Outils avancés pour la statistique – Projet

Travail à effectuer

— Un fichier de script pour la Partie 1 et un fichier de script pour la Partie 2. Dans ces deux fichiers, le code devra pouvoir être exécuté ligne par ligne, en suivant l'ordre des questions.
— Un document PDF contenant les réponses aux questions et décrivant votre démarche lorsque cela est pertinent. Les calculs devront être clairs et détaillés, ainsi que la démarche utilisée.
— Le barème provisoire est donné dans le Tableau 1.

Exercice 1		Exercice 2					
Q1	Q2	Q1	Q2	Q3	Q4	Q5	Q6
6pts	1pt	1pt	1pt	1pt	1pt	1pt	8pts
7pts		13pts					

Tableau 1 – Barème.

1 Algorithme EM et survie

On s'intéresse au temps passé sans emploi sur un échantillon de personnes de taille $n = 12$. L'étude s'arrête au bout de 5 ans et il reste encore 2 personnes qui n'ont pas trouvé d'emploi. Les données sont présentées dans le Tableau 2. On suppose que les temps Y_i sont i.i.d. de loi exponentielle de paramètre inconnu λ.

1. À l'aide d'un algorithme EM, estimer le paramètre de cette loi exponentielle.
2. Existe-t-il une solution explicite à l'estimation de λ ? (C'est-à-dire que l'on peut l'obtenir sans exécuter la boucle de l'algorithme et sans utiliser une autre modélisation.)

Indication : $\sum_{i=1}^{10} Y_i = 36860$**.**

Ind. i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	1274	4860	2376	10567	526	1323	1890	2307	3042	1999	5106	833
Z_i	913	510	1224	63	326	305	486	213	1813	843	1826*	1826*
Y_i	2187	5370	3600	10630	852	1628	2376	2520	4855	2842	NA	NA

Tableau 2 – Nombre de jours passés sans emploi sur un échantillon de 12 individus, sans emploi au début de l'étude. L'individu i était sans emploi depuis X_i jours au début de l'étude. Il est resté Z_i jours sans emploi avant d'en trouver un. Les individus 1 à 10 ont donc passé $Y_i = X_i + Z_i$ jours sans emploi. (Aucun des individus n'a quitté l'emploi qu'il a trouvé au cours de l'étude.) Les individus 11 et 12 n'ont pas trouvé d'emploi au cours de l'étude mais on sait qu'ils auront passé au moins $X_i + Z_i$ jours sans emploi.
* la durée de l'étude est de 1826 jours.

2 Biais corrigé et biais corrigé et accéléré

Pour améliorer les intervalles de confiance (IC) bootstrap percentiles, Efron a proposé :

— les IC bootstrap avec biais corrigé ou BC ;

— les IC avec biais corrigé et accéléré ou BC_a, décrits dans Efron (1987).

En reprenant les notations de l'article d'Efron (1987), on dispose d'un échantillon $(X_1, \ldots, X_n)$ i.i.d. de fonction de répartition F et on veut construire un intervalle de confiance pour un paramètre $\theta = t(F)$. Nous nous intéresserons exclusivement à l'IC non paramétrique, même si l'article propose aussi la construction d'un IC paramétrique. On note Φ la fonction de répartition d'une loi $\mathcal{N}(0\,;\,1)$.

On peut introduire $\hat{\theta} = t\left(\hat{F}_n\right)$ avec

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\{X_j \leq x\}$$

et $\hat{\theta}^* = t\left(\hat{F}_n^*\right)$ avec

$$\hat{F}_n^*(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left\{X_j^* \leq x\right\}$$

où $(X_1^*, \ldots, X_n^*)$ est un échantillon issu d'un tirage avec remise dans $(X_1, \ldots, X_n)$, de sorte que l'on est capable de générer autant de réalisations de $\hat{\theta}^*$ que l'on veut :

— Pour $b = 1, \ldots, B$

— Tirer un échantillon $\left(X_1^{*\{b\}}, \ldots, X_n^{*\{b\}}\right)$ avec remise dans $(X_1, \ldots, X_n)$.

— Calculer $\hat{\theta}^{*\{b\}} = t\left(\hat{F}_n^{*\{b\}}\right)$ avec

$$\hat{F}_n^{*\{b\}}(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left\{X_j^{*\{b\}} \leq x\right\}.$$

On estime la fonction de répartition de $\hat{\theta}^*$ par la fonction de répartition empirique

$$\hat{G}(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{\hat{\theta}^{*\{b\}} \leq t\right\} = \frac{\#\left\{\hat{\theta}^{*\{b\}} \leq t,\, 1 \leq b \leq B\right\}}{B}.$$

$\hat{G}^{-1}(\alpha)$ est alors le quantile empirique d'ordre α calculé sur la base de l'échantillon $\left\{\hat{\theta}^{*\{b\}},\, 1 \leq b \leq B\right\}$, que l'on avait noté $\hat{q}_\alpha^*$ dans le cours. L'intervalle de confiance BC_a non paramétrique pour θ est

$$\left[\hat{G}^{-1}\left(\Phi\left(z[\alpha/2]\right)\right);\, \hat{G}^{-1}\left(\Phi\left(z[1-\alpha/2]\right)\right)\right]$$

où

$$z[\alpha] = z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}$$

et $z^{(\alpha)} = \Phi^{-1}(\alpha)$ est le quantile d'ordre α d'une loi $\mathcal{N}(0\,;\,1)$.

