TISD M1 Pro - Partiel du 9 novembre 2007

Durée : 2 heures
Documents autorisés : sujets de TD, notes de cours manuscriptes
Documents NON autorisés : corrigés de DM
La calculatrice est autorisée
Vous donnerez les résultats numériques avec le nombre de décimales qui vous semblera raisonnable.

On rappelle que pour la loi normale $\mathcal{N}(0,1)$, les quantiles à 95% et 97.5% sont $q_{0.95} = 1.64$ et $q_{0.975} = 1.96$. Pour les lois du χ^2 à 2, 3, 6 degrés de liberté, les quantiles à 95% sont respectivement $q_{0.95}(2) = 5.99$, $q_{0.95}(3) = 7.81$, $q_{0.95}(6) = 12.59$.

Exercice 1 (Ski)
Sur un échantillon de 20 sauteurs à ski, de 20 slalomeurs et de 20 descendeurs on note que respectivement 0, 5 et 10 athlètes présentent un surpoids.

1. Etablir le tableau de contingence du croisement des variables "discipline" et "présence d'un surpoids" (on fera des cases assez grandes...).

2. Etablir les distributions marginales.

3. Etablir les distributions conditionnelles des disciplines sachant que le sportif est/n'est pas en surpoids. Commentaires ?

4. Faire un test d'indépendance du χ^2. Conclure.

Exercice 2 (Taux de non scolarisation)
Sur le site de l'Unesco `http://stats.uis.unesco.org/`, il est possible d'obtenir, par pays, le taux d'enfants non scolarisés en âge d'aller à l'école primaire. A l'Annexe 1, nous donnons quelques statistiques pour cette variable par groupe de pays : Etats Arabes (groupe 1), Europe Centrale et Europe de l'Est (groupe 2), Asie Centrale (groupe 3), Est asiatique et Pacifique (groupe 4), Amérique du Sud et Caraïbes (groupe 5), Amérique du Nord et Europe de l'Ouest (groupe 6), Asie de l'Ouest et du Sud (groupe 7), Afrique du Sud (groupe 8).

1. Calculer le taux moyen de non scolarisation des enfants du primaire.

2. Calculer la moyenne et la variance du taux de non scolarisation pour la réunion des groupes 4 et 7 (constitué des pays asiatiques bordés par les Océans Indiens et Pacifiques).

3. Comparer les niveaux de scolarisation de l'Est asiatique-Pacifique (groupe 3) avec celui de l'Amérique du Sud-Caraïbes (groupe 4).

Exercice 3 (Taches solaires)
Les astronomes ont remarqué très tôt (depuis au moins 2000 ans) des taches sombres sur la surface du soleil (attention, il ne faut jamais regarder directement le soleil sans lunette adaptée !). Elles correspondent à des zones moins chaudes et sont des manifestations de l'activité solaire, dues à des variations de champ magnétique à la surface de celui-ci. Une tache peut avoir une durée de vie de quelques semaines.

Nous disposons d'un relevé du nombre mensuel moyen de taches solaires entre 1749 et 1983, soit 2820 observations (les données entre 1749 et 1960 ont été collectées par l'Observatoire Fédéral Suisse de Zurich, et celles après 1960, par l'Observatoire Astronomique de Tokyo). L'échantillon est noté $(x_t)_{t\in[\![1,T]\!]}$, où $T = 2820$. Nous considérerons dans cet exercice qu'il s'agit de réalisations indépendantes de variables aléatoires de même loi, et nous noterons X une variable aléatoire ayant cette loi.

1. Au vu de la Figure 1 (a), pourquoi l'hypothèse d'indépendance des observations est-elle criticable ? Quelle caractéristique des données ne prend-on pas en compte dans cet exercice ?

(a) (b) (c)

