Master MAS, parcours M2 SMSD, Université Claude Bernard Lyon 1
Remise à niveau en Statistique-Logiciel SAS
Année 2022-2023

Examen du 21 octobre 2022
Durée : 2 heures – 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 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 fichiers de données « *thermometry.csv* » se trouve à l'adresse *:*

https://www.openintro.org/data/csv/thermometry.csv

Les données de ce fichier proviennent du site *http://jse.amstat.org/datasets* et représentent la température corporelle (en degrés Fahrenheit), le sexe (1 = homme, 2= femme) et la fréquence cardiaque (en battements par minute) d'un groupe de 130 individus. Elles proviennent d'une étude visant à déterminer si la température moyenne réelle du corps est de 98.6 degrés Fahrenheit.

1) Créer la table SAS qu'on va appeler *tempnorm* à partir du fichier de données *thermometry.csv* contenant toutes les observations. Les variables de cette table vont s'appeler *tempcorps*, *sexe*, *freqcard.*

2) Tracer l'histogramme ainsi que la densité de la loi Normale s'ajustant au mieux aux données pour la variable *tempcors*. Tester sa normalité. Par défaut l'axe des y est en pourcentage, le mettre en nombre, libeller l'axe en conséquence et rajouter le titre suivant : « Histogramme des données de température corporelle ».

3) La température corporelle moyenne d'une personne en bonne santé est de 98.6 F. En tenant compte du résultat pour le test de normalité de la question précédente, vérifiez si la température moyenne est égale à 98.6F. Pour ceci, réalisez un test de signification bilatéral avec un niveau de signification α = 0,05 (H_0 : μ *tempcorps* = 98.6 H_a : μ *tempcorps* $\neq$ 98.6) .

4) Trier la table *tempnorm* en ordre croissant par rapport aux valeurs de la variable *tempcorps*. Afficher la nouvelle table *tempnorm.*

5) Calculer la moyenne et son l'intervalle de confiance à 95 % (α = 0.05) pour la variable *tempcorps*.

Vérifiez si la valeur de μ=98.6 spécifiée dans H_0 s'y trouve. Si c'est le cas, cela signifie que la p-valeur est supérieure à 0.05 et que nous ne pouvons pas rejeter H_0.

6) A l'aide de la procédure appropriée, calculer le minimum, la moyenne, l'écart-type, les quartiles (Q1, Q2, Q3) et le maximum de la variable *tempnorm*. Dans une table SAS nommée *stat*, récupérer ces 7 indicateurs, en les nommant respectivement *the_min*, *the_mean, the_sd, the_quart1, the_median, the_quart3, the_max*. Supprimer dans la table *stat* les variables *_TYPE_* et *_FREQ_* générées pas la procédure.

7) Créer une table *tempnorm_vm* en exécutant le code ci-dessous afin de rendre aléatoirement certaines valeurs des variable *tempcors* et *freqcard* manquantes.

```
data work.tempnorm_vm;
   set work.tempnorm;
   if rand('Normal', 0.5) < 0.01
   then tempcorps = .;
   if rand('Normal', 0.5) < 0.01
   then freqcard = .;
run;
```

Puis à l'aide de la procédure appropriée calculer la valeur maximum de la variable *tempcorps* selon le sexe de la table contenant des valeurs manquantes *tempnorm_vm*. Dans une table SAS nommée *stat_vm*, récupérer cet indicateur, le nommer *maxx*.

8) Remplacer, selon les valeurs de la variable *sexe*, dans la table *tempnorm_vm* les valeurs manquantes de la variable *tempcorps* par la valeur du maximum dans la table *stat_vm*.

9) Dans la table *tempnorm* ajouter une variable *tempcorps_quant* qui correspond à la discrétisation de la variable numérique *tempcorps* en quatre modalités selon les quartiles. Si *tempcorps* ≤ Q1 *tempcorps_quant=mod1* sinon si ….. Les valeurs des quartiles sont à récupérer depuis la table *stat* au travers de macro variables.

10) Réaliser la statistique descriptive sur les variables *sexe* et *tempcorps_quant.* Réalisez un test d'hypothèse de Chi2 d'indépendance.

EXERCICE 2 (Graphique)

On utilise la table *tempnorm* de l'Exercice 1.

1) *R*eprésenter le boxplot de la variable *tempcorps* en fonction de la variable *sexe* et relier les moyennes des deux boxplots par un trait.
2) Tracer sur un graphique le nuage de points de la variable *tempcorps* en fonction de la variable *freqcard* et selon la variable *sexe*. Les points correspondants à la modalité male de la variable *sexe* doivent être représentée par des symboles en forme de croix (+) et de couleur rouge et ceux correspondants à la modalité *female* par des symboles en forme de rond et de couleur noire. Faire afficher la légende.

EXERCICE 3 (Macros)

1) A partir de la table *stat* crée à la question 6) de l'Exercice 1, créer deux macros variables *moyenne* et *ecart_type* et leur donner la valeur récupérée dans la table *stat*. Afficher le continu des variables *moyenne* et *ecart_type.*

   Puis à partir de la table *tempnorm,* créer une nouvelle table *outliers* ne contenant que les observations pour lesquelles la variable *tempcorps* est en dehors de l'intervalle : « moyenne ± 3 * ecarts_type ». Affichez la table *outliers.* Donner le nombre d'observations de la table *outliers.*

2) Créer un macro-programme nommé ***descriptive*** qui prend en entrée cinq paramètres :

   - *Tab* : une table de données SAS
   - v1 : une variable numérique présente dans la table *Table*
   - *v2* : une variable numérique présente dans la table *Table*
   - v3 : une variable qualitative présente dans la table *Table*
   - *v4* : une variable qualitative présente dans la table *Table*

Ce macro-programme doit d'abord faire de la statistique descriptive sur les variables *v1* et *v2.* Ensuite il faut faire de la statistique descriptive sue les variables *v3* et *v4.*

Tester cette macro sur la table *tempnorm* de l'Exercice 1 et les variables *tempcorps, freqcard, sexe,tempcorps_qual .*

3) ***Question bonus.***

Créer un macro-programme nommé *traitement_vm* qui prend en entrée une variable :

y : une variable numérique présente dans la table à traiter

Ce macro-programme remplace les valeurs manquantes de la variable numérique *y* par la première valeur non absente de la variable y.

Tester cette macro sur la table *tempnorm_vm* avec la variable *tempcorps.* La macro est appelée à l'intérieur d'une étape *Data* et le résultat est mis dans une table *tempnorm_no_vm*.