

Analyse de données  
M1 Statistique et économétrie - 2012-2013  
V. Monbet  
Exploration des données - Analyse factorielle

L'objectif de ce TD sont d'apprendre à mener une analyse descriptive simple : manipuler des données, faire des graphiques.  
D'après un TP de Frédéric Bertrand, IRMA, Strasbourg.

## Exercice 1, fichier de données : europe.csv

Nous vous demandons dans cet exercice de tracer une boîte à moustaches. Pour cela, il faut que vous téléchargiez sur le site de Frédéric Bertrand le fichier de données source correspondant, puis que vous tapiez les lignes de commande suivantes :

```
# Ici il faut indiquer le chemin d'accès au fichier sur le disque de
stockage. Par exemple, pour Frédéric Bertrand c'est :
> Chemin <- "C:/Documents and Settings/Bertrand/Bureau/"
# Lit le fichier. Vous déterminerez en particulier le rôle des
options "dec", "sep" et "quote".
> europe <- read.table(paste(Chemin,"europe.csv",sep=""),dec=".",
+ sep=";",quote="n",header=T)
# Vérification du bon déroulement de l'importation et statistiques descriptives
> head(europe)
> str(europe)
> summary(europe)
> range(europe$Durée.heures.)
> sd(europe$Durée.heures.)
# Quelques représentations graphiques
> plot(europe)
> boxplot(europe$Durée.heures.,ylab="Durée (heures)")
> points(1,mean(europe$Durée.heures.),pch=2)
```

`pch` est une option graphique qui définit le symbole qui représente les observations.

```
# Sauvegarde de la boîte à moustaches au format .pdf
> pdf(file = paste(Chemin,"boxplot.pdf",sep=""),
+ width = 6, height = 6, onefile = TRUE, family = "Helvetica",
+ title = "Europe boxplot", paper = "special")
> boxplot(europe$Durée.heures.,ylab="Durée (heures)")
> points(1,mean(europe$Durée.heures.),pch=2)
> dev.off()
# Sauvegarde de la boîte à moustaches au format .ps
```

```
> postscript(file = paste(Chemin,"boxplot.eps",sep=""),
+ width = 6, height = 6, onefile = TRUE, family = "Helvetica",
+ title = "Europe boxplot", horizontal = FALSE, paper = "special")
> boxplot(europe$Durée.heures.,ylab="Durée (heures)")
> points(1,mean(europe$Durée.heures.),pch=2)
> dev.off()
```

## Exercice 2, fichier de données : iris

Le logiciel R est un ensemble de bibliothèques de fonctions appelées "packages". Chaque bibliothèque contient des jeux de données. Pour connaître par exemple les jeux de données contenus dans le package base, écrire l'instruction suivante :

```
> data(package = "base")
```

Le résultat apparaît dans une fenêtre R data sets. En voici un extrait :

Le résultat apparaît dans une fenêtre R data sets. En voici un extrait :

Data sets in package 'datasets':

AirPassengers.....Monthly Airline Passenger Numbers 1949-1960

BJsales.....Sales Data with Leading Indicator

BJsales.lead (BJsales)..Sales Data with Leading Indicator

BOD.....Biochemical Oxygen Demand

...

iris.....Edgar Anderson's Iris Data

1. Noter la présence du fichier iris. Les données de ce fichier sont célèbres. Elles ont été collectées par Edgar Anderson. Le fichier donne les mesures en centimètres des variables suivantes :

- (i) longueur du sépale (Sepal.Length),
- (ii) largeur du sépale (Sepal.Width),
- (iii) longueur du pétale (Petal.Length) et
- (iv) largeur du pétale (Petal.Width)

pour trois espèces d'iris :

- (i) Iris setosa,
- (ii) Iris versicolor et
- (iii) Iris virginica.

Sir R.A. Fisher a utilisé ces données pour construire des combinaisons linéaires des variables permettant de séparer au mieux les trois espèces d'iris.

2. Pour analyser le fichier iris, il faut le charger. Quelle est l'instruction qu'il faut taper pour charger ce fichier ?
3. Taper une à une chacune des instructions ci-dessous et noter le résultat obtenu si possible. Attention, le logiciel R n'est pas indifférent aux majuscules.

```
> iris
```

```
> dim(iris)
```

```
> names(iris)
```

Quelle(s) différence(s) faites-vous avec la commande `str(iris)` ? Tapez les commandes suivantes :