T. P. n° 2
Statistique descriptive avec le logiciel R

Introduction : Ce T.P. a pour but de vous faire assimiler les notions découvertes de statistiques descriptives univariée et bivariée pendant les cours des mardi 12 février 2008 et mardi 26 février 2008. Donc ce T.P. durera deux semaines consécutives, à savoir le lundi 25 février 2008 et le lundi 3 mars 2008.

Notre objectif pendant ces deux semaines :
- Manipuler les données (et non plus apprendre, qui était le but du premier T.P.).
- Faire des graphiques.
- Et bien sûr continuer à apprendre les bases du langage.

Consignes pour ce T.P. :
- Suivre pas à pas les étapes et voir ce qui se passe.
- Faire les exercices proposés.
- Ne pas hésiter à utiliser l'aide en ligne de R.
- Vous ne comprendrez peut-être pas tous les détails mais la meilleure chose à faire est de taper le code et de voir le résultat produit. Soyez curieux et n'hésitez pas à le modifier pour voir « ce qu'il se passe ».

Quelques notions utiles
- Le symbole > se trouve à chaque début de ligne de commandes (appelé prompt).
- Le symbole $<-$ est le symbole d'affectation.
- Le symbole # signifie le début d'un commentaire.
- Lorsque vous travaillez sous R, il peut être intéressant de conserver les résultats et les graphiques de vos analyses. Le plus simple, dans un premier temps, est de les enregistrer dans un document word à l'aide du copier / coller. Pour ce faire, aller dans le menu « File » ou « Fichier », sélectionner « Copy to the clipboard »« as a Bitmap ». Noter que les graphes peuvent être réduits ou agrandis sans déformation.

Exercice 1 Fichier de données : europe.csv

Dans le chapitre « Statistique descriptive » du cours, il y a une boîte à moustaches qui a été intégrée à ces notes. Nous vous demandons dans cet exercice de la tracer à nouveau et de la comparer avec celle qui figure dans le cours. Pour cela, il faut que vous vous procuriez le fichier de données source. Vous avez le choix entre aller chercher les données dans le polycopié de cours (attention, une nouvelle version de ce chapitre a été mise en ligne le 25 février 2008) ou télécharger depuis mon site dans la partie « Enseignement 2007-2008 », Module « Étude de cas », le fichier « europe.csv ».

1. Choisir une façon de vous procurez le fichier. Si c'est la première, nommer votre fichier « europe » et enregister-le au format « .csv ».
2. Taper les ligne de commande suivantes :

```
# Ici il faut indiquer le chemin d'accès au fichier sur le disque
de stockage
> Chemin <- "C :\\Documents and Settings\\Bertrand\\Bureau\\"

# Lit le fichier. On déterminera en particulier le rôle des
options « dec », « sep »et « quote ».
> europe <- read.table(paste(Chemin,"europe.csv",sep=""),dec=".",
+ sep=" ;",quote="¨",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)

# 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`.

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 :

```
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` et du fichier `women` sur lequel vous avez déjà travaillé (cf T.P. 1). Le fichier `iris` a toute une histoire. La connaissez-vous ? Les données de ce T.P. sont célèbres. Elles ont été collectées par Edgar Anderson[1]. Vous auriez pu le deviner. Pourquoi ? 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[2] 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 et aux minuscules.
   ```
   >iris
   >dim(iris)
   >names(iris)
   >iris$Species
   >iris$Petal.Length
   ```
   Quelle(s) différence(s) faites-vous avec la commande `> str(iris)` ?

