

L3 Mathématiques, Informatique Appliquées
aux Sciences Humaines et Sociales

MIC0503V - Statistique
Utilisation du logiciel de statistique R

2017-2018

Pascal Sarda

1 Présentation - Téléchargement

Le logiciel R est un langage de programmation dédié aux traitements de données. R est un logiciel libre développé à la base comme un clone du logiciel S-Plus, créé par Ross Ihaka et Robert Gentleman à l'Université d'Auckland, Nouvelle Zélande. Il s'agit d'un des logiciels de statistique les plus complets, disponible pour Microsoft Windows, Macintosh et de nombreux systèmes de type Unix. R est écrit en C, C++, FORTRAN et Java et est orienté programmation objet.


R est distribué gratuitement sous les termes de la "GNU", *General Public Licence Version 2*, Juin 1991. Les informations sur le logiciel R ainsi que les différentes distributions, bibliothèques et aides se trouvent à l'adresse <http://www.r-project.org/>. Un système de sites miroirs a été mis en place, permettant le téléchargement via un miroir proche de l'endroit où l'on se trouve.

Pour installer le logiciel R, il faut se rendre à l'adresse <http://www.r-project.org/> puis dans le menu Download, cliquer sur CRAN. On choisit un site miroir : pour la France, il y a 5 ou 6 sites, le plus proche de Toulouse se trouvant à Montpellier (il y avait auparavant un site miroir au centre de calcul de l'Université Toulouse 3 mais ce site a été supprimé). On choisit dans la fenêtre de téléchargement la version correspondant à son système d'exploitation. On réalise ainsi l'installation de base du logiciel comprenant les **packages** les plus courants. On a parfois besoin de packages spécifiques qu'il faut télécharger en reprenant la procédure et en cliquant sur packages dans la fenêtre de téléchargement.

2 Utilisation de R

2.1 Opérations de base

Lancer - Quitter

Pour lancer R double-cliquer sur l'icône , une console R avec un texte qui s'affiche suivi d'un prompt > indiquant que le logiciel est prêt à travailler. Pour quitter R taper la commande :

```
> q()
```

Consultation de l'aide en ligne

```
> help.start()
```

dans une fenêtre pour une fonction particulière

```
> ?read.table
```

ou

```
> help(read.table)
```

Lecture/écriture de données - fichier .csv

changer le répertoire de travail. Pour lire ou enregistrer un fichier à partir de R, il faut lui préciser le répertoire de travail. Il est plus pratique de changer le répertoire de lecture en début de session (ou si nécessaire à tout moment en cours de session). Pour cela, à partir de la console R, aller dans le menu *Changer de Répertoire de Travail...*

lecture de données. Les fichiers de données sont la plupart du temps obtenues à l'aide d'un tableur, Excel, LibreOffice, SPSS, SAS, ... dans un format comme .xls. Il existe des packages spécifiques permettant de lire les fichiers à ces formats (par exemple pour le format .xls). Une autre façon de procéder, évitant le chargement de packages additionnels, est de créer des fichiers de données au format fichier.csv. Pour cela, il suffit d'ouvrir le fichier sous Excel, LibreOffice, ... puis d'en faire une sauvegarde au format .csv. Si besoin, il faut préciser le *séparateur* ";"

On lit ensuite le fichier sous R grâce à l'instruction :

```
> read.csv2("fichier.csv")
```

Pour enregistrer un tableau de données `x` (`data.frame`) :

```
> write.csv(x,"fichier.csv")
```

Affectation

```
> n<-28  
> m=1973
```

Affichage

```
> n  
[1] 28
```

```
> m  
[1] 1973
```

Suppression

```
> rm(n)  
> rm(n,m)
```

2.2 Les objets

R travaille avec des objets, c'est-à-dire des espaces dans lesquels on peut stocker tout ce qui nous intéresse. Un objet se caractérise par un nom, une classe, un mode, une taille.

Principales classes d'objets :

vector
factor
matrix
list
data.frame

Modes :

numeric (nombre réel)
character (chaîne de caractères, entre guillemets)
list (liste d'objets)
logical (booléen)
Autres modes : fonction, expression, formula

- Un vecteur (vector) contient des valeurs ayant le même mode ; il est constitué d'une colonne
- Un facteur (factor) est un vecteur particulier permettant de traiter des variables qualitatives (ou catégorielles)
- Une matrice (matrix) est un tableau pouvant avoir plusieurs colonnes dont les valeurs ont le même mode
- Une liste (list) est un objet permettant de stocker des objets de différents modes et de différentes longueur
- Un data.frame est un tableau dont les composantes ont la même longueur et pouvant être de modes différents (liste particulière)

