Analyse des données

Département Informatique
mardi 16 mars 2021
DI4

Travaux pratiques
Séance 1

Prise en main du logiciel R

Préambule

L'objectif de cette séance est de s'approprier l'essentiel des commandes disponibles dans le logiciel R. Celui-ci appartient à la famille des logiciels de calcul statistique au même titre que statgraphic ou SAS. Il est distribué selon la licence GNU General Public License. Les différentes versions actuellement disponibles sont téléchargeables au travers des archives partagées au sein du réseau CRAN (Comprehensive R Archive Network) dont il existe plusieurs miroirs.

Afin de faciliter l'apprentissage de la syntaxe, il vous est proposé de travailler sur un fichier de données bien connu dans le domaine de l'analyse des données à savoir les iris de Fisher. Celui-ci lui à servir à démontrer l'intérêt de l'utilisation d'outils mathématiques pour effectuer la séparation entre les différentes espèces contenues dans ce fichier. Son contenu est constitué des mesures en centimètres des variables suivantes : la longueur du sépale (Sepal.Length), la largeur du sépale (Sepal.Width), longueur du pétale (Petal.Length) et pour finir la largeur du pétale (Petal.Width) et ce pour trois espèces d'iris, connues sous les noms de : Iris setosa, versicolor et virginica.

Les objets créés lors d'une session sont enregistrés dans un fichier nommé RData situé dans le dossier de travail. L'historique de la session est conservé dans un fichier nommé. Rhistory.
Lancez le programme et vérifiez le chemin du répertoire de travail grâce à la commande :

```
getwd()
```

Pour changer, si travail avec portable perso et pas la VM, faire menu session et Set Working Director ou bien la commande:

```
Setwd (" le chemin en partant de la racine, ex: D:/ arborescence…")
```

Une façon simple de se familiariser avec R est donc de l'utiliser à partir d'un jeu de données tel que celui des iris de Fisher sur lesquelles une analyse pourra être faite. Les différentes syntaxes proposées dans la suite de ce texte sont là pour vous faire découvrir la richesse des fonctionnalités de ce logiciel en termes de calcul et de graphisme. Le déroulé pas à pas vous conduira à regarder avec soin les actions engendraient. La richesse de certaines commandes nécessitera d'utiliser l'aide en ligne ou le lien suivant : https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf
Par ailleurs, la lettre ? précédent une commande vous oriente automatiquement sur le point d'entrée dans l'aide en ligne.
Ceci dit, il vous arrivera sans doute de ne pas tout comprendre les détails mentionnés et dans ce cas le meilleur comportement à adopter est de taper le code et de voir et analyser le résultat produit. Soyez donc curieux. Un compte rendu comportant vos commentaires sera demandé à l'issue de la séance.

Petite précision, le symbole > qui se trouve à chaque début de ligne de commandes correspond au prompt bien connu appelé encore invite de commande.
Le symbole <- signifie une opération d'affectation et le symbole # vous permet d'insérer le début d'un commentaire.

Accès aux fichiers de données

Un grand nombre de jeux de données est fourni dans le logiciel. On les appelle des packages. La commande suivante vous en donnera la liste.

```
data().
```

Vérifier la présence de celui nommé iris. Pour accéder à ces données, il faut les charger en mémoire et pour cela il est nécessaire d'utiliser en autre la commande :

```
data(iris)
```

POLYTECH TOURS

Analyse des données

Département Informatique
mardi 16 mars 2021
DI4

Travaux pratiques
Séance 1

Prise en main du logiciel R

Exercice. Tapez une après l'autre les instructions données ci-après. (Attention à la casse).

```
echo = TRUE
data(iris)
iris
dim(iris)
names(iris)
iris$Species
iris$Petal.Length
```

Faites de même concernant les autres caractéristiques

Analyse du contenu du fichier

La dernière colonne du fichier contient le nom des espèces qui sont donc réparties en 3 catégories (variables) dénommées setosa, versicolor et virginica. Pour accéder à celles-ci, il faut utiliser l'instruction :

```
iris$Species
```