[1] E. Anderson (1935) The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, 59, 2-5

[2] R.A. Fisher, (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II, 179-188

4. La dernière colonne du fichier `iris` contient le nom des espèces réparties en trois catégories : setosa, versicolor et virginica. Pour accéder à celle-ci, il faut utiliser l'instruction `iris$Species`. On dit alors que la dernière colonne contient une variable qualitative à trois modalités[3] appelés « levels » par le logiciel. La fonction « levels » appliquée à la colonne `iris$Species` donne les modalités de la variable. En effet, il suffit de taper :
   ```
   > levels(iris$Species)
   ```
   Pour résumer l'information contenue dans cette variable, on utilise l'instruction :
   ```
   > summary(iris$Species)
   ```
   Quel est le résultat qui s'affiche ?
5. Cette information peut être obtenue en construisant un tableau (`table`) comptabilisant le nombre d'individus par modalité. Pour ce faire, taper l'instruction suivante :
   ```
   > table(iris$Species)
   ```
   Comparer avec le résultat obtenu à la question précédente.
6. Le logiciel R permet de réaliser des graphiques comme vous avez pu le constater lors du premier T.P. Lorsqu'une instruction graphique est lancée, une nouvelle fenêtre « device » est ouverte. Les représentations graphiques liées aux variables qualitatives sont la représentation en secteurs ou camembert (`pie`) et la représentation en bâtons (`barplot`). Taper les lignes de commande suivantes :
   ```
   >pie(table(iris$Species))
   >barplot(table(iris$Species))
   ```
7. Il existe un paramètre permettant de découper la fenêtre graphique :
   `par(mfrow=c(nl,nc))` ou `par(mfcol=c(nl,nc))`.
   `nl` définit le nombre de graphiques en lignes et `nc` définit le nombre de graphiques en colonnes.
   `mfrow` signifie que l'ordre d'entrée des graphiques s'effectue selon les lignes et `mfcol` signifie que l'ordre d'entrée des graphiques s'effectue selon les colonnes.
   Supposons que nous voulions représenter six graphiques dans une fenêtre en deux lignes et trois colonnes. La première instruction conduit à entrer les graphiques selon l'ordre :

1	2	3
4	5	6

   La seconde instruction conduit à entrer les graphiques selon l'ordre :

1	3	5
2	4	6

   Deux botanistes se sont également intéressés aux iris et ont collecté les espèces suivantes :
   ```
   > collection1<-rep(c("setosa","versicolor","virginica"),c(15,19,12))
   ```

[3] Nous reverrons ces définitions et ces notations lors de l'analyse de la variance

```
> collection2<-rep(c("setosa","versicolor","virginica"),c(22,27,17))
```

En utilisant la fonction `par(mfrow=c(2,2))`,

1. Construire les camemberts liés à ces deux nouvelles distributions. Commenter.
2. Construire les représentations en bâtons de ces deux nouvelles distributions. Commenter.
3. Discuter des avantages et des inconvénients de ces deux types de représentations.

8. La troisième colonne du fichier iris contient la longueur du pétale. Il s'agit d'une variable mesurée qualifiée alors de variable quantitative. Pour résumer l'information contenue dans cette variable, nous utilisons la fonction `summary`. Taper la ligne de commande suivante :
`>summary(iris$Petal.Length)`
Quel résultat obtenez-vous ?

   La plus petite (`Min`) longueur de pétale est égale à $1cm$ tandis que la plus grande (`Max`) est égale à $6.9cm$. La moyenne (`Mean`) représente la somme des valeurs de la distribution divisée par le nombre total d'iris. Elle est égale à $3.758\,cm$. Si l'ensemble des 150 longueurs de pétale sont classées par ordre croissant, `1sr Qu.`, `Median` et `3rd Qu.` sont les trois valeurs qui permettent de couper la distribution en quatre parties égales. Nous rappelons que nous les appelons respectivement premier quartile, médiane (ou deuxième quartile) et troisième quartile.

   Essayons de retrouver ces six valeurs de paramètres. Taper les lignes de commande suivantes :
```
> min(iris$Petal.Length)
> max(iris$Petal.Length)
> mean(iris$Petal.Length)
```

   Pour calculer la moyenne nous aurions pu procéder autrement. Taper les lignes de commande suivantes :
```
> sum(iris$Petal.Length)
> length(iris$Petal.Length)
> sum(iris$Petal.Length)/length(iris$Petal.Length)
```
Obtenez-vous le même résultat que précédemment ?

