CENTRALESUPELEC

Première Année 2018-2019

Contrôle I - Statistique et Apprentissage

Sans document.

Exercice 1

Pour l'évaluation du potentiel de production d'une usine d'éoliennes, on modélise la vitesse du vent par une variable aléatoire de distribution de type Weibull, dont la densité est donnée par

$$f_\beta(x) = 2\beta^{-2} x \exp(-x^2/\beta^2)\mathbb{I}_{\mathbb{R}+}(x),$$

où β est un paramètre strictement positif dit paramètre d'échelle.
On admettra que si $X \sim f_1$, alors :

$$\mathbb{E}_1(X) = \int_{\mathbb{R}} x\, f_1(x)\, dx = \frac{\sqrt{\pi}}{2},\ \mathbb{E}_1(X^2) = \int_{\mathbb{R}} x^2 f_1(x)\, dx = 1 \text{ et } \mathbb{E}_1(X^4) = \int_{\mathbb{R}} x^4 f_1(x)\, dx = 2$$

On dispose d'un échantillon $X_1, \ldots, X_N$ de variables aléatoires i.i.d. de loi de densité f_{β^*} où β^* est la vraie valeur du paramètre.

1. Montrer que $\forall r \in \mathbb{N}^*$, $\mathbb{E}_\beta(X^r) = \beta^r \mathbb{E}_1(X^r)$. En déduire $\mathbb{E}_\beta(X)$, $\mathbb{V}_\beta(X)$ et $\mathbb{V}_\beta(X^2)$ pour tout $\beta > 0$.
(N.B. : $\mathbb{E}_\beta$ et $\mathbb{V}_\beta$ désignent l'espérance et la variance sous la loi de densité f_β).

2. a) Donner un estimateur de β^* par la méthode des moments en utilisant le moment d'ordre 1. On notera $\hat{\beta}_1$ cet estimateur.

b) Calculer son biais et son risque quadratique moyen.

c) Montrer que $\hat{\beta}_1$ est convergent.

d) Donner la distribution asymptotique de $\sqrt{N}(\hat{\beta}_1 - \beta^*)$.

e) Donner une fonction asymptotiquement pivotale pour β^*. En déduire un intervalle de confiance de taille asymptotique 98% pour β^* en fonction de $q_{0.99}$, le quantile d'ordre 0.99 pour la loi normale centrée réduite.

3. a) Donner une expression de la vraisemblance du paramètre β en l'échantillon $(x_1, \ldots, x_N)$.

b) Montrer que l'estimateur du maximum de vraisemblance existe, qu'il est unique et qu'il s'exprime comme une fonction du moment empirique d'ordre 2 de l'échantillon. On le notera $\hat{\beta}_2$.

c) Montrer que $\hat{\beta}_2$ est convergent.

d) En utilisant la méthode Delta, déterminer un intervalle de confiance pour β de taille asymptotique 0.98 à partir de $\hat{\beta}_2$ et en fonction de $q_{0.99}$ le quantile d'ordre 0.99 pour la loi normale centrée réduite.

Exercice 2

Soit $(X_1, X_2, \ldots, X_N)$ un échantillon de variables aléatoires indépendantes, identiquement distribuées pour un modèle statistique paramétré par $\theta \in \Theta \subset \mathbb{R}$. Dans cet exercice, nous allons considérer le test du rapport de vraisemblance pour les tests paramétriques de la forme :

$$H_0 : \theta \in \Theta_0 \text{ contre } H_1 : \theta \in \Theta_0^c$$

où $\Theta_0 \subsetneq \Theta$ est donné et Θ_0^c est le complémentaire de Θ_0 dans Θ.

Pour cela, on construit la statistique de test :

$$\lambda(X_1, \ldots, X_N) = \frac{\sup_\Theta \mathcal{L}(\theta; X_1, \ldots, X_N)}{\sup_{\Theta_0} \mathcal{L}(\theta; X_1, \ldots, X_N)}$$

où $\mathcal{L}(\theta; X_1, \ldots, X_N)$ représente la fonction de vraisemblance du paramètre θ pour l'échantillon $(X_1, \ldots, X_N)$. On est alors conduit à une zone de rejet définie par :

$$R_\alpha = \{(x_1, \ldots, x_N) \text{ tels que } \lambda(x_1, \ldots, x_N) > c_\alpha\}$$

où c_α est choisi tel que :

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left((X_1, \ldots, X_N) \in R_\alpha \right) = \alpha \ .$$

1) Soit $X = (X_1, X_2, \ldots, X_N)$, échantillon de variables aléatoires indépendantes, identiquement distribuées pour une loi normale $\mathcal{N}(\mu, 1)$.

a) Calculer $\hat{\mu}$ l'estimateur du maximum de vraisemblance pour μ sur $\Theta = \mathbb{R}$ et donner sa loi.

b) On désire réaliser le test d'hypothèse paramétrique :

$$H_0 : \mu = \mu_0 \text{ contre } H_1 : \mu \neq \mu_0$$

où μ_0 est donné.
Déterminer le rapport de vraisemblance $\lambda(X)$ et en déduire une forme simplifiée de la zone de rejet. Déterminer c_α.

