Correction de l'examen de Statistique du
Mardi 7 décembre 2010

Exercice 1 : Estimation

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

$$f\left(x;\sigma^2\right) = \frac{x}{\sigma^2}\exp\left[-\frac{x^2}{2\sigma^2}\right] I_{\mathbb{R}^+}(x)$$

avec $\sigma > 0$ et où $I_{\mathbb{R}^+}(x)$ est la fonction indicatrice sur $\mathbb{R}^+$ ($I_{\mathbb{R}^+}(x) = 1$ si $x \in \mathbb{R}^+$ et $I_{\mathbb{R}^+}(x) = 0$ si $x \notin \mathbb{R}^+$). Cette loi est appelée loi de Rayleigh et on utilisera la notation habituelle $X_k \sim \mathcal{R}(\sigma^2)$. On admettra que les premiers moments deles résultats suivants

$$E\left[X_k\right] = \sigma\sqrt{\frac{\pi}{2}}, E\left[X_k^2\right] = 2\sigma^2, E\left[X_k^3\right] = \left(3\sqrt{\frac{\pi}{2}}\right)\sigma^3 \text{ et } E\left[X_k^4\right] = 8\sigma^4.$$

Estimateur du Maximum de Vraisemblance

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

$$\begin{aligned} L(x_1, ..., x_n; \theta) &= \prod_{i=1}^n f\left(x_i;\theta\right), \\ &= \prod_{i=1}^n \frac{x_i}{\theta}\exp\left[-\frac{x^2}{2\theta}\right] I_{\mathbb{R}^+}(x_i), \\ &= \frac{\prod_{i=1}^n x_i}{\theta^n}\exp\left[-\frac{1}{2\theta}\sum_{i=1}^n x_i^2\right]\prod_{i=1}^n I_{\mathbb{R}^+}(x_i). \end{aligned}$$

On a alors

$$\ln L(x_1, ..., x_n; \theta) = -n\ln\theta + \sum_{i=1}^n \ln(x_i) - \frac{1}{2\theta}\sum_{i=1}^n x_i^2.$$

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

$$\begin{aligned} \frac{\partial \ln L(x_1, ..., x_n; \theta)}{\partial\theta} &\geq 0 \Longleftrightarrow \frac{-n}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^n x_i^2 \geq 0, \\ &\Longleftrightarrow \theta \leq \frac{1}{2n}\sum_{i=1}^n x_i^2. \end{aligned}$$

La vraisemblance possède donc un maximum global unique obtenu pour $\theta = \frac{1}{2n}\sum_{i=1}^n x_i^2$ d'où

$$\widehat{\theta}_{\text{MV}} = \frac{1}{2n}\sum_{i=1}^n X_i^2.$$

2) La moyenne de l'estimateur $\widehat{\theta}_{\mathrm{MV}}$ est

$$\begin{aligned} E\left[\widehat{\theta}_{\mathrm{MV}}\right] &= \frac{1}{2n}\sum_{i=1}^{n} E\left[X_i^2\right] \\ &= \frac{1}{2n}\sum_{i=1}^{n}\left(2\sigma^2\right) \\ &= \sigma^2 = \theta \end{aligned}$$

L'estimateur $\widehat{\theta}_{\mathrm{MV}}$ est donc un estimateur non biaisé de θ. La variance de l'estimateur $\widehat{\theta}_{\mathrm{MV}}$ est

$$\begin{aligned} \operatorname{var}\left[\widehat{\theta}_{\mathrm{MV}}\right] &= \frac{1}{4n^2}\operatorname{var}\left[\sum_{i=1}^{n} X_i^2\right], \\ &= \frac{1}{4n}\operatorname{var}\left(X_1^2\right), \\ &= \frac{1}{4n}\left(E\left[X_1^4\right] - E\left[X_1^2\right]^2\right), \\ &= \frac{1}{4n}\left(8\sigma^4 - 4\sigma^4\right) \\ &= \frac{\sigma^4}{n} \end{aligned}$$