   Maintenant occupons nous de retrouver les valeurs des trois quartiles. Pour cela, taper la ligne de commande suivante :
`> sort(iris$Petal.Length)`
Que se passe-t-il ?
Puis continuer par taper les lignes de commandes suivantes :
```
> ordLpetal <- sort(iris$Petal.Length)
> ordLpetal # commenter le résultat
```

```
> sum(ordLpetal)/length(ordLpetal)
> ordLpetal[38]
> (ordLpetal[75]+ordLpetal[76])/2
> ordLpetal[113]
```

9. Une des représentations adéquate est l'histogramme (`hist`). Taper la ligne de commande suivante :
   `> hist(iris$Petal.Length,col=grey(0.6),main="Histogramme")`
10. Réaliser le même type d'analyse sur chacune des trois autres variables quantitatives : largeur du pétale, longueur du sépale et largeur du sépale. Notez que vous n'avez pas toutes les instructions à réécrire en utilisant le système de flèches du clavier. $\uparrow$ et $\downarrow$ vous permettent de retrouver les fonctions que vous avez utilisées. $\leftarrow$ et $\rightarrow$ vous permettent de vous déplacer dans la fonction et donc, dans changer certains paramètres.
11. À partir de cette question, nous allons aborder la statistique descriptive bivariée. Si vous en êtes arrivé là le lundi 25 février 2008, reportez-vous alors au cours.
12. Lire le chapitre se reportant à la statistique descriptive bivariée.
13. Une fois réalisée les graphiques pour chaque variable prise séparemment, l'étude peut porter sur la relation entre deux variables. Nous parlons alors de croisement de deux variables ou d'étude bivariée. La représentation graphique liant deux variables quantitatives est le nuage de points. Représentons par exemple la longueur et la largeur du pétale pour les 150 iris contenus dans le fichier de données. Pour cela, exécuter la ligne de commentaire suivante :
   ```
   >plot(iris$Petal.Length, iris$Petal.Width,
   + xlab="Longueur du pétale", ylab="Largeur du pétale",
   + main="Nuage de points", pch=20)
   ```
   Faire un commentaire.

   Dans cette représentation graphique, plusieurs individus peuvent être situés sur un même point. La fonction `sunflowerplot` permet de visualiser ces superpositions. Taper la ligne de commande suivante :
   ```
   > sunflowerplot(iris$Petal.Length, iris$Petal.Width,
   + xlab="Longueur du pétale", ylab="Largeur du pétale",
   + main="Nuage de points", pch=20)
   ```
14. Réaliser l'étude du croisement de deux variables quantitatives de votre choix. Il est clair que le sens biologique de l'étude ne doit pas être négligé.
15. La représentation graphique permettant de lier une variable qualitative et une variable quantitative est la boîte à moustaches (`boxplot`). Représentons par exemple la longueur des pétales en fonction de l'espèce. Pour cela, taper la ligne de commande suivante :
    `> boxplot(iris$Petal.Length ~ iris$Species,col=grey(0.6))`
    Commenter.
16. Choisir une autre variable quantitative, croiser-la avec la variable espèce d'iris et commenter.

