
CORRECTION EXAMEN STATISTIQUE - 1TR

Mercredi 1 Décembre 2014 (8h-9h45)

Partiel sans document (Une feuille A4 recto-verso autorisée)

Exercice 1 : Estimation

On considère n variables aléatoires X_1, \dots, X_n indépendantes suivant la même loi continue de densité

$$p(x; \theta) = \begin{cases} \frac{1}{2} - \theta & \text{si } -1 \leq x < 0 \\ \theta + \frac{1}{2} & \text{si } 0 \leq x < 1 \\ 0 & \text{sinon} \end{cases}$$

avec $\theta \in]-\frac{1}{2}, \frac{1}{2}[$.

1. Montrer que la vraisemblance des observations x_1, \dots, x_n s'écrit

$$p(x_1, \dots, x_n; \theta) = \left(\frac{1}{2} - \theta\right)^{n-y_n} \left(\theta + \frac{1}{2}\right)^{y_n}$$

où y_n est le nombre d'observations x_i vérifiant $x_i \geq 0$. Montrer que cette vraisemblance admet un unique maximum global pour une valeur de θ que l'on déterminera. On définit la variable aléatoire associée Y_n comme suit

$$Y_n = \text{card}\{i \in \{1, \dots, n\} | X_i \geq 0\}.$$

où $\text{card}\{A\}$ est le nombre d'éléments de l'ensemble A . Exprimer l'estimateur du maximum de vraisemblance de θ noté $\hat{\theta}_n$ en fonction de n et Y_n .

Réponse : puisque les variables aléatoires X_i sont indépendantes, la vraisemblance des observations x_1, \dots, x_n est

$$p(x_1, \dots, x_n; \theta) = \prod_{i=1}^n \left[\left(\frac{1}{2} - \theta\right)^{\mathcal{I}_{[-1,0]}(x_i)} \left(\theta + \frac{1}{2}\right)^{\mathcal{I}_{[0,1]}(x_i)} \right] = \left(\frac{1}{2} - \theta\right)^{n-y_n} \left(\theta + \frac{1}{2}\right)^{y_n}.$$

Le sens de variation de la vraisemblance en fonction des valeurs de θ s'obtient en étudiant l'inégalité

$$\frac{\partial \ln p(x_1, \dots, x_n; \theta)}{\partial \theta} \geq 0 \Leftrightarrow \frac{y_n}{\theta + \frac{1}{2}} \geq \frac{n - y_n}{\frac{1}{2} - \theta} \Leftrightarrow \theta \leq \frac{y_n}{n} - \frac{1}{2}.$$

La vraisemblance admet un maximum global unique. L'estimateur du maximum de vraisemblance du paramètre θ est donc

$$\hat{\theta}_n = \frac{Y_n}{n} - \frac{1}{2}.$$

2. Montrer que Y_n suit une loi binomiale $\mathcal{B}(n, p)$ avec $p = \frac{1}{2} + \theta$. En déduire que $\hat{\theta}_n$ est un estimateur sans biais et convergent du paramètre θ .

Réponse : Y_n est le nombre de variables aléatoires X_i qui sont positives. Si le fait d'avoir $X_i \geq 0$ est un succès et le fait d'avoir $X_i < 0$ est un échec, Y_n est le nombre de succès parmi n expériences. Donc Y_n suit une loi binomiale $\mathcal{B}(n, p)$ où p est la probabilité du succès sur une expérience, i.e.,

$$p = P[X_i \geq 0] = \int_0^1 p(x; \theta) dx = \int_0^1 \left(\theta + \frac{1}{2}\right) dx = \theta + \frac{1}{2}.$$

On a donc $E[Y_n] = np$ et $\text{Var}[Y_n] = np(1-p)$. On en déduit

$$E[\hat{\theta}_n] = p - \frac{1}{2} = \theta.$$

L'estimateur $\hat{\theta}_n$ est donc un estimateur non biaisé de θ . De plus

$$\text{Var}[\hat{\theta}_n] = \frac{1}{n^2} \text{Var}[Y_n] = \frac{p(1-p)}{n} = \frac{\frac{1}{4} - \theta^2}{n}.$$

Comme l'estimateur $\hat{\theta}_n$ est un estimateur non biaisé de θ et que sa variance tend vers 0 lorsque n tend vers ∞ , l'estimateur $\hat{\theta}_n$ est convergent.