L'estimateur $\widehat{\theta}_{\mathrm{MV}}$ est un estimateur non biaisé de θ tel que

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

donc l'estimateur $\widehat{\theta}_{\mathrm{MV}}$ est convergent.
3) La borne de Cramer-Rao pour les estimateurs non biaisés du paramètre θ est définie par

$$\begin{aligned} \mathrm{BCR}(\theta) &= \frac{-1}{E\left[\frac{\partial \ln L(X_1,\ldots,X_n;\theta)}{\partial\theta^2}\right]}, \\ &= -E\left[\frac{\partial}{\partial\theta}\left[\frac{-n}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} X_i^2\right]\right]^{-1}, \\ &= -E\left[\frac{n}{\theta^2} - \frac{2}{2\theta^3}\sum_{i=1}^{n} X_i^2\right]^{-1} \\ &= \left[-\frac{n}{\theta^2} + \frac{1}{\theta^3}\left(2n\theta\right)\right]^{-1} = \frac{\theta^2}{n} \end{aligned}$$

Puisque $\widehat{\theta}_{\mathrm{MV}}$ est un estimateur non biaisé de θ et que $\operatorname{var}\left[\widehat{\theta}_{\mathrm{MV}}\right]$ =BCR(θ), l'estimateur $\widehat{\theta}_{\mathrm{MV}}$ est l'estimateur efficace de θ.

Estimation Bayésienne

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

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

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

$$\begin{aligned} f(\theta | x_1, ..., x_n) &\propto f(x_1, ..., x_n | \theta) f(\theta) \\ &\propto \frac{1}{\theta^n} \exp\left[-\frac{1}{2\theta} \sum_{i=1}^n x_i^2\right] \frac{1}{\theta^{\alpha+1}} \exp\left(-\frac{\beta}{\theta}\right) I_{\mathbb{R}^+}(\theta) \\ &\propto \frac{1}{\theta^{n+\alpha+1}} \exp\left\{-\frac{1}{\theta}\left[\frac{1}{2}\sum_{i=1}^n x_i^2 + \beta\right]\right\} I_{\mathbb{R}^+}(\lambda) \end{aligned}$$

Cette densité est la densité d'une loi inverse-gamma

$$\theta | x_1, ..., x_n \sim IG\left(n + \alpha, \frac{1}{2}\sum_{i=1}^n x_i^2 + \beta\right)$$

2) 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

$$\ln[f(\theta | x_1, ..., x_n)] = C - (n + \alpha + 1)\ln\theta - \frac{1}{\theta}\left[\frac{1}{2}\sum_{i=1}^n x_i^2 + \beta\right]$$

d'où

$$\begin{aligned} \frac{\partial \ln[f(\theta | x_1, ..., x_n)]}{\partial \theta} &\geq 0 \Longleftrightarrow -\frac{n+\alpha+1}{\theta} + \frac{1}{\theta^2}\left[\frac{1}{2}\sum_{i=1}^n x_i^2 + \beta\right] \geq 0 \\ &\Longleftrightarrow \theta \leq \frac{1}{n+\alpha+1}\left[\frac{1}{2}\sum_{i=1}^n x_i^2 + \beta\right] \end{aligned}$$

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

$$\widehat{\theta}_{\text{MAP}} = \frac{1}{n+\alpha+1}\left[\frac{1}{2}\sum_{i=1}^n X_i^2 + \beta\right]$$

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

$$\begin{aligned} \widehat{\theta}_{\text{MAP}} &= \frac{1}{1 + \frac{\alpha+1}{n}}\left[\frac{1}{2n}\sum_{i=1}^n X_i^2 + \frac{\beta}{n}\right] \\ &= \frac{1}{1 + \frac{\alpha+1}{n}}\widehat{\theta}_{\text{MV}} + \frac{\beta}{n+\alpha+1} \end{aligned}$$

L'estimateur $\widehat{\theta}_{\text{MAP}}$ se comporte donc $\widehat{\theta}_{\text{MV}}$ comme 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.

