

Traitement statistique de l'information

Examen

Consignes :

- L'épreuve est individuelle et dure 1h30.
- Le travail doit être réalisé à l'aide du langage R.
- Vous avez le droit de consulter uniquement les slides du cours sur madoc, vos notes personnelles de cours et de TP (numériques ou papier), et une fiche de révision A4 recto-verso manuscrite. L'accès à internet est interdit, mis à part madoc et la documentation officielle de R sur rdocumentation.org.

Vous pouvez rédiger votre travail dans un script R (fichier texte), dans un fichier .Rmd de RStudio, ou dans un traitement de texte.

A chaque question, vous devez fournir dans votre réponse :

- votre **code R** (vous pouvez insérer des commentaires avec #)
- votre **interprétation/commentaire** sur le résultat renvoyé par R et/ou la **réponse** à la question du sujet d'examen.

Ces informations sont indispensables pour engranger des points. Si vous rédigez dans un traitement de texte, vous pouvez aussi inclure les résultats textuels ou graphiques renvoyés par R, uniquement pour améliorer la lisibilité de votre document.

Si vous écrivez des fonctions, n'oubliez pas de fournir leur code.

Les questions qui comportent le verbe "calculer" indiquent que vous devez mener vous-même les calculs avec R, et non utiliser une commande qui les fera pour vous. Fournissez tout le détail des calculs.

A la fin de l'épreuve, vous devez déposer votre travail sur madoc dans le dépôt de la section "EXAMEN". **Pensez à préparer votre fichier avant la fermeture du dépôt.**

1 Exercice 1

Le fichier `vins.data` décrit un échantillon de vins issus d'une région viticole européenne selon différents composants : les acides, les chlorides, le sucre, les sulphates, et le degré d'alcool. Les quatre premières variables sont des masses (en g) relevées dans 1 L de vin, tandis que le degré d'alcool est un pourcentage. Chargez le jeu de données dans un dataframe, en prenant en compte l'en-tête.

1.1 Analyse 1D

1. Combien de vins sont présents dans le jeu de données ?
2. Visualisez les distributions des variables à l'aide des représentations de votre choix et répondez aux questions suivantes :

- Quel composant est présent dans le vin en beaucoup plus grandes quantités que les autres composants ?
 - Quelles variables présentent des valeurs extrêmes du côté des petites valeurs ?
3. Représentez la densité continue estimée de la variable **sulphates**. Qualifieriez-vous cette distribution de multimodale ? Justifiez.
 4. A l'aide d'un graphique, vérifiez la normalité de la variable **chlorides**.
 5. Donnez la moyenne et la variance non corrigée de la variable **sucre**, sans utiliser une fonction écrite par vous.
 6. Combien de vins ont un degré d'**alcool** supérieur à 13% ?
 7. Calculez un intervalle de confiance sur la quantité moyenne de **sulphates** dans la population des vins. Utilisez le niveau de confiance de 90%. Vérifiez votre résultat à l'aide d'une commande R.
 8. Est-ce que les données remettent en cause l'hypothèse H_0 d'une quantité moyenne de **sucre** dans la population égale à 15g ? Pour répondre à cette question, calculez la p-valeur, puis concluez le test. Choisissez un seuil de risque de 1%.

1.2 Analyse 2D

1. Quelles variables sont les plus corrélées ? (au sens de la corrélation de Pearson) Répondez d'abord à l'aide d'un graphique, puis à l'aide de valeurs numériques.
2. La corrélation entre les variables **sucre** et **sulphates** est-elle significative ?
3. Calculez la valeur observée de la statistique de test dans le test sur la corrélation entre **acidite** et **chlorides**.
4. Donnez une estimation non biaisée de la corrélation entre **sucre** et **acidite** dans la population

1.3 Un test sur la médiane

Dans la partie 1.1, un test portait sur la quantité moyenne de **sucre**. Etant donné que cette variable a une distribution asymétrique, il vaut mieux utiliser la médiane. Vous allez donc tester l'hypothèse H_0 : "quantité médiane de **sucre** = 15g dans la population". Il s'agit d'un test avec les caractéristiques suivantes :

- la statistique de test est T = "nombre de valeurs observées supérieures à 15"
- si H_0 est vraie, alors T suit la loi normale de moyenne $\frac{n}{2}$ et d'écart-type $\sqrt{\frac{n}{4}}$, où n est la taille de l'échantillon
- le test est bilatéral

Faites le test au seuil de risque 1% en calculant la région critique puis la p-valeur. Concluez.

2 Exercice 2

Le fichier **films.data** est un échantillon de films où les acteurs Ben Affleck, Bradley Cooper, Matt Damon, Johnny Depp ou Liam Neeson tiennent un rôle central. Chargez le jeu de données dans un dataframe, en prenant en compte l'en-tête.

2.1 Analyse 1D

1. Combien de films comporte le jeu de données ?
2. Combien de réalisateurs (**director**) différents sont présents dans le jeu de données ?

3. Visualisez la distribution des catégories de films (**genre**).
4. Les données remettent-elles en cause l'hypothèse H_0 d'une proportion de comédies égale à 30% dans la population ?

2.2 Analyse 2D

On s'intéresse ici au lien entre la catégorie du film (**genre**) et son acteur vedette (**starring**).

1. Donnez la table de contingence qui croise les deux variables.
2. Question théorique : indiquez par une formule comment vous pouvez calculer la table de contingence attendue sous hypothèse d'indépendance.
3. A l'aide d'un graphique, trouvez pour chaque acteur son genre cinématographique de prédilection.
4. Le lien entre les deux variables est-il fort (=intense) ? Est-il significatif ?