Quelle est la nature de ces variables et combien de modalités ont-elles ?
L'accès à celles-ci est possible grâce à la commande :

```
levels(iris$Species)
```

Il est possible d'avoir l'information complète de cette variable, grâce à l'instruction :

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

Cette information peut être obtenue en construisant un tableau incorporant le nombre d'individus par modalité. Tapez à cet effet :

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

Le logiciel R propose d'excellents graphiques. Lorsqu'une instruction graphique est exécutée, une nouvelle fenêtre "device" est ouverte. Plusieurs types de représentations liées aux variables sont possibles.
Illustrons cela en tapant les commandes suivantes et commentez :

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

Vous avez la possibilité de découper la fenêtre en passant par la syntaxe *par(mfrow = c(nl, nc))* ou *par(mfcol = c(nl, nc))* où nl définit le nombre de graphiques en ligne et nc donne le nombre de graphiques en colonne. La commande *mfrow* signifie que l'ordre d'entrée des graphiques se fera selon les lignes d'abord et *mfcol* inversement.
Exemple : Supposons que nous voulions dessiner six graphiques dans une fenêtre en deux lignes et trois colonnes.
Le premier cas, nous amènerait à obtenir le graphisme

1	2	3
4	5	6

Le second cas, nous aurions la configuration suivante :

1	3	5
2	4	6

POLYTECH TOURS

Analyse des données

Département Informatique
mardi 16 mars 2021
DI4

Travaux pratiques
Séance 1

Prise en main du logiciel R

Exercice : En utilisant la commande par(mfrow = c(1, 2)), afficher les deux vecteurs a et b suivants sous les trois formes possibles (camembert, bâtons et Cleveland) vus précédemment.

```
a <- c(1,2,3)
b <- c(4,5,6)
par(mfrow = c(1,2))
barplot(a)
barplot(b)
à compléter…
```

Tester et commentez
Afficher sous la forme (2,2) les variables qualitatives (setosa, virginica et versicolor) contenues dans le fichier iris en utilisant les deux syntaxes mfrow et mfcol et commentez.

La troisième colonne (Petal.Length) du jeu de données iris contient la longueur du pétale. Il s'agit d'une variable mesurée de type quantitative. La commande suivante permet de résumer l'information contenue dans cette variable. Commentez ce qui est fourni.

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

Il est possible de retrouver ces valeurs individuellement grâce aux commandes ci-dessous qui donne dans l'ordre la valeur min, max, la somme de toutes les longueurs, le nombre d'individus et la moyenne. Testez-les :

```
min(iris$Petal.Length)
max(iris$Petal.Length)
sum(iris$Petal.Length)
length(iris$Petal.Length)
sum(iris$Petal.Length)/length(iris$Petal.Length)
```

Il est possible d'effectuer des tris sur les fichiers avec la commande *sort*, par exemple :

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

Commentez le résultat des commandes suivantes :

```
ordlongPetal<-sort(iris$Petal.Length)
ordlongPetal
Sum(ordlongPetal)/length(ordlongPetal)
ordlongPetal[37]
(ordlongPetal[92]+ordlongPetal[37])/2
```

Obtenir une représentation type histogramme est aussi possible via la commande suivante :

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

Exercice : refaire ce travail avec les autres variables quantitatives du fichier Iris (largeur du pétale, longueur et largeur du sépale) en modifiant les paramètres.

Travaux pratiques
Séance 1

Prise en main du logiciel R

Une fois réalisés les graphiques pour chaque variable prise séparément, l'étude peut se prolonger sur la relation entre deux variables quantitatives parmi les 4 présentes, afin de savoir s'il y a un lien potentiel entre elles. On parle de croisement de deux variables ou encore étude bivariée. La représentation graphique dans ce cas est un nuage de points. Prenons l'exemple de la représentation de la longueur et la largeur du pétale pour la 150 iris contenues dans notre fichier de données. Saisissez la commande suivante et commentez en particulier le rôle du paramètre:

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

Il peut arriver que dans ce type de représentation graphique, plusieurs individus puissent être situés sur un même point. La fonction suivante permet de visualiser ces superpositions.