Méthode des Moments

1) La moyenne d'une loi de Rayleigh est

$$E[X] = \sigma\sqrt{\frac{\pi}{2}}$$

donc

$$\theta = \sigma^2 = \frac{2}{\pi}[E[X]]^2$$

L'estimateur des moments résultant de cette dernière égalité est

$$\widehat{\theta}_{\mathrm{M}} = \frac{2}{\pi}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]^2$$

2) La moyenne de l'estimateur $\widehat{\theta}_{\mathrm{M}}$ est

$$\begin{aligned}
E\left[\widehat{\theta}_{\mathrm{M}}\right] &= \frac{2}{\pi n^2}E\left[\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j\right] \\
&= \frac{2}{\pi n^2}\sum_{i,j} E[X_i X_j] \\
&= \frac{2}{\pi n^2}\left[nE\left[X_i^2\right] + \left(n^2 - n\right)E[X_i]^2\right] \\
&= \frac{2}{\pi n^2}\left[2n\theta + \left(n^2 - n\right)\frac{\pi}{2}\theta\right] \\
&= \frac{4 + \pi(n-1)}{\pi n}\theta
\end{aligned}$$

On en déduit un estimateur non-biaisé du paramètre θ

$$\widehat{\theta}^* = \frac{\pi n}{4 + \pi(n-1)}\widehat{\theta}_{\mathrm{M}} = \frac{2n}{4 + \pi(n-1)}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]^2$$

Exercice 2 : Test de Neyman-Pearson

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

$$\begin{aligned}
\frac{L(x_1, ..., x_n; \theta_1)}{L(x_1, ..., x_n; \theta_0)} &> S_\alpha \Leftrightarrow \frac{\frac{\prod_{i=1}^n x_i}{\theta_1^n} \exp\left[-\frac{1}{2\theta_1}\sum_{i=1}^n x_i^2\right]}{\frac{\prod_{i=1}^n x_i}{\theta_0^n} \exp\left[-\frac{1}{2\theta_0}\sum_{i=1}^n x_i^2\right]} > k_\alpha, \\
&\Leftrightarrow \left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)\sum_{i=1}^n x_i^2 > k'_\alpha
\end{aligned}$$

Pour $\theta_1 > \theta_0$, on a $\frac{1}{\theta_0} - \frac{1}{\theta_1} > 0$ et donc on en déduit le test équivalent

$$\text{Rejet de } H_0 \text{ si} \quad \sum_{i=1}^n X_i^2 > s_\alpha.$$

Pour $\theta_1 < \theta_0$, on a $\frac{1}{\theta_0} - \frac{1}{\theta_1} < 0$ et donc le test s'écrit

$$\text{Rejet de } H_0 \text{ si} \quad \sum_{i=1}^n X_i^2 < s_\alpha.$$

La statistique du test de Neyman-Pearson est donc

$$T_n = \sum_{i=1}^n X_i^2.$$

La région critique de ce test est définie par

- Si $\theta_1 > \theta_0$

$$\left\{x \in \mathbb{R}^n \,\middle|\, \sum_{i=1}^n x_i^2 > s_\alpha\right\}.$$

- Si $\theta_1 < \theta_0$

$$\left\{x \in \mathbb{R}^n \,\middle|\, \sum_{i=1}^n x_i^2 < s_\alpha\right\}.$$