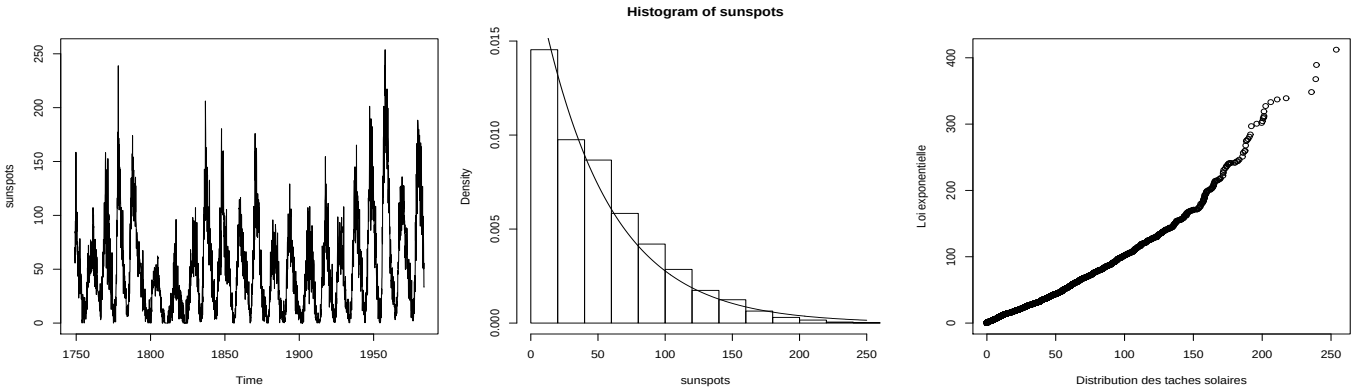

FIG. 1 – *(a) Variation du nombre mensuel moyen de taches solaires entre 1749 et 1983 ; (b) histogramme de la variable "nombre mensuel moyen de taches solaires" ; (c) graphique quantile-quantile de cette variable et d'une loi exponentielle.*

2. En utilisant les sorties de la `proc univariate` de SAS en Annexe 2, donner :
2.1. La moyenne, la médiane, le coefficient d'asymétrie. Commenter.
2.2. Tracer la boîte à moustaches (Boxplot) en indiquant bien les valeurs importantes.

3. L'histogramme des observations est présenté à la Figure 1 (b). Au vu des observations, est-il raisonnable de choisir pour la loi de X l'une des lois paramétriques vues en cours ? Laquelle ? Justifier brièvement.

4. Nous avons tracé à la Figure 1 (c) le graphique quantile-quantile (QQ-plot) des observations et d'une loi exponentielle. Qu'en pensez-vous ?

5. Nous supposons que X suit une loi exponentielle de paramètre $\lambda > 0$.
5.1. Ecrire le modèle statistique. Ce modèle est-il exponentiel ?
5.2. Déterminer l'estimateur du maximum de vraisemblance $\widehat{\lambda}_T$ de λ.
5.3. Cet estimateur $\widehat{\lambda}_T$ est-il convergent lorsque $T \to +\infty$ (préciser en quel sens et justifier la réponse) ?
5.4. Donner la loi asymptotique de $\sqrt{T}(\widehat{\lambda}_T - \lambda)$ lorsque $T \to +\infty$.
5.5. Utiliser ce résultat pour construire un intervalle, fonction de $\widehat{\lambda}_T$ contenant le vrai paramètre λ avec probabilité 95%. Faire l'application numérique.

6. Donner une approximation de la probabilité que le nombre de taches solaires observé dépasse 300 (valeur non observée sur les données pour lesquelles le maximum est 253.8). On obtient ainsi une estimation d'un événement très rare...

Exercice 4 (Répartition salariale sur des données groupées)
Dans une entreprise, les salaires sont les suivants :

Classe de salaire	Salaires mensuels	Nombre de salariés
1	$[500, 1500[$	50
2	$[1500, 2500[$	125
3	$[2500, 5500[$	25

1. Pour chaque classe $i \in \{1, 2, 3\}$ de salaires (notée $[x_{i-1}, x_i[$ et d'effectif n_i), calculer la fréquence empirique f_i, l'amplitude a_i, le centre $c_i = (x_{i-1} + x_i)/2$, la fréquence empirique cumulée F_i (proportion des salaires inférieurs à x_i), la masse salariale approchée $n_i c_i$ et la masse salariale cumulée approchée m_i (approximation de la somme de tous les salaires inférieurs à x_i). On présentera les résultats dans un tableau.

2. Tracer l'histogramme de la variable "salaire". Quelle règle faut-il respecter ?

3. Dessiner la fonction de répartition. Comme la variable de salaire est quantitative continue, on choisira ici la version continue de la fonction de répartition empirique, obtenue par interpolation linéaire des points (x_i, F_i).

4. Quelle est l'équation de la portion de droite représentant la fonction de répartition empirique sur l'intervalle de salaires $[1500, 2500[$? En déduire la médiane.

5. On s'intéresse à la répartition des salaires sur ces données agrégées.
5.1. Tracer la courbe de Lorenz. Pour des données groupées, cette courbe est obtenue en reliant les points

$$\left(F_i, \frac{\sum_{j \leq i} n_j c_j}{\sum_{j=1}^{3} n_j c_j} \right).$$

5.2. Calculer l'indice de Gini.
5.3. Commenter.