

TD/TP d'informatique spécialisée

Exercice 1 :

On s'intéresse à l'âge (en mois), le poids, la taille (en cm) et le sexe (m ou f) de 237 enfants et jeunes (de moins de 21 ans).

Les données sont accessibles via le fichier enfants.csv disponible sur Moodle.

1. (a) Représenter la taille de ces personnes.
La taille est-elle distribuée normalement ?
(b) Représenter la taille de ces personnes suivant le sexe. On pourra utiliser les fonctions graphiques de base puis le package ggplot2 en consultant
<http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization>.
(c) Déterminer la taille moyenne de ces personnes ainsi que la taille moyenne suivant le sexe.
2. A l'aide d'une régression linéaire multiple, justifier que le sexe n'a pas d'influence significative sur le poids des personnes de cette étude.
3. Donner une estimation du poids d'un enfant âgé de 12 ans et mesurant 150 cm.
On pourra donner une estimation ponctuelle puis par intervalle de confiance.

Exercice 2 :

On s'intéresse au déplacement d'un drone dans un repère $(0; \vec{i}, \vec{j}, \vec{k})$.

À l'instant $t = 0$, il est situé à l'origine du repère.

On considère le relevé suivant de positions obtenues, durant 2 s, toutes les $T_e = 0,1$ s.

Les données sont accessibles via le fichier drone.csv disponible sur Moodle.

1. Représenter graphiquement le déplacement du drone dans l'espace.
2. Un modèle de régression multiple de Z en fonction de X et Y est-il envisageable ?
3. Donner une estimation de la position du drone au bout de 10 secondes.

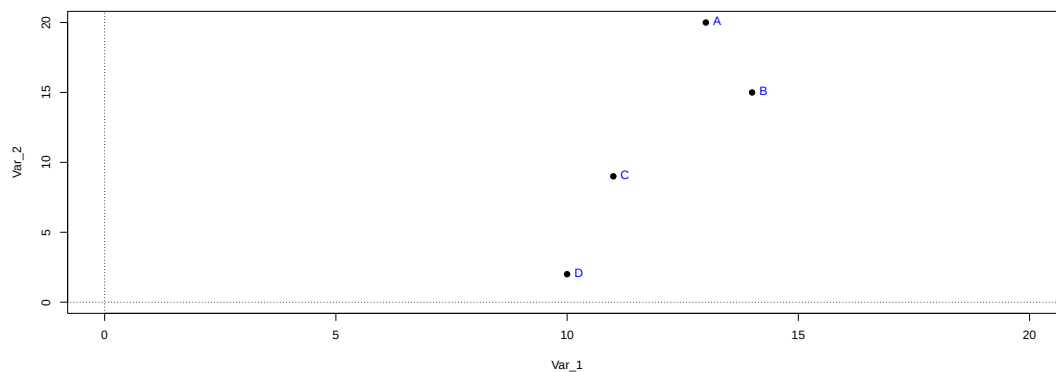
Exercice 3 : Premiers pas autour de la CAH

On considère quatre individus identifiés par deux variables.

1. Créer, à l'aide de R, le tableau suivant :

	Var_1	Var_2
A	13	20
B	14	15
C	11	9
D	10	2

2. Visualiser le jeu de données avec les fonctions *plot*, *text* et *abline*.



3. Calculer les distances $d(A; B)$ et $d(A; C)$.
4. (a) Déterminer toutes les distances entre individus en utilisant la fonction *dist*.
(b) En déduire le premier groupe.
5. (a) Construire le dendrogramme à l'aide de R.
(b) Couper ce dendrogramme pour obtenir la partition en deux classes.
6. (a) Construire un nouveau dendrogramme en utilisant des variables centrées réduites.
(b) Couper ce nouveau dendrogramme pour obtenir la partition en deux classes.

Exercice 4 :

On considère différents fromages caractérisés par différentes variables : nombre de calories (pour 100 g), la quantité de sodium, de calcium, ...

Les données sont accessibles via le fichier *fromage.csv* accessible sur Moodle.

1. Statistique descriptive
 - (a) Représenter le nombre de calories pour tous les fromages.
 - (b) Représenter le nombres de calories (pour 100 g) à l'aide d'un diagramme en barres.
On pourra distinguer ceux qui font moins de 250 kcal de ceux qui font plus de 300 kcal (pour 100 g).
2. Classification ascendante hiérarchique
 - (a) Réaliser une CAH des fromages suivant uniquement le critère du nombre de calories.