17. Et pour finir...Le nuage de points comme les boîtes à moustaches montrent que les données morphologiques des iris semblent liées à l'espèce. Il pourrait donc être intéressant de réaliser des graphiques différents pour chacune des modalités Iris setosa, Iris Versicolor et Iris Virginica ou de superposer l'information espèce dans le graphique des nuages de points. Nous vous proposons ici quelques développements. Libre à vous, de les refaire ou d'en trouver d'autres... Taper alors les lignes de commande suivantes :

```
# Tracé des histogrammes des longueurs des pétales de l'ensemble
des iris, des I. setosa, des I. versicolor et des I. virginica
> par(mfrow=c(2,2))
> br0=seq(0,8,le=20)
> hist(iris$Petal.Length, main="Ensemble des 150 iris",
+ xlab="Longueur du pétale", br=br0)
> hist(iris$Petal.Length[iris$Species=="setosa"], main="Setosa",
+ xlab="Longueur du pétale", br=br0)
> hist(iris$Petal.Length[iris$Species=="versicolor"],
+ main="Versicolor", xlab="Longueur du pétale", br=br0)
> hist(iris$Petal.Length[iris$Species=="virginica"],
+ main="Virginica", xlab="Longueur du pétale", br=br0)

> par(mfrow=c(2,2))
> plot(iris$Petal.Length, iris$Petal.Width,
+ xlab="Longueur du pétale", ylab="Largeur du pétale",
+ main="Nuage de points", pch=20)
> plot(iris$Petal.Length[iris$Species=="setosa"],
+ iris$Petal.Width[iris$Species=="setosa"],
+ xlim=c(1,6.9), ylim=c(0.1,2.5), xlab="",ylab="",
+ main="iris setosa", pch=20)
> plot(iris$Petal.Length[iris$Species=="versicolor"],
+ iris$Petal.Width[iris$Species=="versicolor"],
+ xlim=c(1,6.9), ylim=c(0.1,2.5), xlab="", ylab="",
+ main="iris versicolor", pch=20)
> plot(iris$Petal.Length[iris$Species=="virginica"],
+ iris$Petal.Width[iris$Species=="virginica"],
+ xlim=c(1,6.9), ylim=c(0.1,2.5), xlab="", ylab="",
+ main="iris virginica", pch=20)
```

Et maintenant à vous sans les lignes de commande.

Exercice 3 Données brutes ou groupement en classes.

Comme vous l'avez découvert dans le chapitre 2 du polycopié de cours, il arrive que lorsque la série statistique qui étudie un caractère quantitatif, comporte un grand nombre de valeurs, nous préférons alors regrouper par classes puis ensuite remplacer chaque classe par son milieu. Mais les résultats en sont légèrement modifiés, ce que vous pouvez imaginer. D'ailleurs certains auteurs suggèrent des corrections par

exemple en ce qui concerne la variance, comme la correction de Sheppard, comme le déclarent Couty, Debord et Fredon dans leur livre « Mini manuel de probabilités et statistique », Dunod. D'ailleurs de ce livre, nous allons extraire le jeu de données qui va nous permettre de faire cet exercice.
Nous considérons une série statistique de 60 taux d'hémoglobine dans le sang (g/L) mesurés chez des adultes présumés en bonne santé :

Femmes	105	110	112	112	118	119	120	120	125	126
	127	128	130	132	133	134	135	138	138	138
	138	142	145	148	148	150	151	154	154	158
Hommes	141	144	146	148	149	150	150	151	153	153
	153	154	155	156	156	160	160	160	163	164
	164	165	166	168	168	170	172	172	176	179

1. On considère le groupement en classes suivant :

$$]104;114];]114;124];]124;134];]134;144];]144;154];]154;164];$$
$$]164;174];]174;184].$$

   Pour chacune des deux séries : femmes et hommes, déterminer les effectifs et les fréquences de chaque classe.
2. Effectuer une réprésentation graphique adaptée des deux distributions groupées en classe de la question 1.
3. Calculer les moyennes des trois distributions initiales : ensemble, femmes, hommes.
4. Calculer les moyennes des trois distributions (ensemble, femmes, hommes) après le regroupement en classes de la question 1., en remplaçant chaque classe par son milieu.
5. Calculer les médianes des trois distributions initiales : ensemble, femmes, hommes.
6. Calculer l'écart interquartile pour chacune des trois distributions initiales : ensemble, femmes, hommes.
7. Calculer les variances et les écart-types des trois distributions initiales : ensemble, femmes, hommes.
8. Calculer les variances et les écart-types des trois distributions après le regroupement en classes de la question 1., en remplaçant chaque classe par son milieu.
9. Pour la distribution des femmes, calculer les moments jusqu'à l'ordre 4 puis déduire les moments centrés jusqu'à l'ordre 4 puis les paramètres de caractéristique de forme de Fisher.