Master M2 SITN & Data Science, Université Claude Bernard Lyon 1
Logiciel SAS, Année 2018-2019

Examen du 19 octobre 2018
Durée : 1 heure 30 – Documents autorisés

NB : Chaque étudiant enregistre son programme dans un fichier nommé « nom_prénom.sas ». Ecrire en en-tête du programme en commentaire le nom et le prénom. A la fin de l'examen, le fichier sera enregistré sur la clé USB fournie. Ecrivez en commentaire où commence chaque exercice et la réponse à chaque question.

EXERCICE 1 (Etape data et statistiques descriptives)

1. Importer sous SAS le fichier *distance.csv* disponible à une des adresses suivantes : http://math.univ-lyon1.fr/homes-www/perrut/distance.csv http://math.univ-lyon1.fr/~gciuperca/enseign.html, rubrique « Fichiers de données », à l'aide de l'instruction INFILE. Ce fichier contient les distances en km entre le domicile et le lieu de travail de 160 personnes. Le lieu de résidence ainsi que l'âge des individus sont également indiqués.
2. Corriger la colonne *lieu* en éliminant l'espace qui s'est glissé par erreur dans le mot « Paris ». Effacer la ligne de l'individu avec un âge de 7 ans.
3. Créer une variable *categorie_age* qui prend la valeur 1 si l'âge est compris entre 20 et 40 ans (strictement inférieur) et 2 si l'âge est compris entre 40 et 60 ans.
4. A l'aide de la procédure appropriée, calculer la moyenne, l'écart-type, les percentiles 25% et 75% de la variable *distance* pour chacune des deux villes de la colonne *lieu*. Dans une table SAS nommée *stat*, récupérer ces 4 indicateurs, en les nommant respectivement *moy*, *s*, *Q1* et *Q3*. Ajouter à cette table la variable *IQR* qui est égale à Q3-Q1.
5. Créer une nouvelle table nommée *Tukey*, qui contient la table initiale *distance* et les colonnes de la table *stat*. Cette table contient donc 159 lignes et 7 colonnes (*lieu*, *age*, *distance*, *moy*, *s*, *Q1*, *Q3*).
6. D'après le test de Tukey, une valeur peut être considérée comme outlier si elle est strictement inférieure à Q1-1.5*IQR ou strictement supérieure à Q3+1.5*IQR. Ajouter une variable *outlier* à la table Tukey qui prend la valeur 0 si la valeur n'est pas un outlier, ou 1 si elle peut être considérée comme un outlier.

EXERCICE 2 (procédures)

1. A l'aide de la procédure appropriée calculer les fréquences (ou les nombres) d'individus en fonction de la catégorie d'âge et du lieu géographique (tableau à double entrée en fonction des deux variables *categorie_age* et *lieu*). Effectuer le test du Khi-deux indépendance des deux variables *catégorie_age* et *lieu*.

2. A l'aide de la procédure appropriée, réaliser un test de Student (t-test) pour comparer les moyennes de la variable distance en fonction du lieu (Lyon ou Paris), sans prendre en compte la valeur considérée comme un outlier. Récupérer l'intervalle de confiance de la différence des deux moyennes dans une table SAS. Modifier le risque alpha (à défaut fixé à 0.05) à 0.1.

3. Afin de visualiser l'effet du lieu, réaliser les boxplots de la variable *distance* en fonction de la variable *lieu*.

EXERCICE 3 (Macro)

Créer une macro nommée *boxplot*, qui prend en entrée 4 variables :

- *Table* : une table de données SAS
- *Rep* : une variable numérique présente dans la table *Table*
- *Mod* : une variable alphabétique présente dans la table *Table*
- *Ligne* : une valeur numérique

Cette macro doit tracer les boxplots de la variable *Rep* présente dans la table *Table* en fonction des modalités de la variable *Mod*, et qui ajoute une ligne horizontale à la valeur numérique présente dans la variable *Ligne*.
Tester cette macro sur la table *Tukey* de la manière suivante :
%boxplot (Table=Tukey, Rep=distance, Mod=Lieu, ligne=30);