Épreuve rendue - Fondements statistiques

M1 SDSC-SDIA

Arthur et Leonardo : Arthur Teisseire et Leonardo Nassabain

- Ce sujet comporte 2 pages.
- La rédaction de vos réponses est à rendre au format PDF sur Moodle au plus tard le 28/12/2020.
- Il est conseillé d'utiliser RMarkdown pour rédiger votre rapport (envoyez-moi un mail si vous avez des questions quant à son utilisation).
- Illustrez vos propos par des graphiques bien choisis ou des résultats numériques dès que cela est pertinent.
- Les fichier devront être nommés suivant la convention
  - `ArthuretLeonardo.pdf`

  ou
  - `ArthuretLeonardo1.pdf`
  - `ArthuretLeonardo2.pdf`

  si vous choisissez de rendre un document par exercice.
- Il appartient aux candidats de prendre leurs dispositions pour rendre le sujet dans le temps imparti. Aucun délai supplémentaire ne sera accordé.

Les données présentées dans ce sujet sont synthétiques, aléatoires et individuelles. Toute ressemblance avec une situation réelle ou des personnes existantes serait purement fortuite.

Exercice 1

On dispose d'un jeu de données `alim` sur les habitudes alimentaires de $I = 150$ individus. Ce jeu de données possède 18 variables :

- 15 variables quantitatives représentant, par catégorie d'aliment, la masse (en kg) consommée en moyenne par mois : poulet, boeuf, porc, poisson, fruits de mer, légume vert, légume racine, autre légume, féculent, crudités, pâtisseries, encas sucré, encas salé, soda, alcool,
- 3 variables qualitatives taux fer, taux vitamines, cholesterol relatives à des données biologiques.

On considère comme variables actives les 15 variables quentitatives de ce jeu de données, et on prend les 3 variables qualitatives comme variable supplémentaires.

En vous appuyant sur les données disponibles, proposez une analyse de ce jeu de données. Vous articulerez votre analyse en trois parties :

ACP Y a-t-il des ressemblances ou des oppositions entre les individus ? Pouvez-vous dresser une typologie ? Certains aliments sont-ils correlés ? Peut-on résumer des variables fortement correlées par des variables synthétiques ? Le nombre d'axes factoriels retenus est-il pertinent ? Selon quel critère ?

CAH Quels sont les différents types de comportements alimentaires ? Qu'est-ce qui les caractérise ? Le premier plan factoriel suffit-il à résumer leurs différences ? Le nombre de clusters choisi est-il pertinent ? D'après quel critère ?

ADisc Quelles sont les variables permettant de discriminer les différents groupes ? Quelles sont celles qui décrivent au mieux les groupes ?

Exercice 2

On a relevé le nombre d'occurences de 31 termes dans les discours de personnages politiques. Ces données sont disponibles dans le jeu de données `lexique`. On considère comme ligne supplémentaire le terme "écologie", et la première colonne contient les noms de lignes.

Analysez ce jeu de données en vous basant sur les résultats de l'analyse factorielle des correspondance que vous mènerez.

- Vous détaillerez le contenu du discours de Hameau et Lachambre,
- ainsi que les apparitions des termes "transparence" et "emploi".
- Vous exploiterez les données disponibles pour déterminer les attirances et répulsions entre modalités.
- Vous examinerez la projection du terme "écologie" et les relations qu'il entretient avec les autres modalités.