3. Déterminer la borne de Cramer-Rao d'un estimateur non biaisé de θ . L'estimateur $\hat{\theta}_n$ est-il l'estimateur efficace du paramètre θ ?

Réponse : on a

$$\frac{\partial^2 \ln p(x_1, \dots, x_n; \theta)}{\partial \theta^2} = -\frac{y_n}{(\theta + \frac{1}{2})^2} - \frac{n - y_n}{(\frac{1}{2} - \theta)^2}$$

d'où

$$E \left[-\frac{\partial^2 \ln p(X_1, \dots, X_n; \theta)}{\partial \theta^2} \right] = \frac{np}{(\theta + \frac{1}{2})^2} + \frac{n - np}{(\frac{1}{2} - \theta)^2} = \frac{n}{\frac{1}{4} - \theta^2}.$$

La borne de Cramer-Rao d'un estimateur non biaisé de θ est donc

$$\text{BCR}(\theta) = \frac{\frac{1}{4} - \theta^2}{n} = \text{Var}[Y_n].$$

L'estimateur Y_n est donc l'estimateur efficace du paramètre θ .

4. On désire maintenant estimer le paramètre $a = \theta + \frac{1}{2} \in]0, 1[$.

- Quel est l'estimateur des moments du paramètre a noté \tilde{a}_n construit à partir de $E[X_i]$?

Réponse : on a

$$E[X_i] = \int_{-1}^0 x \left(\frac{1}{2} - \theta \right) dx + \int_0^1 x \left(\frac{1}{2} + \theta \right) dx = \theta = a - \frac{1}{2}$$

et donc

$$a = E[X_i] + \frac{1}{2}.$$

L'estimateur des moments du paramètre a construit à partir de $E[X_i]$ est donc

$$\tilde{a}_n = \frac{1}{n} \sum_{i=1}^n X_i + \frac{1}{2}.$$

- Quel est l'estimateur du maximum de vraisemblance du paramètre a noté \hat{a}_n ?

Réponse : en utilisant le principe d'invariance fonctionnelle, on obtient directement

$$\hat{a}_n = \hat{\theta}_n + \frac{1}{2} = \frac{Y_n}{n}.$$

- On suppose qu'on dispose d'une information a priori sur le paramètre a résumée dans la loi beta de densité

$$p(a) = \frac{a^{\alpha-1}(1-a)^{\beta-1}}{B(\alpha, \beta)} \mathcal{I}_{]0,1[}(a)$$

où $\mathcal{I}_{]0,1[}(a)$ est la fonction indicatrice sur l'intervalle $]0, 1[$ et où $B(\alpha, \beta)$ est la fonction beta dont l'expression n'est pas importante pour cet exercice. Montrer que la loi a posteriori de $a|x_1, \dots, x_n$ est une loi beta dont on précisera les paramètres. Quel est l'estimateur du maximum a posteriori du paramètre a noté a_n^* ?

Réponse : la loi a posteriori de a est telle que

$$p(a|x_1, \dots, x_n) \propto p(x_1, \dots, x_n|a)p(a).$$

En remplaçant θ par $a - \frac{1}{2}$ dans l'expression précédente de la vraisemblance, on obtient

$$p(x_1, \dots, x_n | a) = (1 - a)^{n - y_n} a^{y_n}.$$

On en déduit

$$p(a | x_1, \dots, x_n) \propto (1 - a)^{n - y_n + \beta - 1} a^{y_n + \alpha - 1} \mathcal{I}_{]0,1[}(a).$$