La constante de biais z_0 et la constante d'accélération a sont calculées de la manière suivante

$$z_0 = \Phi^{-1}\left(\hat{G}\left(\hat{\theta}\right)\right) \quad \text{et} \quad a = \frac{\sum_{i=1}^{n} U_i^3}{6\left(\sum_{i=1}^{n} U_i^2\right)^{3/2}}$$

avec

$$U_i = \lim_{\Delta \to 0} \frac{t\left((1-\Delta)\hat{F}_n + \Delta D_i\right) - t\left(\hat{F}_n\right)}{\Delta}$$

où [1] $D_i(x) = \mathbf{1}\{X_i \leq x\}$. Les intervalles de confiance BC sont obtenus pour $a = 0$.

1. Montrer que lorsque $z_0 = a = 0$, on retrouve l'intervalle de confiance percentile d'Efron vu en cours.
2. Montrer que pour $\theta_1 = t_1(F) = \mathbf{E}[X]$ que l'on estime par $\hat{\theta}_1 = t_1\left(\hat{F}_n\right) = \overline{X}_n = n^{-1}\sum_{j=1}^n X_j$, on a
$$U_{i,1} = X_i - \overline{X}_n.$$
3. Montrer que pour $\theta_2 = t_2(F) = \mathrm{Var}(X) = \sigma^2(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right]$ que l'on estime par $\hat{\theta}_2 = t_2\left(\hat{F}_n\right) = n^{-1}\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^2$, on a
$$U_{i,2} = \left(X_i - \overline{X}_n\right)^2 - \frac{1}{n}\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^2.$$
4. Montrer que pour $\theta_3 = t_3(F) = \sigma(X) = \sqrt{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}$ que l'on estime par
$$\hat{\theta}_3 = t_3\left(\hat{F}_n\right) = \sqrt{n^{-1}\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^2},$$
on a
$$U_{i,3} = \frac{U_{i,2}}{2\hat{\theta}_2}.$$
5. Montrer que pour $\theta_4 = t_4(F) = \dfrac{\mathbf{E}\left[(X - \mathbf{E}[X])^3\right]}{\sigma^3(X)}$ que l'on estime par
$$\hat{\theta}_4 = t_4\left(\hat{F}_n\right) = \frac{n^{-1}\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^3}{\left[n^{-1}\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^2\right]^{3/2}},$$
on a
$$U_{i,4} = \frac{\left(X_i - \overline{X}_n\right)^3 - \dfrac{3}{n}\left(X_i - \overline{X}_n\right)\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^2 - \dfrac{1}{n}\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^3}{\left[n^{-1}\sum_{j=1}^n \left(X_j - \overline{X}_n\right)^2\right]^{3/2}}.$$
6. Faire une étude des longueurs et des taux de couverture des intervalles de confiance percentile d'Efron, BC et BC_a sur la base de $N = 1000$ échantillons i.i.d. de taille $n = 20$ et de loi $\Gamma(5\,;\,1/2)$, et pour les paramètres θ_1, θ_2, θ_3, θ_4 et θ_5. On utilisera un nombre d'échantillons bootstrap adéquat et on construira des intervalles à 90%.

1. Ici c'est bien D_i et non pas la masse de Dirac δ_i en X_i, qu'il faut ajouter.

Remarque 1. La notation $t(F)$ signifie que le paramètre ne dépend que de F. Par exemple, pour t_2 on a bien

$$\begin{aligned} t_2(F) &= \int x^2 F(dx) - \left[\int x F(dx)\right]^2 \\ &= \int x^2 f(x)\, dx - \left[\int x f(x)\, dx\right]^2 \end{aligned} \tag{1}$$

où f est la densité de X, égale à la dérivée de F presque partout. La quantité $t_2\left(\hat{F}_n\right)$ s'obtient en remplaçant chaque occurence de F par $\hat{F}_n$ dans (1), ce qui donne

$$\begin{aligned} t_2\left(\hat{F}_n\right) &= \int x^2 \hat{F}_n(dx) - \left[\int x \hat{F}_n(dx)\right]^2 \\ &= n^{-1}\sum_{j=1}^{n} X_j^2 - \left[n^{-1}\sum_{j=1}^{n} X_j\right]^2. \end{aligned}$$

Concernant le mélange $(1-\Delta)\hat{F}_n + \Delta D_i$ on a

$$t_2\left((1-\Delta)\hat{F}_n + \Delta D_i\right) = \frac{1-\Delta}{n}\sum_{j=1}^{n} X_j^2 + \Delta X_i^2 - \left[\frac{1-\Delta}{n}\sum_{j=1}^{n} X_j + \Delta X_i\right]^2.$$

Références

Bradley Efron. Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.*, 82(397) :171–200, 1987. ISSN 0162-1459. URL `http://links.jstor.org/sici?sici=0162-1459(198703)82:397<171:BBCI>2.0.CO;2-H&origin=MSN`. With comments and a rejoinder by the author.