Exemple. Les célèbres données *iris* ont été collectées par Edgar Anderson. Ce sont les mesures en centimètres des variables suivantes : longueur du sépale (Sepal.Length), largeur du sépale (Sepal.Width), longueur du pétale (Petal.Length) et largeur du pétale (Petal.Width) pour trois espèces d'iris : *Iris setosa*, *I. versicolor* et *I. virginica*.

```
> class(iris)

[1] "data.frame"

> mode(iris)

[1] "list"

> names(iris)

[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
[5] "Species"

> length(iris)

[1] 5
```

indique que *iris* possède 5 colonnes.

```
> dim(iris)

[1] 150  5
```

indique que *iris* possède 150 lignes et 5 colonnes.

Conversion de mode. On peut convertir le mode d'un objet à l'aide des fonctions :

```
> as.character
> as.list
> as.logical
> as.numeric
> as.factor
```

et tester son mode avec :

```
> is.character
> is.list
> is.logical
> is.numeric
> is.factor
```

2.3 Graphiques

La principale fonction pour obtenir un graphique est la fonction *plot*. D'autres fonctions permettent d'obtenir différents types de graphiques. Voici quelques exemples :

Diagramme en bâtons. Le tableau suivant indique la fréquentation et le sexe de 28 étudiants.

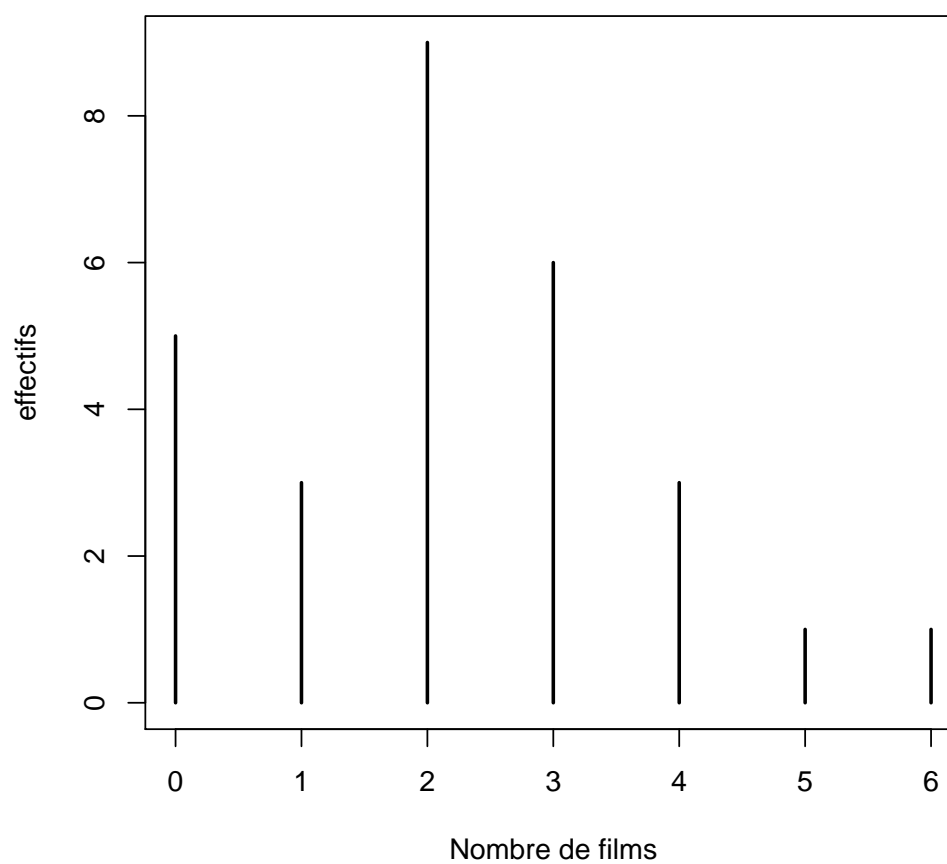
	A	B	C
1	Etudiants	Nombre de films	Sexe
2	1		1 F
3	2		3 M
4	3		4 M
5	4		2 F
6	5		1 F
7	6		4 F
8	7		0 M
9	8		2 M
10	9		2 F
11	10		0 F
12	11		4 F
13	12		5 M
14	13		2 F
15	14		1 M
16	15		2 F
17	16		2 M
18	17		2 F
19	18		3 F
20	19		3 M
21	20		3 M
22	21		6 M
23	22		3 M
24	23		0 F
25	24		0 F
26	25		0 M
27	26		2 M
28	27		3 M
29	28		2 F
30			