On reconnaît une loi beta de paramètres $y_n + \alpha$ et $n - y_n + \beta$. Le maximum de cette loi a posteriori vérifie

$$\frac{\partial \ln p(a | x_1, \dots, x_n)}{\partial a} = 0 \Leftrightarrow \frac{-n + y_n - \beta + 1}{1 - a} + \frac{y_n + \alpha - 1}{a} = 0 \Leftrightarrow a = \frac{y_n + \alpha - 1}{n + \alpha + \beta - 2}.$$

L'estimateur du maximum a posteriori du paramètre a est donc

$$\hat{a}_{\text{MAP}} = \frac{Y_n + \alpha - 1}{n + \alpha + \beta - 2}.$$

Exercice 2: Tests Statistiques

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

$$p(x; \theta) = \begin{cases} \frac{1}{2} - \theta & \text{si } -1 \leq x < 0 \\ \theta + \frac{1}{2} & \text{si } 0 \leq x < 1 \\ 0 & \text{sinon} \end{cases}$$

avec $\theta \in]-\frac{1}{2}, \frac{1}{2}[$. Dans un premier temps, on cherche à effectuer le test d'hypothèses simples

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

avec $\theta_1 > \theta_0$.

1. Montrer que la statistique du test de Neyman Pearson est $Y_n = \text{card}\{i \in \{1, \dots, n\} | X_i \geq 0\}$ et déterminer la région critique du test. Commenter la forme de ce test à l'aide de l'expression de la densité $p(x; \theta)$.

Réponse : le test de Neyman Pearson est défini par

$$\text{Rejet de } H_0 \text{ si } \frac{p(x_1, \dots, x_n | \theta_1)}{p(x_1, \dots, x_n | \theta_0)} > S_\alpha$$

où S_α est un seuil dépendant du risque de première espèce α . Mais

$$\frac{p(x_1, \dots, x_n | \theta_1)}{p(x_1, \dots, x_n | \theta_0)} > S_\alpha \Leftrightarrow \frac{(1 + \theta_1)^{Y_n} (1 - \theta_1)^{n - Y_n}}{(1 + \theta_0)^{Y_n} (1 - \theta_0)^{n - Y_n}} > S_\alpha \Leftrightarrow Y_n \left[\ln \left(\frac{1 + \theta_1}{1 + \theta_0} \frac{1 - \theta_0}{1 - \theta_1} \right) \right] > K_\alpha.$$

Puisque $\theta_1 > \theta_0$, on a $\frac{1 + \theta_1}{1 + \theta_0} > 1$ et $\frac{1 - \theta_0}{1 - \theta_1} > 1$. En conséquence, le test de Neyman Pearson peut s'écrire

$$\text{Rejet de } H_0 \text{ si } Y_n > \mu_\alpha$$

où μ_α est un autre seuil dépendant du risque de première espèce α . Le test consiste donc à rejeter l'hypothèse H_0 lorsque le nombre de données positives est "grand". Mais la probabilité d'avoir une donnée positive est $\theta + \frac{1}{2}$. Comme $\theta_1 > \theta_0$, l'hypothèse H_0 correspond à θ "petit" et l'hypothèse H_1 correspond à θ "grand". La règle de décision est donc logique.

2. En remarquant que $Y_n = \sum_{i=1}^n Z_i$, où Z_i est une variable aléatoire binaire telle que $Z_i = 1$ si $X_i \geq 0$ et $Z_i = 0$ si $X_i < 0$, donner la loi approchée de Y_n pour n "grand" découlant de l'application du théorème de la limite centrale. On supposera que cette approximation est suffisamment précise pour être utilisée dans la suite de ce exercice.

