Analyse des données

TP 1 : Analyse en composantes principales (ACP) avec R (packages FactoMineR et Factoshiny, fonction PCA())

Le fichier étudié concerne 25 pays de l'Union Européenne (source : Eurostat 2002). Les variables considérées sont les suivantes :

- espérance de vie à la naissance pour un homme (en années)
- espérance de vie à la naissance pour une femme (en années)
- population (en milliers d'habitants)
- taux d'activité (en pourcentage)
- produit intérieur brut par habitant (en standards de pouvoir d'achat)
- taux d'inflation (en pourcentage)
- taux d'emploi (en pourcentage)
- taux de chômage (en pourcentage)
- taux de chômage longue durée (en pourcentage)
- nombre de mariages (pour 1000 personnes)
- nombre d'abonnés aux services de téléphonie mobile (en milliers)

1. Importer le fichier `pays-eu.txt` sous `R`.
2. Réaliser une étude univariée rapide des données à l'aide d'indicateurs numériques et de graphiques.
3. Donner la matrice des corrélations et les nuages de points associés. Commenter.
4. L'ACP vous paraît-elle justifiée ?
5. Déterminer le nombre de composantes principales à retenir pour cette ACP. Justifier votre réponse.
6. Interpréter les composantes principales retenues à l'aide des variables initiales. Donner un (des) graphique(s) permettant de visualiser l'interprétation.
7. Quels sont les pays bien représentés sur chacun des axes retenus ? Justifier votre réponse.
8. Commenter les contributions des pays aux premier axe.
9. Réaliser le(s) graphique(s) des pays et commenter l'ACP.

Annexe : Initiation à R

R utilise des **fonctions** ou des opérateurs qui agissent sur des objets (vecteurs, matrices, data-frames etc.).

Importation de données
Lecture d'un fichier texte : fonction read.table(.)

```
employes=read.table("employes.txt",header=T)
employes[,-1]#enlève la première colonne
head(employes)# début du fichier
employes[,c(2,3,4)] # donne les colonnes 2, 3 et 4 du fichier
employes[1:10,]# donne les 10 premières lignes du fichier
```

employes est un data-frame (format par défaut sous R, format obtenu par la lecture de fichiers externes).

Les jeux de données de R
R contient plusieurs jeux de données, qui peuvent être chargés par la fonction data. Pour voir leur liste taper : data()

```
data(Orange)  # charge le jeu de données Orange
help(Orange)   # donne des informations sur le jeu de données
```

Edition intéractive des données
La fonction edit permet de modifier les données avec la souris. En cliquant sur le nom de la colonne, on peut aussi modifier son type (real=numérique=quantitative, character=facteur=qualitative).

```
data(airquality) # charge le jeu de données airquality
aq=edit(airquality) # édite le jeu de données et le stocke dans aq
```

Ainsi, le jeu de données initial airquality n'est pas modifié. On peut également utiliser la fonction fix, mais les modifications qu'on apporte écrasent le tableau originel.
Pour créer un jeu de données en entrant les données au clavier :

```
donnees=data.frame()  # crée le tableau donnees
fix(donnees)  # édite le tableau vide
```

Résumé numérique des variables
Les fonctions suivantes donnent les statistiques descriptives usuelles pour les variables quantitatives :

```
attach(airquality) # rend les variables appelables directement
mean(Ozone) # moyenne de Ozone
sd(Ozone)  # écart-type
var(Ozone)  # variance
median(Ozone) # médiane
quantile(Ozone)  # donne le min, le max et les quartiles
summary(Ozone)  # min, max, moyenne et quartiles
summary(airquality)  # statistiques pour le data frame tout entier
cor(Ozone,Temp) # corrélation entre Ozone et Temp
```

Si on veut que R calcule ces statistiques en présence de données manquantes, il faut rajouter l'argument na.rm=T aux fonctions ci-dessus.
Pour avoir le nombre de données non manquantes : sum(!is.na(Ozone))

Si la variable est de type facteur, summary donne simplement les effectifs des différentes modalités, ce que l'on peut aussi obtenir par la fonction table.

Comme cela n'a pas de sens de calculer une moyenne ou une variance pour une variable qualitative, toute variable qualitative codée en numérique doit être mise au type facteur (fonction factor).

```
summary(employes)# noter ce qu'il fait pour la variable sexe
employes$sexe=factor(employes$sexe,labels=c("F","M"))#modifie le type
de la variable sexe dans le tableau
#cette opération se fait également en éditant le fichier
# et en changeant son type
summary(employes)
table(employes$sexe) #fait la même chose que summary pour une qualitative
round(prop.table(table(employes$stat_pro)),digits=2)#tableau des fréquences
relatives arrondies à 2 décimales

#Tableau de contigence
employes$stat_pro=factor(employes$stat_pro, labels=c("employé de bureau",
+   "agent de sécurité","manager")) # transforme la variable qualitative
 codée stat_pro en facteur
 tab= table(employes$sexe,employes$stat_pro) # crée le tableau de contingence et
le stocke dans tab
```

Graphiques
Pour les variables qualitatives :
Diagramme circulaire : pie

```
attach(employes)
table(stat_pro)
pie(table(stat_pro))
barplot(prop.table(table(stat_pro)),col=1:3)
```

Diagramme en colonnes

```
barplot(table(stat_pro))
barplot(prop.table(table(stat_pro)),col=1:3)
```

Pour les variables quantitatives discrètes :
Diagramme en bâtons (respecte l'espacement des valeurs) :

```
#diagramme en effectif
plot(table(educ),lwd=5,col="red",xlab="Nombre d'années d'études",
ylab="effectif",main="Diagramme en bâtons de la variable educ")

# diagramme en fréquence
n=length(educ)
plot(table(educ)/n,lwd=5,col="red",xlab="Nombre d'années d'études",
ylab="effectif",
main="Diagramme en bâtons de la variable educ")
```

Pour les variables quantitatives continues :
Histogramme :

```
hist(salaire)
```

Boîte à moustaches :

```
boxplot(salaire)
```

Diagramme de dispersion (nuage de points)

```
plot(salaire,age,xlab="âge",ylab="salaire")
# salaire en ordonnée, âge en abscisse
avec les labels des axes
```

Exercice (à faire en autonomie)

Le fichier de données `employes.txt` concerne 474 employés d'une entreprise américaine. Les variables relevées sont les suivantes :

- sexe (2 pour masculin, 1 pour féminin)
- éducation : nombre d'années passées à l'école
- statut professionnel (1 si "employé de bureau", 2 si "agent de sécurité", 3 si "manager")
- salaire annuel à l'embauche dans l'entreprise (en dollars)
- salaire annuel courant (en dollars)
- ancienneté dans l'entreprise en nombre de mois
- expérience passée (en nombre de mois)
- nationalité (1 pour américaine, 0 sinon)
- âge (en années)

1. Importer le fichier `employes.txt` sous `R`.
2. Transformer les variables qualitatives du fichier en facteurs et mettre des labels.
3. Donner des résumés numériques et des graphiques représentant les distributions des variables du fichier "employes".
4. Donner la matrice des corrélations (arrondie à 2 décimales) entre les variables quantitatives du fichier. Commenter.
5. Sauver le script.