Devoir de statistique mathématique

Partie 1. Modèle exponentiel pour le temps de survie dans un essai randomisé

Considérons un modèle exponentiel pour le temps de survie dans un essai randomisé et contrôlé par placebo de l'azathioprine pour la cirrhose biliaire primitive. Les temps de survie (en jours) des n patients du groupe placebo sont donnés par les variables aléatoires X_i avec les indicatrices de censure $\gamma_i \quad i = 1, \ldots, n$, tandis que les temps de survie des m patients du groupe de contrôle sont désignés par Y_i et ont des indicatrices de censure $\delta_i \quad i = 1, \ldots, m$. Les temps de survie non censurés (en partie non observés) suivent des modèles exponentiels avec les paramètres η et $\eta\theta$ dans les groupes placebo et traités respectivement ($\eta, \theta > 0$).

a. Interpréter η et θ et montrer que la log-vraisemblance des données est donnée par

$$\ell(\eta, \theta) = (n\bar{\gamma} + m\bar{\delta}) \log(\eta) + m\bar{\delta} \log(\theta) - \eta(n\bar{x} - m\theta\bar{y}),$$

où $\bar{\gamma}, \bar{\delta}, \bar{x}$ et $\bar{y}$ sont les moyennes empiriques des γ_i, δ_i, x_i et y_i respectivement.

b. Calculer le MLE $\left(\hat{\eta}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}}\right)$.

c. Proposer une approximation de l'information de Fisher $I_{m,n}\left(\hat{\eta}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}}\right)$ apportée par toutes les observations en remplaçant $\mathbb{E}(\gamma_i)$ et $\mathbb{E}(\delta_i)$ par les estimateurs appropriés[1]. Dans le cas d'un vecteur de paramètres l'information de Fisher est une matrice symétrique où $-\mathbb{E}\left(\frac{\partial^2}{\partial\eta^2}\ell(\eta, \theta)\right)$ et $-\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\ell(\eta, \theta)\right)$ sont les éléments diagonaux et $-\mathbb{E}\left(\frac{\partial^2}{\partial\theta\partial\eta}\ell(\eta, \theta)\right)$ est l'élément anti-diagonal.

d. Montrer que $\hat{\theta}_{\mathrm{ML}}\sqrt{\frac{n\bar{\gamma}+m\bar{\delta}}{n\bar{\gamma}m\bar{\delta}}}$ est une approximation de l'écart type asymptotique de $\hat{\theta}_{\mathrm{ML}}$ et proposer une formule générale d'un intervalle de confiance $\alpha \cdot 100\%$ asymptotique.

e. Considérer la transformation $\psi = \log(\theta)$. À l'aide de la delta méthode, proposer un intervalle de confiance $\alpha \cdot 100\%$ asymptotique pour ψ.

Partie 2. Un mélange de deux lois

On se donne deux fonctions de répartition F_1 et F_2 de deux variables aléatoires Z_1 et Z_2 et un nombre de positif $p \in]0, 1[$. Alors $F(t) = pF_1(t) + (1-p)F_2(t)$ est encore une fonction de réparition. On dit que la nouvelle loi obtenue est le *mélange* de la loi de Z_1 et la loi de Z_2. Attention cela n'a rien à avoir avec la loi de $pZ_1 + (1-p)Z_2$.

a. Prouver que F est la fonction de répartition de la v.a X suivante :

$$X = \begin{cases} Z_1 & \text{si } Y = 1 \\ Z_2 & \text{si } Y = 0 \end{cases}$$

où Y est une variable de Bernoulli de paramètre p, indépendante du couple (Z_1, Z_2).

On se place maintenant de le cas où $Z_1 \sim \mathcal{U}([0, 1])$ et $Z_2 \sim \mathcal{U}([0, \theta])$ avec $\theta \geq 1$. On note $F_{\theta,p}$ la fonction de répartition de X dans ce cas.

b. On suppose qu'on observe un n-échantillon de même loi que X, et qu'on cherche à estimer à la fois p et θ. Montrer que si l'on exclut une valeur de θ à préciser, ce modèle statistique est identifiable (c'est-à-dire que la fonction $(\theta, p) \mapsto F_{\theta,p}$ est injective). Pour cela, il faut procéder comme suit

[1] On considère la moyenne empirique est une bonne approximation de la moyenne théorique pour les X_i et Y_i. Dans cette question nous n'avons pas besoin de faire un calcul d'espérence.

1. Exprimer $F_{\theta,p}$ sous la forme

$$F_{\theta,p}(x) = \begin{cases} ? & \text{si } x \leq 0 \\ ? & \text{si } 0 < x \leq 1 \\ ? & \text{si } 1 < x \leq \theta \\ ? & \text{si } \theta < x. \end{cases}$$

2. Identifier une valeur particulière de θ pour laquelle on peut trouver différentes valeurs de p qui conduisent à la même fonction $F_{\theta,p}$.

3. Montrer l'identifiablité du modèle en exluant cette valeur particulière de θ.

c. Montrer que $\widehat{\theta} = \max_{i \leq i \leq n} X_i$ est un estimateur qui converge en probabilité vers θ[2].

d. Écrire la densité de X. Calculer $\mathbb{E}(X)$, en déduire un estimateur $\widehat{p}$ de p. Montrer que si $\theta > 1$, $\widehat{p}$ converge en porbabilité vers p[3].

e. Pour $\theta > 1$, montrer que $n\left(\theta - \widehat{\theta}\right)$ converge en loi vers une loi exponentielle $\mathcal{E}\left(\frac{(1-p)}{\theta}\right)$[4]. En déduire que

$$\left[\widehat{\theta}, \widehat{\theta} + \frac{\widehat{\theta}\ln(1/\alpha)}{n(1-\widehat{p})}\right]$$

est un intervalle de confiance asymptotique pour θ de niveau $1 - \alpha$[5].

f. Proposer une procédure de test pour tester $\theta = 1$ contre $\theta > 0$[6].

[2]indications : on utilisera directement la définition de la convergence en probabilité en exprimant $\mathbb{P}\left(|\widehat{\theta} - \theta| > \varepsilon\right)$ en fonction de $F_{\theta,p}$

[3]indication : c'est une estimation par la méthode des moments, obtenue en remplaçant $\mathbb{E}(X)$ par $\bar{X}_n$ et θ par son estimateur

[4]indication : utiliser la limite suivante $\lim_{n\to+\infty}\left(1 + \frac{x}{n}\right)^n = e^x$

[5]indication: construire à l'aide du théorème de Slutsky une variable aléatoire proportionnelle à $n\left(\theta - \widehat{\theta}\right)$ de loi asymptotique $\mathcal{E}(1)$

[6]indication : penser à un seuil de type $c = (1-\alpha)^{\frac{1}{n}}$