2) Le changement de variables $Y_k = \frac{X_k^2}{\sigma^2}$ est bijectif de $\mathbb{R}^+$ dans $\mathbb{R}^+$. De plus

$$y_k = \frac{x_k^2}{\sigma^2} \Longleftrightarrow x_k = \sigma\sqrt{y_k}$$

Le Jacobien de la transformation est donc

$$J = \frac{dx_k}{dy_k} = \frac{\sigma}{2\sqrt{y_k}}$$

La densité de Y_k est donc

$$g(y_k) = \frac{\sigma\sqrt{y_k}}{\sigma^2} \exp\left(-\frac{y_k}{2}\right) \frac{\sigma}{2\sqrt{y_k}} I_{\mathbb{R}^+}(y_k)$$

soit

$$g(y_k) = \frac{1}{2} \exp\left(-\frac{y_k}{2}\right) I_{\mathbb{R}^+}(y_k)$$

On en déduit que $Y_k = \frac{X_k^2}{\sigma^2}$ suit une loi du χ_2^2 à deux degrés de liberté. La fonction caractéristique de $\frac{T_n}{\sigma^2}$ est

$$\begin{aligned}
\phi(u) &= E\left[e^{iu\frac{T_n}{\sigma^2}}\right] \\
&= E\left[e^{iu\sum_{k=1}^{n}\frac{X_k^2}{\sigma^2}}\right] \\
&= E\left[\prod_{k=1}^{n} e^{iuY_k}\right]
\end{aligned}$$

En utilisant le fait que les variables aléatoires Y_k sont indépendantes, on obtient

$$\begin{aligned}
\phi(u) &= \prod_{k=1}^{n} \phi_{Y_k}(u) \\
&= \prod_{k=1}^{n} \frac{1}{1-2iu} \\
&= \frac{1}{(1-2iu)^n}
\end{aligned}$$

On en déduit que $\frac{T_n}{\sigma^2}$ suit une loi du khi-deux à $2n$ degrés de liberté

$$\frac{T_n}{\sigma^2} \sim \chi_{2n}^2$$

3) Le risque de première espèce α est défini par

$$\begin{aligned}
\alpha &= P\left[\text{Rejeter } H_0 \mid H_0 \text{ vraie}\right], \\
&= P\left[\sum_{i=1}^{n} X_i^2 > s_\alpha \middle| \theta = \theta_0\right], \\
&= P\left[\frac{T_n}{\sigma_0^2} > \frac{s_\alpha}{\sigma_0^2} \middle| \sum_{i=1}^{n} \frac{X_i^2}{\sigma_0^2} \sim \chi_{2n}^2\right], \\
&= G_{2n}\left(\frac{s_\alpha}{\sigma_0^2}\right).
\end{aligned}$$

De la même façon

$$\begin{aligned}
\beta &= P\left[\text{Rejeter } H_1 \mid H_1 \text{ vraie}\right], \\
&= P\left[\sum_{i=1}^{n} X_i^2 \leq s_\alpha \middle| \theta = \theta_1\right], \\
&= 1 - P\left[\frac{T_n}{\sigma_1^2} > \frac{s_\alpha}{\sigma_1^2} \middle| \sum_{i=1}^{n} \frac{X_i^2}{\sigma_1^2} \sim \chi_{2n}^2\right], \\
&= 1 - G_{2n}\left(\frac{s_\alpha}{\sigma_1^2}\right).
\end{aligned}$$

4) 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

$$\begin{aligned}
\pi &= G_{2n}\left(\frac{s_\alpha}{\sigma_1^2}\right), \\
&= G_{2n}\left[\frac{\sigma_0^2}{\sigma_1^2} G_{2n}^{-1}(\alpha)\right].
\end{aligned}$$

On voit donc que la performance du test dépend des variances σ_0^2 et σ_1^2 uniquement via la quantité $\frac{\sigma_0^2}{\sigma_1^2}$ (la puissance dépend aussi de α bien entendu). Si on fixe σ_0^2 et le risque α, plus σ_1^2 est grand, plus $\frac{\sigma_0^2}{\sigma_1^2}$ est petit et donc plus la puissance du test est grande.

