Ecole Normale Supérieure
FIMFA – Statistique 2011-2012

Feuille d'exercices n°7.

Exercice 1. Dans le logiciel Scilab, l'instruction `rand(1,10,'uniform')` permet de générer $n = 10$ nombres pseudo-aléatoires de loi $\mathcal{U}[0, 1]$. Voici le résultat donné lors de l'appel de cette fonction :

0.2113249	0.7560439	0.0002211	0.3303271	0.6653811
0.6283918	0.8497452	0.6857310	0.8782165	0.0683740

Mettre en place un test de niveau 5% pour tester l'hypothèse H_0 : "la loi de cet échantillon est bien uniforme sur $[0, 1]$" contre H_1 : "il s'agit d'une autre loi".

A.N. : on donne $\mathbb{P}(D_{10} \leq 0.4092) \approx 0.95$, où D_n désigne la statistique de Kolmogorov-Smirnov associée à un échantillon de taille n.

Exercice 2. (Test de Cramér-Von-Mises)

1. Soit X une variable aléatoire réelle, F sa fonction de répartition et $F^{(-1)}$ son inverse généralisée :
$$F^{(-1)}(x) = \inf\{y \in \mathbb{R} : F(y) \geq x\}, \quad x \in [0, 1].$$
Démontrer les résultats classiques suivants :
   (a) $x \leq F(y) \Leftrightarrow F^{(-1)}(x) \leq y, \quad \forall x \in [0, 1], y \in \mathbb{R}$.
   (b) Si U est uniformément distribuée sur $[0, 1]$, alors $F^{(-1)}(U)$ a même loi que X.
   (c) Si F est continue sur $\mathbb{R}$, alors $F(X)$ est uniformément distribuée sur $[0, 1]$. Que se passe-t-il si F n'est pas continue ?
2. On considère un échantillon $(X_1, ..., X_n)$ de fonction de répartition continue F. On introduit la statistique suivante
$$nI_n = n \int_{\mathbb{R}} (F_n(x) - F(x))^2 \, \mathrm{d}F(x).$$
   (a) Montrer que
$$nI_n = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(X_{(i)}) \right)^2.$$
   (b) Expliquer comment construire un test d'adéquation à partir de la statistique I_n.

Exercice 3. On admet dans cet exercice que si $X_1, \ldots, X_n$ sont des variables aléatoires réelles i.i.d. de fonction de répartition F et fonction de répartition empirique F_n, on a, pour $t > 0$,

$$\mathbb{P}\Big(\|F - F_n\|_\infty > t\Big) \leq 2\exp(-2nt^2).$$

1. On considère des observations i.i.d. $X_1, \ldots, X_n$ distribuées suivant une loi $\mathcal{E}(\theta)$, $\theta > 0$, et on note F_θ leur fonction de répartition. Soit $h > -1$ et $\theta' = (1+h)\theta$.

   1.a Calculer soigneusement
   $$g(h) = \|F_\theta - F_{\theta'}\|_\infty$$
   en fonction de h.

   1.b Montrer que g est équivalente à $|h|/e$ lorsque $h \to 0$.

2. On appelle F_n la fonction de répartition empirique des X_i et on définit un estimateur $\hat{\theta}_n$ de θ par
   $$\|F_{\hat{\theta}_n} - F_n\|_\infty = \inf_\theta \|F_\theta - F_n\|_\infty.$$
   Soit θ_0 la vraie valeur du paramètre et $\mathbb{P}_{\theta_0}$ la probabilité correspondante. On pose
   $$\frac{\hat{\theta}_n}{\theta_0} = 1 + \hat{h}_n$$
   et on admet que g est une fonction strictement décroissante pour $h < 0$ et strictement croissante pour $h > 0$.

   2.a Montrer que si $g(\hat{h}_n) > t$, alors $\|F_{\theta_0} - F_n\|_\infty > t/2$.

   2.b Déduire de la question précédente une majoration explicite de
   $$\mathbb{P}_{\theta_0}(g(\hat{h}_n) > t).$$

   2.c Déduire alors des propriétés de g que la suite $\hat{\theta}_n$ est convergente en probabilité.

3. Donner finalement un intervalle de confiance de niveau $1 - \alpha$ pour θ_0 sous forme implicite. Donner la forme explicite approchée de cet intervalle lorsque n est grand.

Exercice 4. Soit $(X_1, ..., X_n)$ un échantillon de loi $\mathcal{E}(\theta)$, $\theta > 0$. Pour $k \in \mathbb{N}^*$, on pose

$$S_k = \sum_{i=1}^{k} X_i.$$

1. Déterminer la densité jointe du n-uplet $(S_1, ..., S_n)$. En déduire que le vecteur
   $$T_{1,n} = \left(\frac{S_1}{S_n}, ..., \frac{S_{n-1}}{S_n}\right)$$
   suit la loi de la statistique d'ordre d'un $(n-1)$-échantillon de loi uniforme sur $[0, 1]$.

2. Déterminer la loi jointe de la statistique d'ordre $(X_{(1)}, ..., X_{(n)})$ puis vérifier que

$$\left(nX_{(1)}, (n-1)\left(X_{(2)} - X_{(1)}\right), (n-2)\left(X_{(3)} - X_{(2)}\right), ..., \left(X_{(n)} - X_{(n-1)}\right)\right)$$

   est un n-uplet (i.i.d) de loi exponentielle $\mathcal{E}(\theta)$.

3. En déduire la loi de la statistique libre

$$T_{2,n} = \left(\frac{nX_{(1)}}{S_n}, .., \frac{X_{(1)} + .. + X_{(k-1)} + (n-k+1)X_{(k)}}{S_n}, .., \frac{X_{(1)} + .. + X_{(n-2)} + 2X_{(n-1)}}{S_n}\right).$$

4. Proposer deux suites de tests non-paramètriques de niveau asymptotique α, fondées respectivement sur $T_{1,n}$ et $T_{2,n}$, permettant de décider s'il est raisonnable de penser que $H_0 : \mathbb{P}_X \in \{\mathcal{E}(\theta), \theta > 0\}$ est vraie ou s'il faut rejeter cette hypothèse.

Exercice 5. On a relevé dans deux forêts les hauteurs en mètres de 12 et 14 arbres respectivement. On obtient les résultats suivants :

Forêt 1 :	23.4	24.4	24.6	24.9	25.0	26.2	26.3	26.8	26.9	27.0	27.6	27.7		
Forêt 2 :	22.5	22.9	23.7	24	24.4	24.5	25.3	26.0	26.2	26.4	26.7	26.9	27.4	28.5

Tester l'homogénéité des deux forêts.

Exercice 6. Dans une population de 500 personnes (300 hommes et 200 femmes), on a mesuré la tension artérielle de chaque individu, ce qui a donné les effectifs suivants :

Effec. observés	Hypertension	Tension normale	Hypotension
Hommes	72	192	36
Femmes	38	118	44

A-t-on, au risque de 5%, une liaison entre le sexe de l'individu et sa tension artérielle ?