Analyse de données
TP 1

L3 MIASHS

05/09/19

Les données considérés dans ce TP sont les performances réalisées par des athlètes lors de deux compétitions de décathlon, le Décastar de 2004 et les Jeux Olympiques de 2004. Les 10 épreuves sont abrégées de la façon suivante :

100m : course de 100 mètres

Long.jump : saut en longueur

Shot.put : lancer du poids

High.jump : saut en hauteur

400m : course de 400 mètres

110m.hurdle : course du 110 m haies

Discus : lancer du disque

Pole.vault : saut à la perche

Javeline : lancer du javelot

1500m : course de 1500 mètres

En plus des 10 épreuves, le jeu de données comporte une variable "Rank" qui correspond au classement de l'athlète et une variable "Points" qui contient le nombre de points obtenus sur chaque épreuve.

Dans R, on peut accéder à l'aide en tapant ? suivi du nom de la fonction qu'on utilise.

1 Les données

a) Charger le package FactoMineR et le tableau de données "decathlon".

b) Visualiser les données. Qui sont les individus ? Quelles sont les variables ?

c) Que vaut n ? Que vaut p ?

2 Etude des données

a) Calculer la moyenne et l'écart-type de chacune des variables des disciplines (10 premières colonnes).

b) A l'aide de la fonction **boxplot**, visualiser la répartition des différentes variables selon la compétition considérée. Par exemple en entrant la commande suivante :
> boxplot(decathlon$Long.jump~decathlon$Competition)
Commenter les résultats obtenus.

3 Rappels sur la matrice de corrélation

a) Rappeler comment est défini le coefficient de corrélation linéaire entre deux variables $X = (x_1, \cdots, x_n)^T$ et $Y = (y_1, \cdots, y_n)^T$. Montrer que $r \in [-1, 1]$. A quelle(s) condition(s) a-t-on $r = 1$? A quelle(s) condition(s) a-t-on $r = -1$? Donner une interprétation graphique sur un nuage de points selon si $r = 0, -1$ ou 1.

b) Donner un exemple de variables aléatoires X et Y qui ne sont pas indépendantes et qui vérifient $cov(X, Y) = 0$. Comment interpréter un coefficient de corrélation proche de 0 ?

4 Etude de la matrice de corrélation

a) Calculer la matrice de corrélation S (*n.b.* on ne considère que les 10 premières variables). Commenter (certaines variables sont-elles corrélées positivement/négativement ?).

b) Taper la commande
> image(z=S[,seq(10,1,-1),zlim=c(-1,1),col=gray((0 :32)/32))
Interpéter le graphique obtenu.

c) Réaliser des nuages de points afin de visualiser la relation entre les différentes variables des disciplines. Vous pouvez soit réaliser des graphiques différents pour chacune des compétitions *Decastar* et *OlympicG* ou bien essayer de superposer l'information *Competition* dans le graphique des nuages de points (avec un code couleur). Commenter les graphiques obtenus : la relation entre les variables est-elle linéaire ?