Statistique

Examen, M1 Statistique & Économétrie

Décembre 2015

L'examen dure 2 heures, aucun document n'est autorisé
La calculatrice n'est pas autorisée non plus.

Tous les exercices étaient tirés de feuilles d'exercice, et la plupart avaient été faits en classe. En bleu figurent des éléments de correction, qui peuvent contenir encore quelques erreurs de calcul. Je renvoie aux notes prises pendant les exercices pour plus de détails (ce sont juste ici des éléments de correction).

1. On dispose d'un échantillon $\{x_1, \cdots, x_n\}$ tiré suivant un mélange

$$X = \begin{cases} U_1 \sim \mathcal{U}([0,a]) \text{ avec probabilité } \theta \\ U_2 \sim \mathcal{U}([0,b]) \text{ avec probabilité } 1-\theta \end{cases}$$

avec $\theta \in (0,1)$ et $a < b$. Soit N le nombre d'observations entre 0 et a dans l'échantillon. Quelle est la loi de N ? Quel est l'estimateur du maximum de vraisemblance de θ.

Calculons la probabilité que X tombe entre 0 et a,

$$\mathbb{P}(X \in [0,a]) = \sum_{i=1,2} \mathbb{P}(X \in [0,a] | X = U_i) = \sum_{i=1,2} \mathbb{P}(U_i \in [0,a] | X = U_i)$$

i.e.

$$\mathbb{P}(X \in [0,a]) = 1 \cdot \theta + \frac{a}{b} \cdot (1-\theta)$$

de telle sorte que

$$N \sim \mathcal{B}(n,p) \text{ avec } p = \frac{a + (b-a)\theta}{b}.$$

Notons n_a le nombre d'observations inférieures à a. Alors

$$\mathcal{L} = \prod_{i=1}^{n} \binom{n}{n_a} p^{n_a} [1-p]^{n-n_a}$$

On écrit alors la log-vraisemblance, que l'on maximise en θ. Sinon, on peut se souvenir que le maximum de vraisemblance de p est

$$\widehat{p} = \frac{n_a}{n}$$

or on sait que si $\theta = g(p)$ alors $\widehat{\theta} = g(\widehat{p})$ dès lors que g est strictement croissante. Comme c'est le cas ici,

$$\widehat{\theta} = \frac{b}{b-a}[\widehat{p} - a] = \frac{b}{b-a}\left[\frac{n_a}{n} - a\right].$$

2. Considérons un échantillon tiré suivant une loi de densité

$$f_\theta(x) = \frac{2x}{\theta^2} \text{ pour } x \in [0, \theta].$$

Trouver l'estimateur du maximum de vraisemblance de la médiane de la distribution.

Essayons décrire la médiane de notre disposition,

$$F_\theta(x) = \int_0^x f_\theta(t)dt = \frac{1}{\theta^2}\int_0^x 2tdt = \left(\frac{x}{\theta}\right)^2$$

sur $[0, \theta]$. Aussi, $F_\theta(m) = 1/2$ si $m/\theta = 1/\sqrt{2}$ soit $m = \theta/\sqrt{2}$. On peut récrire la densité sous la forme

$$f_m(x) = \frac{x}{m^2} \text{ pour } x \in [0, \sqrt{2}m].$$

La vraisemblance est ici

$$\mathcal{L} = \prod_{i=1}^n \frac{x_i}{m^2}\mathbf{1}(x_i \leq \sqrt{2}m) = \frac{\prod_{i=1}^n x_i}{m^{2n}}\mathbf{1}(\max\{x_i\} \leq \sqrt{2}m)$$

qui est maximale lorsque $\sqrt{2}m = \max\{x_i\}$, i.e.

$$\widehat{m} = \frac{\max\{x_i\}}{\sqrt{2}}.$$

3. Considérons un échantillon $\{x_1, \cdots, x_n\}$ tiré suivant une loi $\mathcal{N}(\theta, 1)$. Malheureusement, seules des indicatrices, indiquant que les observations étaient positives, ont été gardées, i.e. $\{y_1, \cdots, y_n\}$, avec $y_i = \mathbf{1}(x_i > 0)$. Quelle est l'estimateur du maximum de vraisemblance de θ?

Notons que

$$p = \mathbb{P}(Y = 1) = \mathbb{P}(X > 0) = \mathbb{P}(X - \theta > -\theta) = \Phi(-\theta)$$

et rappelons que $Y \sim \mathcal{B}(p)$. Soit n_0 le nombre d'observations positives, i.e. $n_0 = Y_1 + \cdots + Y_n$. La condition du premier ordre est ici

$$\frac{\log \mathcal{L}(p)}{\partial p} = \frac{n_0}{p} - \frac{n - n_0}{1 - p}$$

de telle sorte que $\widehat{p} = n_0/n = \overline{Y}$, et donc l'estimateur du maximum de vraisemblance, en θ, est

$$\widehat{\theta} = -\Phi^{-1}\left(\frac{n_0}{n}\right)$$

4. Considérons un échantillon $\{x_1, \cdots, x_n\}$ tiré suivant une loi Beta $\mathcal{B}(\theta, \theta)$. (1) Trouver une statistique exhaustive pour θ pour un échantillon tiré suivant une loi Beta $\mathcal{B}(\theta, \theta)$. (2) Trouver un test de niveau α de $H_0 : \theta = 1$ contre $H_1 : \theta = 2$.