Réponse : puisque les variables aléatoires X_i sont indépendantes et identiquement distribuées, il en est de même pour les variables aléatoires Z_i . On peut donc appliquer le théorème de la limite centrale à la variable aléatoire $Y_n = \sum_{i=1}^n Z_i$. On obtient alors

$$\frac{Y_n - E[Y_n]}{\sqrt{\text{Var}[Y_n]}} = \frac{Y_n - np}{\sqrt{npq}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

avec $p = P[X_i \geq 0] = \theta + \frac{1}{2}$ et $q = 1 - p = \frac{1}{2} - \theta$. Pour n "grand", on peut donc approcher la loi de Y_n par une loi normale

$$\mathcal{N}(np, npq) = \mathcal{N}\left(n\left(\theta + \frac{1}{2}\right), n\left(\frac{1}{4} - \theta^2\right)\right).$$

3. En utilisant la loi approchée déterminée à la question précédente, exprimer les risques de première et seconde espèce α et β en fonction du seuil déterminant la région critique du test, des paramètres n , θ_0 et θ_1 , et de la fonction $\phi(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$.

Réponse : Les risques α et β sont définis par

$$\alpha = P[\text{Rejeter } H_0 | H_0 \text{ vraie}], \quad \beta = P[\text{Rejeter } H_1 | H_1 \text{ vraie}].$$

On en déduit

$$\alpha = P[Y_n > \mu_\alpha | \theta = \theta_0] = P\left[\frac{Y_n - np_0}{\sqrt{np_0q_0}} > \frac{\mu_\alpha - np_0}{\sqrt{np_0q_0}}\right] = \phi\left[\frac{\mu_\alpha - np_0}{\sqrt{np_0q_0}}\right]$$

avec $p_0 = \theta_0 + \frac{1}{2}$ et $q_0 = 1 - p_0 = \frac{1}{2} - \theta_0$. De même

$$\beta = P[Y_n < \mu_\alpha | \theta = \theta_1] = P\left[\frac{Y_n - np_1}{\sqrt{np_1q_1}} < \frac{\mu_\alpha - np_1}{\sqrt{np_1q_1}}\right] = 1 - \phi\left[\frac{\mu_\alpha - np_1}{\sqrt{np_1q_1}}\right]$$

avec $p_1 = \theta_1 + \frac{1}{2}$ et $q_1 = 1 - p_1 = \frac{1}{2} - \theta_1$.

4. Déterminer les courbes COR et analyser leur comportement en fonction de n . Quelles sont les deux autres quantités notées $A(\theta_0, \theta_1)$ et $B(\theta_0, \theta_1)$ dont dépendent les courbes COR ?

Réponse : Les courbes COR sont les courbes traçant les variations de la puissance $\pi = 1 - \beta$ en fonction de α . En utilisant les résultats de la question précédente, on obtient

$$\mu_\alpha = np_0 + \phi^{-1}(\alpha)\sqrt{np_0q_0}$$

et donc

$$\pi = \phi\left[\frac{\mu_\alpha - np_1}{\sqrt{np_1q_1}}\right] = \phi\left[\sqrt{n}\frac{p_0 - p_1}{\sqrt{p_1q_1}} + \phi^{-1}(\alpha)\sqrt{\frac{p_0q_0}{p_1q_1}}\right].$$

Comme $p_0 - p_1 = \theta_0 - \theta_1 < 0$, on observe que π est une fonction croissante de n , ce qui est logique car la performance du test est d'autant meilleure que le nombre d'observations est grand. On observe également que les courbes COR dépendent des deux quantités

$$A(\theta_0, \theta_1) = \frac{p_0 - p_1}{\sqrt{p_1q_1}} = \frac{\theta_0 - \theta_1}{\sqrt{\frac{1}{4} - \theta_1^2}} \quad \text{et} \quad B(\theta_0, \theta_1) = \frac{p_0q_0}{p_1q_1} = \frac{\frac{1}{4} - \theta_0^2}{\frac{1}{4} - \theta_1^2}.$$

Barème

Exercice 1 (11 points)

1. 3 pts
2. 3 pts
3. 1 pt
4. 1 pt + 1pt + 2pts

Exercice 2 (9 points)

1. 2 pts
2. 2 pts
3. 2 pt
4. 3pts