Exercice 3 : **Test d'ajustement**

1) Nous allons effectuer un test du χ^2 construit à partir des 6 classes suivantes

$$C_1 = \{1\}, C_2 = \{2\}, C_3 = \{3\}, C_4 = \{4\}, C_5 = \{5\}, C_6 = \{6\}$$

Les deux hypothèses sont définies par

$$\begin{aligned} H_0 &: \text{ le dé est non truqué} \\ H_1 &: \text{ le dé est truqué} \end{aligned}$$

L'hypothèse H_0 est caractérisée par

$$P(C_i) = p_i = \frac{1}{6}, \qquad \forall i = 1, .., 6$$

tandis que pour l'hypothèse H, au moins une des probabilités p_i est différente de $\frac{1}{6}$. La statistique du test du χ^2 est

$$\begin{aligned} \phi &= \sum_{i=1}^{6} \frac{(N_i - np_i)^2}{np_i} \\ &= \frac{6}{n} \sum_{i=1}^{6} \left(N_i - \frac{n}{6}\right)^2 \\ &= \frac{1}{20} [4 + 4 + 4 + 1 + 0 + 9] \\ &= \frac{22}{20} = 1.1 \end{aligned}$$

La régle de décision du test du χ^2 est

$$\text{Rejet de l'hypothèse } H_0 \text{ si } \phi > s_\alpha,$$

avec

$$\begin{aligned} \alpha &= P\left[\text{Rejeter } H_0 \mid H_0 \text{ vraie}\right], \\ &= P\left[\phi > s_\alpha \mid H_0 \text{ vraie}\right]. \\ &= P\left[\phi > s_\alpha \mid \phi \sim \chi_5^2\right]. \end{aligned}$$

On sait que ϕ est distribuée suivant une loi du χ^2_{K-1} sous l'hypothèse H_0, où $K = 6$ est le nombre de classes. Si $g_n(u)$ est la densité d'une loi du χ^2_n, on pose

$$\int_x^\infty g_n(u) du = G_n(x),$$

et on obtient

$$\alpha = G_3\left(s_\alpha\right) \Longrightarrow s_\alpha = G_5^{-1}\left(\alpha\right).$$

Pour $\alpha = 0.05$, les tables de la loi du χ_5^2 donnent

$$s_\alpha = 11.071$$

On observe que

$$\phi < s_{0.05}$$

et donc on accepte l'hypothèse que le dé est parfait avec le risque $a = 0.05$.
2) L'application du théorème de la limite centrale donne

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - E[X_i]}{\sqrt{\operatorname{var}[X_i]/n}} \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0,1)$$

Pour n grand, on peut donc approcher la loi de $\frac{1}{n}\sum_{i=1}^n X_i$ par la loi normale

$$\mathcal{N}\left(E[X_i], \frac{\operatorname{var}[X_i]}{n}\right)$$

Pour une loi uniforme sur l'ensemble $\{1, ..., 6\}$ (ce qui correspond à l'hypothèse H_0), on a

$$\begin{aligned} E[X_i] &= \frac{1 + ... + 6}{6} = \frac{7}{2} \\ \operatorname{var}[X_i] &= E[X_i^2] - E[X_i]^2 = \frac{1^2 + ... + 6^2}{6} - \frac{49}{4} = \frac{35}{21} = \nu^2 \end{aligned}$$

donc

$$U_n = \frac{\frac{1}{n}\sum_{i=1}^n X_i - \frac{7}{2}}{\nu/\sqrt{n}} \simeq \mathcal{N}(0,1)$$

où $\simeq$ signifie asymptotiquement distribué. On en déduit

$$\begin{aligned} \alpha &= P\left[\text{Rejeter } H_0 \mid H_0 \text{ vraie}\right], \\ &= P\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{7}{2}\right)^2 > s_\alpha \mid H_0 \text{ vraie}\right], \\ &= P\left[\frac{\nu^2}{n}U_n^2 > s_\alpha \mid H_0 \text{ vraie}\right], \\ &= P\left[U_n^2 > \frac{n}{\nu^2}s_\alpha \mid U_n^2 \sim \chi_1^2\right], \\ &= G_1\left(\frac{n}{\nu^2}s_\alpha\right) \end{aligned}$$

On en déduit

$$s_\alpha = \frac{\nu^2}{n}G_1^{-1}(\alpha).$$