Rappelons que

$$f_\theta(x) = \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} \cdot x^{\theta-1}(1-x)^{\theta-1} \text{ avec } x \in [0,1].$$

Selon le théoème de factorisation, on a une statistique exhaustive si

$$\mathcal{L}(\theta) = h(\boldsymbol{x}) \cdot g_\theta(s(\boldsymbol{x})),$$

Or ici

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} \cdot x_i^{\theta-1}(1-x_i)^{\theta-1} = g_\theta(s(\boldsymbol{x}))$$

avec comme statistique exhaustive

$$s(x_1, \cdots, x_n) = \prod_{i=1}^{n} x_i \cdot (1-x_i)$$

et

$$g_\theta(s) = \left(\frac{\Gamma(2\theta)}{\Gamma(\theta)^2}\right)^n \cdot s^{\theta-1}$$

Utilisons ici le test du rapport de vraisemblance. On va rejeter $H_0 : \theta = \theta_0$ au profit de $H_1 : \theta = \theta_1$ si

$$\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} > k$$

autrement dit

$$\frac{s(\boldsymbol{x})^{\theta_1-1}}{s(\boldsymbol{x})^{\theta_0-1}} > k'$$

i.e.

$$s(\boldsymbol{x})^{\theta_1-\theta_0} > k$$

or $\theta_1 - \theta_0 > 0$ donc

$$s(\boldsymbol{x}) > c$$

Le test sera de niveau α si

$$\mathbb{P}(s(X_1, \cdots, X_n) > c | X_i \sim \mathcal{B}(\theta_0, \theta_0)) = \alpha.$$

Il faut faire un peu de calculs pour trouver la valeur du seuil, mais ce n'était pas demandé pour avoir les points... Dans le cas particulier où $\theta = 1$, on a des lois uniformes, ce qui simplifie les calculs...

5. On dispose de n observations suivant une loi exponentielle $\mathcal{E}(\theta)$. On souhaite tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$. Écrire la vraisemblance et donnez la région de rejet du test de Neyman-Pearson de niveau $\alpha = 5\%$. Quelle est la puissance de ce test.

La vraisemblance est ici

$$\mathcal{L} = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right)$$

Avec un test à la Neyman Pearson, on regarde le rapport de vraisemblance. On va rejeter $H_0 : \theta = \theta_0$ au profit de $H_1 : \theta = \theta_1$ si

$$\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} > k$$

autrement dit

$$\left(\frac{\theta_1}{\theta_0}\right)^n \exp\left((\theta_1 - \theta_0)\sum_{i=1}^{n} x_i\right) > k$$

Le test eut être basé sur la somme des observations, ou leur moyenne, $\overline{x}$. Si on suppose que $\theta_1 > \theta_0 > 0$, on va rejeter $H_0 : \theta = \theta_0$ au profit de $H_1 : \theta = \theta_1$ si $\overline{x} > c$. Comme on l'avait vu en classe, on peut utiliser un test exact, en utilisant le fait que la somme de lois exponentielles (indépendante) est une loi Gamma. Une alternative peut être d'utiliser le théorème central limite. Rappelons que l'on veut

$$\mathbb{P}\left(\overline{X} > c | H_0\right) = \alpha$$

où le conditionnement par H_0 signifie que $X_i \sim \mathcal{E}(\theta_0)$. Dans ce dernier cas, $\mathbb{E}[X_i] = \theta_0^{-1}$ alors que $\mathrm{Var}[X_i] = \theta_0^{-2}$. Aussi, $\mathbb{E}[\overline{X}] = \theta_0^{-1}$ alors que $\mathrm{Var}[\overline{X}] = \theta_0^{-2}/n$

$$\mathbb{P}\left(\overline{x} > c | H_0\right) = \mathbb{P}\left(\sqrt{n}\frac{\overline{X} - \theta_0^{-1}}{\theta_0^{-1}} > \sqrt{n}\frac{c - \theta_0^{-1}}{\theta_0^{-1}} \middle| H_0\right) = \alpha.$$

D'après le théorème central limite garantie que quand $n \to \infty$, sous H_0,

$$\sqrt{n}\frac{\overline{X} - \theta_0^{-1}}{\theta_0^{-1}} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,1).$$

alors $\alpha = \mathbb{P}\left(\overline{x} > c | H_0\right) = \mathbb{P}\left(Z > \sqrt{n}\frac{c - \theta_0^{-1}}{\theta_0^{-1}}\right) = 1 - \Phi\left(\sqrt{n}\frac{c - \theta_0^{-1}}{\theta_0^{-1}}\right)$ soit

$$\sqrt{n}\frac{c - \theta_0^{-1}}{\theta_0^{-1}} = \Phi^{-1}(1-\alpha) \sim 1.64$$

de telle sorte que

$$c = \frac{1}{\theta_0}\left(1 + \frac{1.64}{\sqrt{n}}\right).$$

Pour la puissance du test, on calcule $\mathbb{P}(\overline{X} > c|H_1)$, Sous H_1, $\mathbb{E}[\overline{X}] = \theta_1^{-1}$ alors que $\text{Var}[\overline{X}] = \theta_1^{-2}/n$, et la puissance est alors

$$\mathbb{P}\left(Z > \sqrt{n}\frac{c - \theta_1^{-1}}{\theta_1^{-1}}\right)$$

avec $Z \sim \mathcal{N}(0, 1)$. Après, on peut continuer un peu les calculs, simplifier, etc.