Récupération des données

```
> films<-read.csv2("films.csv")
```

Diagramme en bâtons de la variable nombre de films

```
> plot(table(films$Nombre.de.films),xlab="Nombre de films",ylab="effectifs")
```



Histogramme de la variable `Iris$Petal.Length`

```
> hist(iris$Petal.Length,freq=FALSE)
```

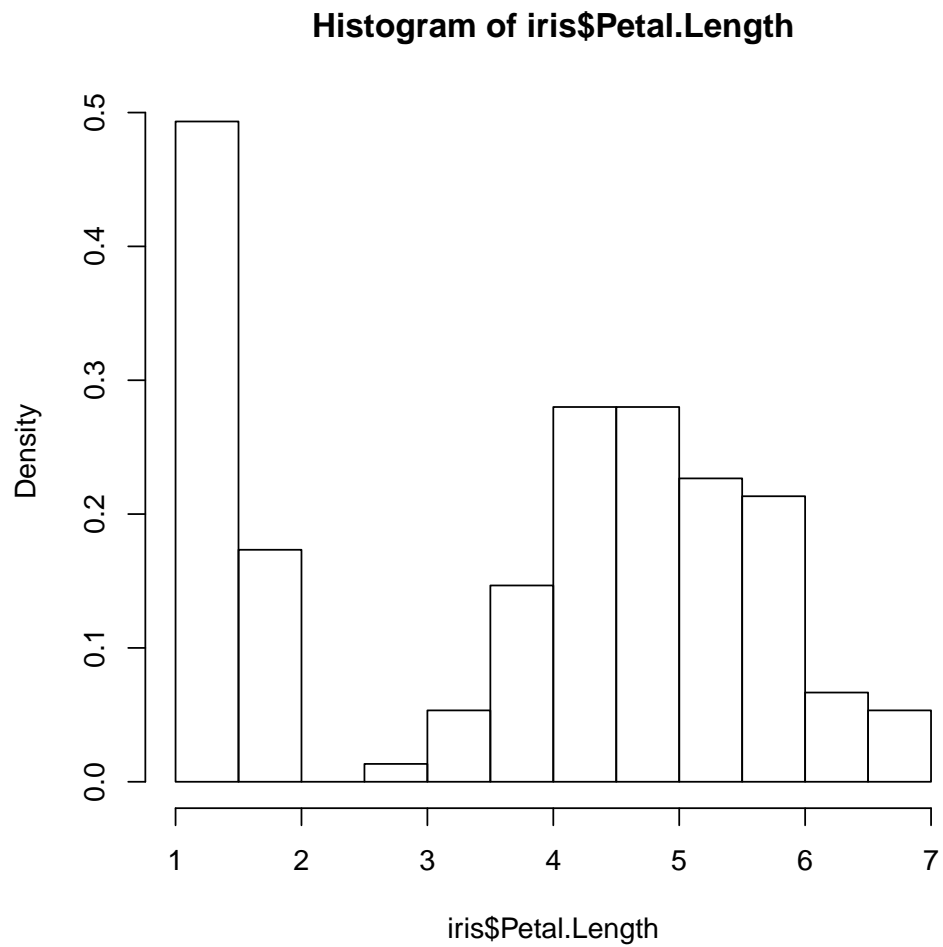
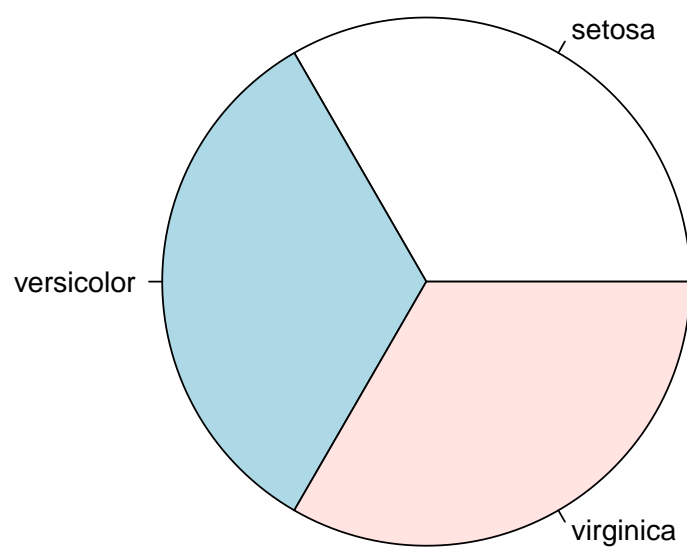


Diagramme en secteurs de la variable `Iris$Species`

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



Bibliographie

Adler, J. *R l'essentiel*, Pearson Eds.