, t^{n-1} I_{\mathbb{R}^+}(t)$$

et que le changement de variables $U_n = \frac{1}{\theta T_n}$ est bijectif de $\mathbb{R}^+$ dans $\mathbb{R}^+$, la densité de U_n est définie par

$$\begin{aligned} h_n(u) &= \frac{\theta^n}{\Gamma(n)} \exp\left(-\frac{1}{u}\right) \frac{1}{\theta^{n-1}u^{n-1}} \left|\frac{-1}{\theta u^2}\right| I_{\mathbb{R}^+}(u) \\ &= \frac{1}{\Gamma(n)} \exp\left(-\frac{1}{u}\right) \frac{1}{u^{n+1}} I_{\mathbb{R}^+}(u) \end{aligned}$$

qui est bien une densité indépendante de θ.

- Le risque de première espèce α est défini par

$$\begin{aligned} \alpha &= P[\text{Rejeter } H_0 \mid H_0 \text{ vraie}] \\ &= P\left[\sum_{i=1}^{n} Y_i < s_\alpha \middle| \theta = \theta_0\right] \\ &= P[T_n < s_\alpha | \, \theta = \theta_0] \\ &= P\left[\frac{1}{\theta U_n} < s_\alpha \middle| \theta = \theta_0\right] \\ &= P\left[U_n > \frac{1}{\theta_0 s_\alpha} \middle| U_n \text{ suit la loi indépendante de } \theta \text{ définie ci-dessus}\right] \\ &= 1 - H_n\left(\frac{1}{\theta_0 s_\alpha}\right). \end{aligned}$$

De la même façon

$$\begin{aligned} \beta &= P[\text{Rejeter } H_1 \mid H_1 \text{ vraie}] \\ &= P\left[\sum_{i=1}^{n} Y_i \geq s_\alpha \middle| \theta = \theta_1\right] \\ &= P\left[\frac{1}{\theta U_n} \geq s_\alpha \middle| \theta = \theta_1\right] \\ &= H_n\left(\frac{1}{\theta_1 s_\alpha}\right). \end{aligned}$$

- Les courbes caractéristiques opérationnelles du récepteur (courbes COR) expriment la puissance du test $\pi = 1 - \beta$ en fonction de α. Dans le cas présent, on a

$$\begin{aligned} \pi &= 1 - H_n\left(\frac{1}{\theta_1 s_\alpha}\right), \\ &= 1 - H_n\left[\frac{\theta_0}{\theta_1} H_n^{-1}(1-\alpha)\right]. \end{aligned}$$

On voit donc que la performance du test dépend du rapport des paramètres $\frac{\theta_0}{\theta_1}$ (la puissance dépend aussi de α bien entendu). Si on fixe θ_0 et le risque α, plus θ_1 est grand, plus $\frac{\theta_0}{\theta_1}$ est petit et donc plus la puissance du test est grande.

3) Test de Kolmogorov

- La fonction de répartition associée à la densité

$$f(x;\theta) = \theta x^{\theta-1} I_{[0,1]}(x)$$

est

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u;\theta)\, du \\ &= \begin{cases} 0 \text{ si } x < 0 \\ \int_0^x \theta u^{\theta-1} du = x^\theta \text{ si } x \in [0,1] \\ 1 \text{ si } x > 1 \end{cases} \end{aligned}$$

- La statistique du test de Kolmogorov est

$$\begin{aligned} K_n &= \sup_{x\in\mathbb{R}} \left| F(x) - \widehat{F}_n(x) \right| \\ &= \sup_{i=1,\ldots,n} \left\{ \max\left(E_i^-, E_i^+\right) \right\} \end{aligned}$$

avec

$$\begin{aligned} E_i^- &= F(x_i) - \frac{i-1}{n} \\ E_i^+ &= F(x_i) - \frac{i}{n} \end{aligned}$$

La règle de décision est

$$\text{Rejet de l'hypothèse } H_0 \text{ si } K_n > k_\alpha$$

où s_α est un seuil dépendant du risque de première espèce α. En effet

$$\begin{aligned} \alpha &= P\left[\text{Rejeter } H_0 \mid H_0 \text{ vraie}\right] \\ &= P\left[K_n > k_\alpha \mid H_0 \text{ vraie}\right] \\ &\simeq 1 - G_n(k_\alpha) \end{aligned}$$

où G_n est la fonction de répartition de la loi de Kolmogorov. On en déduit

$$k_\alpha = G_n^{-1}(1-\alpha).$$