CPP - la prépa des INP (2 ème année). Bordeaux, 16/04/2014.

Devoir de statistiques: CORRIGE
durée 2h

Données: On rappelle que si Z suit une loi $\mathcal{N}(0,1)$, on a

$$\mathbb{P}(Z \leq 1.96) \simeq 0,975 \text{ et } \mathbb{P}(Z \leq 1.65) \simeq 0,95.$$

Exercice 1. On considère une variable aléatoire X de densité f_θ avec $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$ et

$$f_\theta(x) = \begin{cases} \frac{1}{2} - \theta & \text{si } x \in [-1, 0[, \\ \frac{1}{2} + \theta & \text{si } x \in [0, 1], \\ 0 & \text{sinon} \end{cases}$$

1. Représenter f_θ et justifier que f_θ est bien une densité de probabilité pour $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$.

2. Calculer l'espérance et la variance de X. (Les résultats sont $\mathbb{E}[X] = \theta$ et $\mathrm{Var}(X) = \frac{1}{3} - \theta^2$).

On considère maintenant $X_1, \dots X_n$ des variables aléatoires indépendantes et de même loi. On suppose que la loi commune est de densité f_θ avec $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$ inconnu. On va chercher à estimer θ.

3. Proposer un esitmateur $\hat{\theta}$ de θ basé sur la méthode des moments. On prend $\hat{\theta}_n = \frac{1}{n} \sum X_i$.

4. Calculer son biais et son risque quadratique. Il est sans biais ($\mathbb{E}_\theta[\hat{\theta}_n] = \theta$) et de risque quadratique:

$$\mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] = \mathrm{Var}_\theta(\hat{\theta}_n) = \frac{\mathrm{Var}(X_1)}{n} = \frac{\frac{1}{3} - \theta^2}{n}.$$

On va maintenant s'intéresser à l'estimateur du maximum de vraisemblance de θ.

5. On note $N_1 = \sum_{i=1}^{n} \mathbf{1}_{\{X_i \geq 0\}}$ et $N_2 = \sum_{i=1}^{n} \mathbf{1}_{\{X_i < 0\}}$. Soient $x_1, \dots, x_n \in [-1, 1]$ et $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$, écrire la vraisemblance du modèle au point $(x_1, \dots, x_n; \theta)$ en fonction de N_1, N_2 et θ.

6. Montrer que l'estimateur du maximum devraisemblance existe et vaut:

$$\hat{\theta}_{MV} = \frac{N_1 - N_2}{2(N_1 + N_2)} = \frac{N_1}{n} - \frac{1}{2}.$$

7. On pose $1_{\{X_i \geq 0\}}$, pour $1 \leq i \leq n$. Calculer l'espérance et la variance des variables aléatoires Y_i. En déduire l'espérance et le risque quadratique de $\hat{\theta}_{MV}$.

8. Quel estimateur vaut-il mieux utiliser entre $\hat{\theta}$ et $\hat{\theta}_{MV}$? Justifier.

9. 0n veut maintenant tester : l'hypothèse nulle: $H_0 = \{\theta = 0\}$ contre l'hypothèse alternative: $H_1 = \{\theta > 0\}$.

   On suppose ici que la taille de l'échantillon est suffisament grande pour pourvoir utiliser le théorème central limite. Proposer un test de niveau de confiance asymptotique 0,95 basé sur l'estimateur $\hat{\theta}$ ou $\hat{\theta}_{MV}$. On justifiera notamment la forme de la région d'acceptation ou de rejet du test.

Exercice 2. On considère $(X_1, \ldots, X_n)$ des variables aléatoires indépendantes de loi géométrique de paramètre p inconnu $(0 \leq p \leq 1)$. On pose $q = 1 - p$. On rappelle X est à valeurs dans $\mathbb{N}^*$ et que pour $k \geq 1$, $\mathbb{P}(X = k) = q^{k-1}p$.

1. Soit X une variable aléatoire de loi géométrique de paramètre p. Montrer que $\mathbb{E}[X] = \frac{1}{p}$. *Indication:* On rappelle que pour $|x| < 1$, on a

$$\sum_{k \geq 1} k x^{k-1} = \left(\sum_{k \geq 0} x^k \right)' = \left(\frac{1}{1-x} \right)'.$$

   Pour la suite, on admettra que $\mathrm{Var}(X) = \frac{q}{p^2}$.

2. Proposer un estimateur $\hat{\theta}_n$ de $\theta = \frac{1}{p}$ à l'aide de la méthode des moments.

3. Proposer un intervalle de confiance asymptotique de niveau 0,95 à l'aide du théorème central limite.

4. 0n veut maintenant tester l'hypothèse nulle: $H_0 = \{\theta = 2\}$ contre l'hypothèse alternative: $H_1 = \{\theta \neq 2\}$. A l'aide de la variable aléatoire $Y_n = \sqrt{\frac{n}{2}} \left(\frac{1}{n} \sum_{i=1}^n X_i - 2 \right)$, construire un test de niveau asymptotique 0,95 pour θ.

5. Sur un échantillon de taille n, on trouve que $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i = 2,3$. Que peut-on conclure si la taille de l'échantillon est de 70? de 200?

6. Proposer un script Matlab permettant de vérifier que l'erreur de première espèce est bien de 0,05 et calculant la puissance du test lorsque $(\theta \neq 2)$ lorsque la taille de l'échantillon n est de 70. On considerera que la commande *geometrique(N,p)* a déjà été programmée et qu'elle renvoie N variables aléatoires indépendantes de loi géométrique de paramètre p.