Master 1 GSI - Mentions ACCIE et RIM - ULCO, La Citadelle, 2013/2014

Mesures et Analyses Statistiques de Données - Probabilités

TP 1.3 - Statistiques descriptives avec le logiciel `R`

1 Introduction

Ce TP a pour objectif de vous faire assimiler les représentations numériques et graphiques des statistiques descriptives univariée et bivariée. Voici quelques remarques d'ordre général :

- 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, vous allez dans le menu `File` ou `Fichier` et vous sélectionnez `Copy to the clipboard` et `as a Bitmap`.
- Il est à noter que les graphes peuvent être réduits ou agrandis sans déformation.

2 Statistique descriptive univariée

2.1 Présentation de quelques fonctions statistiques avec le logiciel `R`

Fonction	Description
`summary()`	Donner divers paramètres statistiques
`cumsum()`	Calculer les effectifs cumulés
`sum()`	Calculer l'effectif total
`mean()`	Calculer la moyenne
`max()`	Calculer la valeur maximum
`min()`	Calculer la valeur minimum
`range()`	Calculer les valeurs minimum et maximum
`median()`	Calculer la médiane
`var()`	Calculer la variance
`sd()`	Calculer l'écart-type
`plot()`	Tracer le nuage de points
`boxplot()`	Tracer la boîte à moustaches pour une variable quantitative

2.2 Exercices

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 récupériez le fichier de données source correspondant, puis que vous tapiez les lignes de commandes adéquates :

```
>europe<-read.table("europe.csv",dec=".",sep=";",
+quote="/"", header=TRUE)
```

Cette commande lit le fichier. Attention au + dans la ligne de commandes ci-dessus ! Ce n'est pas un symbole à intégrer dans cette dernière. Il est simplement là pour dire que nous allons à la ligne par manque de place et que tout s'écrit à la suite.

1. Pouvez-vous expliquer le rôle des options `dec, sep, quote` ?

Il y a aussi une autre façon de lire un fichier avec la commande `file.choose` que vous rencontrerez par la suite.

Vérifions maintenant le bon déroulement de l'importation du jeu de données :

```
>head(europe)
>str(europe)
```

`head` pour entête et `str` pour structure.
Voici maintenant comment obtenir quelques statistiques descriptives :

```
>summary(europe)
>range(europe$Durée.heures.)
>sd(europe$Durée.heures.)
```

2. Pouvez-vous expliquer le rôle de $?

On donne ensuite 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.
On indique ensuite comment sauvegarder la boîte à moustaches au format `pdf` :

```
> pdf(boxplot.pdf
+width=6,height=6,onfile=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()
```

On indique enfin comment sauvegarder la boîte à moustaches au format `ps` :

```
> postscript(boxplot.eps
+width=6,height=6,onfile=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 spécifiques. Par exemple, pour connaître le jeux de données dans le `package base`, écrivez 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. Notez la présence du fichier `iris` . Les données de ce fichier sont célèbres. Elles ont été collectées par Edgar Anderson[1]. Le fichier donne les mesures en centimètres des variables suivantes :
   (i) longueur du sépale `(Sepal.Length)`,
   (ii) largeur du sépale `(Sepal.Width)`,

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

(iii) longueur du pétale (`Petal.Length`),

(iv) largeur du pétale (`Petal.Width`)

pour trois espèces d'iris :

(a) Iris setosa,

(b) Iris versicolor,

(c) 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. Tapez une à une chacune des instructions ci-dessous et notez le résultat obtenu si possible.
Attention, le logiciel `R` n'est pas indifférent aux majuscules et aux minuscules, comme nous l'avons déjà souligné dans le TP1.

```
>iris
>dim(iris)
>names(iris)
```

Quelle(s) différence(s) faites-vous entre ces deux dernières lignes de commandes et la commande

```
>str(iris)
```

Tapez les commandes suivantes :

```
>iris$Petal.Length
>iris$Species
```

Qu'observez-vous ?

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
```

comme vous venez de le constater à la question précédente. Nous disons alors que la dernière colonne contient une variable qualitative à trois modalités appelées `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, il est conseillé d'utiliser l'instruction :

```
>summary(iris$Species)
```

Quel résultat s'affiche alors ?

5. Cette information peut être obtenue en construisant un tableau (`table`) comptabilisant le nombre d'individus par modalité. Pour ce faire, tapez l'instruction suivante :

```
>table(iris$Species)
```

Comparez la réponse avec le résultat obtenu à la question précédente.

6. Le logiciel `R` permet de réaliser des graphiques. 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`).
Tapez les lignes de commandes suivantes :

```
>pie(table(iris$Species))
>barplot(table(iris$Species))
```

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

7. Il existe un paramètre permettant de découper la fenêtre graphique :

   `par(mfrow=c(nl,nc))` ou `par(mfcol=c(nl,nc))`

   où
   - `nl` définit le nombre de graphiques en lignes,
   - `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 vous vouliez 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))
>collection2<-rep(c("setosa","versicolor","virginica"),
+c(22,27,17))
```

   En utilisant la fonction `par(mfrow=c(2,2))`,
   (a) Construisez les camemberts liés à ces deux nouvelles distributions. Commentez les résultats.
   (b) Construisez les représentations en bâtons de ces deux nouvelles distributions. Commentez les résultats.
   (c) Discutez des avantages et 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, vous pouvez à nouveau utiliser la fonction `summary`.
   Tapez la ligne de commandes suivante :

```
>summary(iris$Petal.Length)
```

   On obtient :