2) On considère cette fois une famille de lois exponentielles avec des densités de la forme :

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

Soit $X = (X_1, X_2, \ldots, X_N)$, échantillon de variables aléatoires indépendantes, identiquement distribuées selon $f(x;\theta)$. On teste :

$$H_0 : \theta \leq \theta_0 \text{ contre } H_1 : \theta > \theta_0$$

où θ_0 est donné.

a) Calculer $\hat{\theta}$ l'estimateur du maximum de vraisemblance pour θ sur $\Theta = \mathbb{R}$ en faisant apparaître $X_{(1)} := \min_{1 \leq i \leq N} X_i$.

b) Déterminer $\lambda(X)$ pour le test paramétrique considéré et en déduire une forme simplifiée de la zone de rejet. Déterminer c_α.

Exercice 3

Nous rappelons la définition d'une Loi multinomiale d'ordre K. Soit $N \in \mathbb{N}^*$, $p \in]0;1[^K$ telle que $\sum_{i=1}^K p_i = 1$. On appelle loi multinomiale de paramètres (N,p), la loi de probabilité sur $\{0;1;\ldots;N\}^K$ définie par la fonction de masse :

$$P(x_1,\ldots,x_K) = \begin{cases} \dfrac{N!}{\prod_{i=1}^K x_i!} \prod_{i=1}^K p_i^{x_i}, \text{ si } (x_1,\ldots,x_K) \in \{0;1;\ldots;N\}^K \text{ tel que } \Sigma_{i=1}^K x_j = N \\ \\ 0 \text{ sinon.} \end{cases}$$

On note $X \sim M(N,p)$.
Nous rappelons également la définition d'une loi de Dirichlet d'ordre K.
Soit $a = (a_1,\ldots,a_K) \in \left(\mathbb{R}_+^*\right)^K$. On appelle loi de Dirichlet de paramètre a, la loi de probabilité de support $\mathcal{S} = \left\{ x \in [0;1]^K : \sum_{i=1}^K x_i = 1 \right\}$, définie par la densité :

$$p(x_1,\ldots,x_K) = \begin{cases} \dfrac{1}{\beta(a)} \prod_{i=1}^K {x_i}^{a_i - 1} \text{ si } (x_1,\ldots,x_K) \in \mathcal{S} \\ 0 \text{ sinon} \end{cases}$$

Elle est notée $\mathrm{Dir}(a)$. La loi de Dirichlet d'ordre 2 est la loi Bêta, $\mathrm{Dir}(a_1,a_2) = \mathrm{Bêta}(a_1,a_2)$.
1) Sans réaliser le calcul, dire comment déterminer la fonction $a \mapsto \beta(a)$. On la supposera connue pour la suite.

2) Soit Y, une variable aléatoire qui suit une loi Multinomiale d'ordre K, $K \geq 3$, et de paramètres (N,θ), où N est connu et $\theta = (\theta_1,\ldots,\theta_K)$ inconnu. Soit $y = (y_1,\ldots,y_K)$

une observation de la variable Y. On se place dans le cadre de l'estimation Bayésienne pour θ. On suppose la distribution a priori $\pi = \mathrm{Dir}(a)$, avec $a = (a_1, \ldots, a_K) \in \left(\mathbb{R}_+^*\right)^K$. Déterminer la loi de la distribution a posteriori $p(\theta|y)$.

3) a) Montrer que si $(X_1, \ldots, X_{K-1}, X_K)$ suit une loi de Dirichlet de paramètres $(a_1, \ldots, a_{K-1}, a_K)$, alors $(X_1, \ldots, X_{K-2}, X_{K-1}+X_K)$ suit une loi de Dirichlet de paramètres $(a_1, \ldots, a_{K-2}, a_{K-1}+a_K)$.

b) On note $a_r = \sum_{i=3}^K a_i$ et $y_r = \sum_{i=3}^K y_i$. Déduire de la question précédente que

$$p(\theta_1, \theta_2|y) \propto \theta_1^{a_1+y_1-1}\theta_2^{a_2+y_2-1}(1-\theta_1-\theta_2)^{a_r+y_r-1} \ .$$

4) On réalise le changement de variable ϕ :

$$(\alpha_1, \alpha_2) = \left(\frac{\theta_1}{\theta_1+\theta_2}, \theta_1+\theta_2\right) = \phi(\theta_1, \theta_2).$$

a) Montrer que ϕ est un $\mathcal{C}^1$ difféomorphiseme de $]0;1[^2$ dans $]0;1[^2$.

b) En déduire la densité conditionnelle $p(\alpha_1, \alpha_2|y)$ à une constante de normalisation près.

c) En déduire finalement la loi associée à la densité conditionnelle $p(\alpha_1|y)$.