



département
Mathématiques

Master MAS, parcours M2 SMSD, Université Claude Bernard Lyon 1

Remise à niveau en Statistique - Logiciel SAS,

Année 2024-2025

Examen du 18 octobre 2024
Durée 2 heures: documents autorisés

*NB: Chaque étudiant enregistre son programme dans un fichier sauvé sous le nom: "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 envoyé par e-mail à l'enseignant surveillant de votre salle (ne quittez pas la salle tant que l'enseignant ne vous a pas confirmé la réception de votre mail avec le programme SAS).** Ecrivez en commentaire où commence chaque exercice et la réponse à chaque question.*

Salle	Enseignant	Adresse mail de l'enseignant
Mathesis	Gilles COHEN	gilles.cohen@univ-lyon1.fr
TDmath	Gabriela CIUPERCA	Gabriela.Ciuperca@univ-lyon1.fr

EXERCICE 1

Le fichier de données, appelé *heart_failure_clinical_records_dataset.csv*, peut être téléchargé à l'adresse suivante:

https://archive.ics.uci.edu/ml/machine-learning-databases/00519/heart_failure_clinical_records_dataset.csv

La description des données se trouve après le présent sujet. Les variables sont dans l'ordre donné dans cette description.

- 1) Enregistrez le fichier sur votre ordinateur, en gardant le même nom pour le fichier. Dans le fichier, à quelle observation commencent les données? Comment (par quel caractère) les valeurs des variables sont séparées?
Note: répondre à cette question dans votre programme .SAS, comme un commentaire
- 2) Créez un tableau SAS, appelé **heart_failure**, à partir du fichier *heart_failure_clinical_records_dataset.csv*, contenant toutes les observations et toutes les variables. Quel est le nombre total d'observations et de variables dans le tableau **heart_failure**?
- 3) Affichez les 5 premières observations du tableau SAS **heart_failure**.
- 4) Divisez le tableau **heart_failure** en deux nouveaux tableaux :
 - **hf_age** : contient toutes les observations pour lesquelles l'âge du patient est connu (âge supérieur à 0).
 - **hf_no_age** : contient toutes les observations pour lesquelles l'âge du patient est inconnu (âge manquant ou égal à 0).

Ces deux nouveaux tableaux ne doivent plus contenir la variable **age**. Spécifiez pour chacun des deux tableaux le nombre d'observations et de variables.

- 5) Pour les observations du tableau **hf_age**, effectuez une analyse descriptive univariée sur les variables

`serum_creatinine`, `ejection_fraction`, et `platelets`. Affichez les statistiques descriptives (moyenne, médiane, minimum, maximum, écart-type).

6) Pour les observations du tableau `hf_age`, testez si les variables `serum_creatinine` et `ejection_fraction` sont corrélées. Utilisez une procédure adéquate pour calculer le coefficient de corrélation de Pearson.

7) Pour les observations du tableau `hf_age`, réalisez un test de normalité pour les variables `serum_creatinine` et `ejection_fraction`. Affichez les histogrammes de ces deux variables avec la densité de la loi normale ajustée aux données.

8) À partir du tableau `hf_age`, créez deux nouvelles variables :

- `log_creatinine` : le logarithme de la variable `serum_creatinine`.
- `log_ejection_fraction` : le logarithme de la variable `ejection_fraction`.

9) Quelle est la moyenne des variables `log_creatinine` et `log_ejection_fraction` ?

10) Créez une nouvelle variable appelée `ejection_category` dans le tableau `hf_age`, définie de la manière suivante en fonction de la variable `ejection_fraction` :

- `faible` : si `ejection_fraction` < 25.
- `moyen` : si `ejection_fraction` >= 25 et `ejection_fraction` < 50.
- `élevé` : si `ejection_fraction` >= 50.

Réalisez une analyse descriptive pour la variable `ejection_category`. Testez si les trois proportions de la variable `ejection_category` sont égales.

EXERCICE 2

Pour les observations du tableau `hf_age`, utilisez la procédure `TRANSPOSE` pour transformer les variables `serum_creatinine`, `ejection_fraction`, et `platelets`, afin qu'elles deviennent des lignes regroupées par la variable `DEATH_EVENT`. La nouvelle table sera appelée `hf_transposed`. Affichez la table `hf_transposed`.

EXERCICE 3

1) Pour les observations du tableau `heart_failure` réalisez un test de Student pour comparer la moyenne de la variable `ejection_fraction` entre les patients décédés (`DEATH_EVENT` = 1) et ceux qui ne le sont pas (`DEATH_EVENT` = 0).

2) Pour les observations du tableau `heart_failure`, représenter les boxplot de la variable `ejection_fraction` en fonction de la variable `DEATH_EVENT` et relier les moyennes des deux boxplots par un trait.

3) Pour les observations du tableau `heart_failure` calculez la moyenne des variables `serum_creatinine` et `ejection_fraction` par rapport à la variable `time`. Sauvegardez les deux moyennes dans un tableau `heart_failure2` où les deux variables qui contiendront la moyenne portent le nom `serum2` et `ejection2`.

4) On utilise le tableau `heart_failure2`. Tracez sur le même graphique les variables `serum2` et `ejection2` en ordonnée et `time` en abscisse. Les points de représentations doivent être "o", en rouge, pour `ejection2` et "+", en bleu, pour `serum2`. Les points du graphique doivent être reliés par interpolation.

EXERCICE 4

1) Écrivez un macro-programme nommé `stat_description` qui prend en entrée une table de données et deux variables notées V1 et V2. La variable V1 est numérique. Le macro-programme doit calculer la moyenne pour la variable V1 par rapport aux valeurs de la variable V2. Dans la PROC MEANS utilisée dans le macro-programme, il faut également créer un tableau nommé `sortie` en sortie qui contiendra dans la variable `moy` la moyenne de la variable V1 par rapport aux valeurs de la variable V2.

Testez cette macro sur la table `heart_failure` et les variables `V1=creatinine_phosphokinase`, `V2=age`.
2) Créez un autre macro-programme nommé `plot_evol` qui prend en entrée un tableau de données et deux variables `V3` et `V4`. Dans ce macro-programme il faut tracer le graphique de la variable `V3` par rapport à la variable `V4`. Les points du graphique doivent être reliés et de couleurs bleue.
Testez ce macro-programme sur la table `sortie` et les variables `V3=moy`, `V4=age`.

Le jeu de données **Heart Failure Prediction** contient les variables suivantes :

- `age`: âge du patient
- `anaemia`: 1 si le patient a de l'anémie, 0 sinon
- `creatinine_phosphokinase`: niveau de CPK dans le sang (indicateur de dommages au muscle cardiaque)
- `diabetes`: 1 si le patient est diabétique, 0 sinon
- `ejection_fraction`: pourcentage de sang pompé par le coeur à chaque battement
- `high_blood_pressure`: 1 si le patient a de l'hypertension, 0 sinon
- `platelets`: niveau de plaquettes dans le sang
- `serum_creatinine`: niveau de créatinine dans le sang
- `serum_sodium`: niveau de sodium dans le sang
- `sex`: 1 pour homme, 0 pour femme
- `smoking`: 1 si le patient fume, 0 sinon
- `time`: temps de suivi en jours
- `DEATH_EVENT`: 1 si le patient est décédé lors du suivi, 0 sinon