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

Si le nombre de points confondus est trop important, on peut être amené alors à représenter la densité des points plutôt que les points eux-mêmes. La commande suivante répond à cela :

```
library(MASS)
densite<-kde2d(iris$Petal.Length,iris$Petal.Width)
filled.contour(densite, color=topo.colors, xlab="longueur du pétale", ylab="largeur du pétale")
```

Tester et commentez en particulier le rôle des paramètres.

Exercice : refaire cette étude pour d'autres combinaisons de variables quantitatives.

Parfois, on a besoin de représenter graphiquement le lien entre deux variables de types différents (qualitative et quantitative), dans ce cas, la boite dite à moustaches rend cela possible.
Représentons par exemple la longueur des pétales en fonction de l'espèce à l'aide de la commande :

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

Commentez l'affichage obtenu et l'effet du paramètre. Penser à utiliser la commande *mfrow* et utiliser l'onglet "*zoom*".

Exercice. Choisissez les deux autres variables quantitatives, comparez, pour chacune d'elles, avec la variable espèce et commentez.

Vous avez remarqué que le nuage de points ainsi que les boites à moustaches montrent que les données morphologiques des iris semblent être liées à l'espèce. Il pourrait donc être intéressant de représenter les graphiques pour chacune des modalités : setosa, versicolor et virginica mais aussi de superposer l'information qualitative « espèce » dans le graphique des nuages de points. Voici quelques propositions de commandes répondant à ces remarques :

On refait le point sur le contenu du fichier :

```
summary (iris)
```

Nous allons afficher les histogrammes de la variable quantitative « longueur pétale » pour chacune des trois espèces.

Travaux pratiques
Séance 1

Prise en main du logiciel R

```
par(mfrow = c(2, 2))
brk = seq(from = 0, to = 8, length = 20)
hist(iris$Petal.Length, main = "Ensemble des 150 iris", xlab = "Longueur du petale", breaks = brk)
hist(iris$Petal.Length[iris$Species == "setosa"], main = "Setosa", xlab = "Longueur du petale", breaks = brk)
hist(iris$Petal.Length[iris$Species == "versicolor"], main = "Versicolor", xlab = "Longueur du petale", breaks = brk)
hist(iris$Petal.Length[iris$Species == "virginica"], main = "Virginica", xlab = "Longueur du petale", breaks = brk)
```

Exercice : refaire la même chose avec l'autre variable quantitative, observer et interpréter.

Nous allons maintenant afficher en 2D cette fois (nuage de points) la longueur et la largeur pour les 3 espèces confondues puis prises séparément. Tapez les commandes suivantes puis observer les graphes et interpréter.

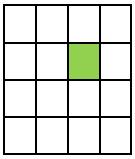
```
par(mfrow = c(2, 2))
plot(iris$Petal.Length, iris$Petal.Width, xlab = "Longueur du petale", ylab = "Largeur du petale", main = "Nuage de points", pch = 20)
plot(iris$Petal.Length[iris$Species == "setosa"], iris$Petal.Width[iris$Species =="setosa"], xlim = c(1, 6.9), ylim = c(0.1, 2.5), xlab = "",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

Pour finir ce petit tour d'horizon, nous allons afficher toutes les comparaisons possibles des 2 modalités (longueurs, largeurs) entre toutes les espèces. Tableau 4x4.

```
#nouveau tableau sans colonne Species
new_data <- unclass(iris$Species)
pairs(iris[1:4], main = "Les iris de Fisher -- 3 especes", pch = c(21, 25, 24)[new_data], bg = c("red", "green3", "blue")[new_data], las = 2, gap = 0, labels = c("Longueur\nSepale", "Largeur\nSepale", "Longueur\nPetale","Largeur\nPetale"))
```

Avant de saisir cette commande, avec l'aide en ligne, donner le rôle de chaque paramètre contenu dans cette commande, une fois fait, l'exécuter. Puis observer et en faire l'interprétation.
Question: que représente le graphe situé dans la partie verte?



Fin ! ☺