```
   Min.   1stQu.   Median   Mean   3rdQu.   Max.
   1.000  1.600    4.350    3.758  5.100    6.900
```

   La plus petite `(Min.)` longueur de pétale est égale à 1.000cm tandis que la plus grande `(Max.)` est égale à 6.900cm. La moyenne `(Mean)` représente la somme des valeurs de la distribution divisée par le nombre total d'iris. Elle est égale à 3.758cm. Si l'ensemble des 150 longueurs de pétale sont classés par ordre croissant `1stQu.`, `Median` et `3rdQu.` sont les trois valeurs qui permettent de couper la distribution en quatre parties égales. Rappelons que nous les appelons respectivement premier quartile, médiane (ou deuxième quartile) et troisième quartile.
   Essayons de retrouver ces six valeurs de paramètres. Nous allons procéder en deux étapes :

   - Première étape : commençons par taper les lignes de commandes suivantes :

```
>min(iris$Petal.Length)
>max(iris$Petal.Length)
>mean(iris$Petal.Length)
```

   Remarque 2.1 Pour calculer la moyenne, nous aurions pu procéder autrement. Tapez les lignes de commandes suivantes :

```
>sum(iris$Petal.Length)
>length(iris$Petal.Length)
>sum(iris$Petal.Length)/length(iris$Petal.Length)
```

Obtenez-vous le même résultat que précédemment ?

- Deuxième étape : Occupons-nous maintenant de retrouver les valeurs des trois quartiles : pour cela, tapez la ligne de commandes suivante :

```
>sort(iris$Petal.Length)
```

Que se passe-t-il ?
Tapez ensuite 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 adéquates est l'histogramme fourni par la fonction `hist`. Regardez l'aide de `hist()` puis tapez la ligne de commandes suivantes :

```
>hist(iris$Petal.Length,col=grey(0.6),main="Histogramme")
```

`main` étant l'option qui permet d'afficher un titre dans un graphique.

10. Réalisez 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 car les flèches du clavier ↑ et ↓ vous permettent de retrouver les fonctions que vous avez utilisées. Les flèches ← et → vous permettent de vous déplacer dans la fonction et donc changer certains paramètres.

Exercice 3 **Données brutes ou groupement en classes**

Parfois, lorsque nous étudions une série statistique sur un caractère quantitatif qui 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, celle de Sheppard comme le développent Couty, Debord et Fredon dans leur livre « Mini manuel de probabilités et statistiques », édité chez Dunod. D'ailleurs nous allons extraire de ce livre 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. Nous considérons 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, déterminez les effectifs et les fréquences de chaque classe.

2. Effectuez une représentation graphique adaptée des deux distributions groupées en classe de la question 1.
3. Calculez les moyennes des trois distributions initiales : ensemble, femmes, hommes.
4. Calculez 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. Calculez les médianes des trois distributions initiales : ensemble, femmes, hommes.
6. Calculez l'intervalle interquartile pour chacune des trois distributions initiales : ensemble, femmes, hommes.
7. Calculez les variances et les écart-types des trois distributions initiales : ensemble, femmes, hommes.
8. Calculez 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.

3 Statistique descriptive bivariée

Exercice 4 Nous allons reprendre les données de l'exercice 2 et le continuer. Voici la suite de l'énoncé.

11. Une fois les graphiques réalisés pour chaque variable prise séparément, l'étude peut porter sur la relation entre les 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, exécutez la ligne de commandes suivante :

    ```
    >plot(iris$Petal.Length,iris$Petal.Width,
    +xlab="Longueur du pétale",ylab="Largeur du pétale",
    +main="Nuage de points",pch=20)
    ```

    Faites un commentaire.
    Dans cette représentation graphique, plusieurs individus peuvent être situés sur un même point. La fonction `sunflowerplot` permet de visualiser ces superspositions. Tapez la ligne de commandes suivante :

    ```
    >sunflowerplot(iris$Petal.Length,iris$Petal.Width,
    +xlab="Longueur du pétale",ylab="Largeur du pétale",
    +main="Nuage de points",pch=20)
    ```

12. Réalisez 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.

13. La représentation graphique permettant de lier une variable qualitative et une variable quantitative est la boîte à moustaches définie par la fonction `boxplot`. Représentons par exemple la longueur des pétales en fonction de l'espèce. Pour cela, tapez la ligne de commandes suivante :

    ```
    > boxplot(iris$Petal.Length~iris$Species,col=grey(0.6))
    ```

    Commentez.

14. Choisissez une autre variable quantitative, croisez-la avec la variable « espèce d'iris » et commentez.

15. 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 », « 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 trouvre d'autres... Tapez alors les lignes de commandes suivantes :

    ```
    >#  Tracé des histogrammes de longueurs de pétales de l'ensemble des iris,
    >#  des iris setosa, des iris versicolor, des iris virginica.
    >par(mfrow=c(2,2))
    >br0=seq(0,8,le=20)
    >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)

    >#  Tracé des nuages des points et de la largeur du pétale en fonction de la longueur
    >#  des pétales de l'ensemble des iris,
    >#  des iris setosa, des iris versicolor, des iris virginica.
    >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)
```

16. Et pour finir, tapez la ligne de commandes suivante :

```
># Représentation graphique de toutes les possibilités de varaiables par variables
># variables par variables
>pairs(iris[1 :4],main="Anderson's Iris Data - 3 species",
+pch=21,bg=c("red","green3","blue")[unclass(iris$species)])
```

La fonction `pairs` reproduit tous les graphiques variables par variables possibles sur une seule fenêtre graphique et `bg` est une option graphique pour définir la couleur.

Références

[1] Frédéric BERTRAND. *Initiation au logiciel R - Master Statistique 2ème année.* http ://www-irma.u-strasbg.fr/~fbertran/enseignement/Statistique_Master2SA_2009.html