Master M2 SITN, Université Claude Bernard, Lyon 1
Logiciel SAS, année 2013-2014

Examen du 19 Novembre 2013
Durée 3 heures: documents autorisés

NB: Chaque étudiant enregistre son programme dans un fichier sauvé sous le nom: "nom_prénom.sas"
Ne pas oublier d'écrire en tête de 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.

Le fichier de données provient de l'adresse internet

www.umass.edu/statdata/statdata/stat-survival.html

rubrique "Carcinoma of the Oropharynx", où vous avez à la fois la description des données et le fichier de données "pharynx.dat".
Vous avez également la description des données sur la dernière page du présent sujet.

EXERCICE 1

Le fichier contient des données concernant des essais cliniques sur des patients souffrant de cancer de pharynx. Les variables du fichier sont, dans ordre: *case, inst, sex, tx, grade, age, cond, site, t_stage, n_stage, entry_dt, status, time* . Les variables *case, age, time* sont de type numérique. La variable *entry_dt* est une date, mais vous la déclarez de type numérique. Les variables *inst, sex, tx, grade, cond, site, t_stage, n_stage, status* sont de type caractère(catégoriel).

1) Enregistrez le fichier sur votre ordinateur, en gardant le même nom pour le fichier. Ensuite, créez un tableau SAS, appelé *pharynx*, à partir du fichier *pharynx.dat*, contenant toutes les observations.
2) Affichez le tableau SAS. Quel est le nombre d'observations du fichier?
3) Réalisez une analyse descriptive pour les variables *inst, sex, tx, grade, cond, site, t_stage, n_stage, status*. Réalisez un test de χ^2 pour tester si chacune de ses variables suit une loi uniforme discrète? Il y a-t-il des variables qui suivent une loi uniforme discrète?
4) Réalisez une analyse descriptive en croisant les variables *t_stage* et *n_stage*. Réalisez un test d'indépendance. Ces deux variables sont-elles indépendantes?
5) Réaliser une analyse descriptive univariée (moyenne, variance, écart-type, min, max, ...) pour chacune des variables *age, time*.

6) Calculez la corrélation entre les variables *age, time*.
7) Testez si la variable aléatoire *age* suit une loi Normale.
8) Appuyez la conclusion de la question précédente, en réalisant l'histogramme et le box-plot de la variable étudiée.
9) Tracez le nuage de points de la variable *time* fonction de la variable *age*, les points du graphique prenant les valeurs correspondantes de la variable *sex*.
10) Créez une nouvelle variable (qualitative), ses valeurs dépendant des valeurs possibles du couple de variables *(tx, status)*. Le nom de la nouvelle variable et les noms de ses modalités sont à votre choix.
11) A partir du tableau créé à la question 10), créez deux nouveaux tableaux SAS, appelés *numeriques* et *caracteres*. Le tableau *numeriques* contient les variables *age, time*. Le tableau *caracteres* contient toutes les variables de type caractère. A partir de ces deux nouveaux tableaux créer des **fichiers texte externes**, avec le même nom.

EXERCICE 2 (à utiliser PROC IML)
On utilise le tableau *pharynx* créé à l'Exercice 1, question 1).

1) Lire, dans une matrice, qu'on va noter X, toutes les observations et les variables *age, time* du tableau SAS *pharynx*.
2) On s'est rendu compte que le patient numéro 93 n'a pas 67 ans mais 57 ans. Faites le changement dans les données. Vérifiez que vous avez fait le changement demandé.
3) On veut transformer la variable *age* dans une variable catégorielle. Créez une variable V qui aura les valeurs:

$$\begin{gathered} 1 \text{ si } age \leq 25; \\ 2 \text{ si } age \in]25, 50[; \\ 3 \text{ si } age \in [50, 75]; \\ 4 \text{ si } age > 75. \end{gathered}$$

4) Créer une nouvelle matrice, notée Y, en concaténant la colonne de la matrice X qui contient les valeurs de l'*age* et le vecteur V. Vérifiez (par affichage) que vous avez fait la bonne transformation à la question précédente.
5) Créer un tableau SAS appelé Y qui va contenir toutes les variables de la matrice Y.

EXERCICE 3
1) A partir du tableau de données crée à l'Exercice 1, question 10) et du tableau Y crée à l'Exercice 2, question 5), créez un nouveau tableau, appelé *final* avec toutes les variables, la variable *age* contenant la valeur correcte pour l'observation 93. Il faut qu'à partir du tableau Y vous obtenez des variables avec le nom *age* et *tr*.
2) On s'est rendu compte que pour le patient numéro 125 les données sont incorrectes. En conséquence, on va enlever de la base de données cette observation. Créez un nouveaux

tableau appelé *correct*.

EXERCICE 4 (à utiliser PROC GPLOT)
On utilise le tableau créé à l'Exercice 3, question 2).
Tracer le graphique de la variable *time* en abscisse, fonction de la variable *age* en ordonnée. Les points de représentation sont "*" et la courbe, obtenue par interpolation, sera de couleur bleu. Donner un titre à ce graphique et pour les axes, spécifier les noms des variables (le nom de l'*age* sera en rouge et le nom du *time* sera en en bleu).

EXERCICE 5 (à utiliser les macros SAS)

1) Créer une macro-variable *m_tab* associée au nom du tableau *correct*. Créer également les macro-variables:
- *m_age* correspondant à la variable *age*.
- *m_time* correspondant à la variable *time*.
- *m_status* correspondant à la variable *status*.
- *m_stage* correspondant à la variable *t_stage*.

2) Programmer une macro paramétrée qui utilise le tableau *&m_tab* pour réaliser la corrélation entre deux variables numériques. Application pour les variables *age* et *time*.
3) Programmer une macro paramétrée qui utilise le tableau *&m_tab* pour réaliser la corrélation entre deux variables qualitatives. Application pour les variables *status* et *t_stage*.