

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](https://www.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 souligné "**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.

En général les questions sont indépendantes, donc si vous bloquez, passez à la suivante.

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 `chocolat.data` fournit les résultats de tests sensoriels portant sur différentes tablettes de chocolat vendues dans le commerce. Des consommateurs ont jugé sur une échelle¹ de 0 à 10 s'ils percevaient que le chocolat :

- avait une odeur de cacao ou de lait ;
- avait un goût amer, un goût de cacao ou de lait ;
- était fondant.

Chargez les données dans un dataframe et "attachez" les variables.

1.1 Variables numériques

1.1.1 Analyse unidimensionnelle

1. Visualisez la distribution de chaque variable à l'aide de la représentation graphique de votre choix. Commentez les distributions qui sortent du lot.
2. Affichez le dataframe réduit aux 5 chocolats jugés les plus fondants.
3. Calculez le coefficient de variation sur la variable de votre choix. Que mesure cet indicateur ? Quel est son intérêt dans le cas général ? dans le cas particulier des données sur le chocolat ?
4. En effectuant des tests, indiquez quelle variable est la plus compatible avec une distribution normale entre `Fondant` et `Amertume`.
5. On considère qu'un chocolat est amer si sa note est supérieure (strictement) à 5. Calculez la proportion de chocolats qui ont été jugés amers.
6. Les données remettent-elles en cause l'hypothèse qu'il y a autant de chocolats jugés amers et non amers dans la population ?
7. **Calculez** un intervalle de confiance (au niveau 90%) sur la note moyenne en `Fondant` dans la population. En déduire le résultat du test de l'hypothèse $H_0 : \mu_{Fondant} = 5$.

1.1.2 Analyse bidimensionnelle

1. Fournissez la matrice des corrélations. Commentez les corrélations les plus fortes.
2. Visualisez les distributions 2D et commentez. Identifiez un couple de variables qui semble proche de l'indépendance.
3. Le nuage de points (`Amertume`, `Lait`) présente une asymétrie intéressante. Énoncez cette asymétrie en une phrase.
4. A l'aide du test adéquat, indiquez si la corrélation entre `Fondant` et `OdeurCacao` est significative ou pas.
5. **Calculez** la p-valeur du test précédent.

1.2 Variables catégoriques

Créez deux variables `NiveauFondant` et `NiveauAmertume` en exécutant le script `variablesCateg.R` disponible sur madoc. Dans la suite, vous ne manipulez **que ces deux variables**.

1. Visualisez les distributions de chaque variable à l'aide d'une représentation graphique appropriée. Que constatez-vous ?
2. Visualisez le lien entre les deux variables. Quelle tendance générale constatez-vous ? Qu'est-ce qui va à l'encontre de cette tendance ?

1. 0 signifie que le caractère évalué n'est pas présent dans le chocolat, 10 signifie que le caractère est extrêmement présent.

3. Fournissez la table de contingence qui croise les deux variables, puis la table de contingence sous hypothèse d'indépendance.
4. **Calculez** la mesure du χ^2 sans exécuter de boucle.
5. Le lien entre les deux variables est-il fort ?
6. Faites le test du χ^2 en **calculant** la région critique au seuil de risque de 5%. Concluez.

2 Exercice 2

Le fichier `agro.data` décrit une promotion d'étudiants en première année d'école d'ingénieurs en agronomie dans la filière Agro-écologie. La variable `cursus` indique la formation d'origine des étudiants (PI désigne la prépa intégrée), tandis que la variable `note` indique la moyenne générale sur l'année.

Chargez les données dans un dataframe et "attachez" les variables. Si nécessaire, stockez des variables dans des facteurs.

2.1 Analyses 1D et 2D

1. Combien d'étudiants sont décrits dans le jeu de données ?
2. Visualisez la distribution de chaque variable à l'aide de la représentation graphique de votre choix. Commentez.
3. Donnez les quantiles d'ordre 30% et 70% de la variable `note`.
4. Donnez la variance non corrigée de la variable `note`.
5. Visualisez la distribution des notes selon le cursus, puis selon le genre. Commentez.
6. L'effet du cursus sur les notes est-il significatif? est-il fort ?
7. Mêmes questions sur l'effet du genre sur les notes.

2.2 Un test sur la médiane

Vous allez tester l'hypothèse H_0 : "note médiane = 12 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 12"
- 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.