- (b) Couper ce dendrogramme pour obtenir la partition en deux classes.
 - (c) Représenter ces données à l'aide d'un diagramme en barres en tenant compte de ces résultats.
 - (d)
 - i. Réaliser une CAH des fromages en tenant compte de tous les critères après avoir justifié qu'il est préférable de centrer et réduire les différentes variables quantitatives.
 - ii. Couper ce dendrogramme pour obtenir la partition en trois classes.
 - iii. Donner les fromages associés à chacune des classes.
3. Mise en place d'une ACP
- (a) Déterminer la matrice de corrélation associée aux différentes variables permettant de caractériser ces fromages.
Afficher également le corrélogramme.
 - (b) Y a-t-il une relation linéaire (cf. régression linéaire multiple) entre les calories et les autres variables quantitatives.
Quelles sont les variables dont l'effet est significatif sur le nombre de calories d'un fromage ?
 - (c) Réaliser une ACP sur ces données et donner les valeurs propres ainsi que le graphe des éboulis.
 - (d) Représenter les variables en fonction des deux "premiers" axes principaux.
Donner le niveau de corrélation de chaque variable avec les axes principaux.
 - (e) Quels produits conseillerez-vous pour avoir un apport en calories, protéines, cholestérol et lipides faible ?
Que peut-on dire du fromage de chèvre à pâte molle ?

Exercice 5 :

On s'intéresse à un jeu de données disponible sous Moodle liées aux joueurs des clubs de football suivants :

Manchester City, Real Madrid, Inter (de Milan), PSG et FC Barcelone.

Les données sont extraites de critères/informations utilisées par un célèbre jeu vidéo de football en 2023.

On retrouve ainsi l'âge des joueurs, leur nationalité, la note (globale) qui leur est attribuée, leur pied préférentiel, leur taille (en cm), leur poids (en kg), ...

1. Déterminer la répartition des nationalités des joueurs dans ces cinq équipes.
Proposer une représentation graphique adaptée.
2. (a) Est-il judicieux de réaliser une CAH sur ce jeu de données en considérant les données brutes ?
(b) Réaliser une CAH adaptée en utilisant 3 classes.
Qu'est-ce qui caractérise (principalement) les individus de chaque classe ?
3. Réaliser une CAH sur les données liées au PSG en utilisant, une nouvelle fois, 3 classes.
Comparer les groupements obtenus avec ceux de la question précédente.
4. Déterminer la matrice de corrélation ainsi que le corrélogramme.
Interpréter les résultats qui vous paraissent significatifs.

5. (a) Y a-t-il une relation lineaire (cf. régression linéaire multiple) entre la valeur d'un joueur et les variables suivantes :
Note globale, Réputation internationale, Notes sur la capacité à effectuer des gestes techniques, taille, poids et âge.
- (b) Quelles sont les variables qui influent réellement sur la valeur d'un joueur ?
Préciser le type d'influence de chacune de ces variables.
6. (a) Réaliser une ACP sur les individus d'une équipe, par exemple, Manchester City.
- (b) Préciser les variables fortement corrélées (positivement) avec l'axe 1.
Qu'en est-il pour l'axe 2 ?
Quel lien existe-t-il entre l'axe 2 et les variables poids et taille ?
- (c) Représenter les joueurs de Manchester City dans le plan associé aux axes 1 et 2.
Que peut-on dire concernant Kévin De Bruyne ? Erling Haaland ?
- (d) Représenter les joueurs dans le même plan en tenant compte de leur "bon" pied.
Que constate-t-on ?
7. Recommencer l'étude sur les deux clubs espagnols : FC Barcelone et Real Madrid.
8. Recommencer l'étude sur tous les clubs.
Que peut-on dire concernant Lionel Messi et Neymar Jr ?
9. À l'aide d'une nouvelle étude, comparer les effectifs des deux grands clubs espagnols : le Real Madrid et le FC Barcelone.

Exercice 6 :

L'objectif est d'étudier les données fournies dans le fichier intitulé phones.csv disponible sur Moodle dans lequel sont fournis, pour des smartphones sortis en 2022 la marque, l'OS, la tailles (en pouces), la capacité de la batterie (en mAh), le poids (en g), la capacité de stockage (en Go), le prix (en dollars américains).

1. Proposer un graphique illustrant le nombre de téléphones sortis en 2022 pour chaque marque.
2. Déterminer la matrice de corrélation ainsi que le corrélogramme associés aux variables quantitatives. Interpréter les résultats obtenus.
3. Déterminer un modèle linéaire explicitant le prix d'un smartphone en fonction des autres variables quantitatives.
4. (a) Réaliser une CAH sur le prix des smartphones de la marque Apple.
Commenter ces résultats.
- (b) Réaliser une CAH sur le prix des smartphones des marques Apple, Samsung et Xiaomi.
Commenter ces résultats.
5. (a) Réaliser une ACP sur ces données.
Que dire des deux premiers axes principaux ?
- (b) Représenter les individus dans le plan engendré par les deux premiers axes principaux.
Que peut-on dire de l'iPhone SE (2022) ?