



département
Mathématiques

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

Remise à niveau en Statistique - Logiciel SAS,

Année 2025-2026

Examen du 17 octobre 2025
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 votre nom et le prénom. **A la fin de l'examen, le fichier sera envoyé par e-mail à gabriela.ciuperca@univ-lyon1.fr (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.

EXERCICE 1 (9.5 points)

- 1) Importer sous SAS le fichier *distance.csv* disponible à l'adresse suivante :
<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. (1.5 points)
- 2) Corriger la colonne *lieu* en éliminant l'espace qui s'est glissé par erreur dans le mot "Paris". Effacer la ligne de la personne avec un âge de 7 ans. (1.5 points)
- 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. (1 point)
- 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*. (2 points)
- 5) Créer à partir de la table *stat* une nouvelle table *stat1* qui contiendra en plus une nouvelle variable *IQR* qui est égale à $Q3 - Q1$. (1 point)
- 6) Créer par rapport au lieu une nouvelle table nommée *Tukey*, qui contient la table initiale *distance* et les colonnes de la table *stat1*. Cette table contient donc 159 lignes et 7 colonnes (*lieu*, *age*, *distance*, *moy*, *s*, *Q1*, *Q3*). (1 point)
- 7) D'après le test de Tukey, une valeur peut être considérée comme outlier si elle est strictement inférieure à $Q1 - 1.5 \cdot IQR$ ou strictement supérieure à $Q3 + 1.5 \cdot IQR$. Ajouter à la table *Tukey* une variable appelée *outlier* qui prend la valeur 0 si la valeur n'est pas un outlier, ou 1 si elle peut être considérée comme un outlier. (1.5 points)

EXERCICE 2 (7 points)

Pour les quatre questions de cet exercice, on utilise la table *Tukey*.

- 1) A l'aide d'une 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*). Tester si les deux variables *categorie_age* et *lieu* sont indépendantes. (1.5 points)
- 2) A l'aide de la procédure GLM, 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 les sommes des carrés du tableau d'analyse de variance dans une table SAS, appelée *results*. Afficher la table *results*. (2 points)

3) Afin de visualiser l'effet du lieu, réaliser les boxplots de la variable *distance* en fonction de la variable *lieu*. (1 point)

4) Tracer la variable *distance* en ordonnée et *age* en abscisse. Les points de représentations doivent être "o", en rouge. Les points du graphique doivent être reliés par interpolation. Mettre une légende au graphique. (2.5 points)

EXERCICE 3 (3.5 points)

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

- *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*, la variable numérique *distance*, la variable alphabétique *lieu*, et la valeur numérique 30.