

Se former au logiciel R : initiation et perfectionnement

François Rebaudo

2018-07-27

Contents

1	Préambule	5
2	Remerciements	7
3	Licence	9
4	Introduction	11
4.1	Pourquoi se former à R	11
4.2	Ce livre	11
4.3	Lectures complémentaires en français	11
I	Concepts de base	13
5	Premiers pas	15
5.1	Installation de R	15
5.2	R comme calculatrice	15
5.3	La notion d'objet	22
5.4	Les scripts	23
6	Choisir un environnement de développement	27
6.1	Editeurs de texte et environnement de développement	27
6.2	RStudio	27
6.3	Notepad++ avec Npp2R	29
6.4	Geany (pour Linux, Mac OSX et Windows)	33
6.5	Autres solutions	34
6.6	Conclusion	34
7	Les types de données	37
7.1	Le type <code>numeric</code>	37
7.2	Le type <code>character</code>	39
7.3	Le type <code>factor</code>	40
7.4	Le type <code>logical</code>	41
7.5	A propos de <code>NA</code>	42
7.6	Conclusion	42
8	Les conteneurs de données	43
8.1	Le conteneur <code>vector</code>	43
8.2	Le conteneur <code>list</code>	51
8.3	Le conteneur <code>data.frame</code>	62
8.4	Le conteneur <code>matrix</code>	67
8.5	Le conteneur <code>array</code>	72
8.6	Conclusion	74

9 Les fonctions	77
9.1 Qu'est-ce qu'une fonction	77
9.2 Les fonctions les plus courantes	77
9.3 Autres fonctions utiles	94
9.4 Quelques exercices	98
9.5 Ecrire une fonction	100
9.6 Conclusion	100
10 Importer et exporter des données	101
10.1 Lire des données depuis un fichier tableur	101
10.2 Sauver des données pour R	101
10.3 Exporter des données	101
11 Les boucles	103
11.1 Pourquoi faire des boucles	103
11.2 La boucle if	103
11.3 La boucle switch	103
11.4 La boucle for	103
11.5 La boucle while	103
11.6 repeat, next, break, stop	103
11.7 Les boucles de la famille apply	104
II Les graphiques	105
12 Graphiques simples	107
12.1 plot	107
12.2 hist	108
12.3 barplot	108
12.4 boxplot	108
12.5 image et contour	108
13 La gestion des couleurs	109
13.1 colors()	109
13.2 RGB	109
13.3 Paletas	109
14 Graphiques composés	111
14.1 mfrow	111
14.2 layout	111
15 Manipuler les graphiques	113
15.1 Inkscape	113
15.2 The Gimp	113
III Statistiques	115
16 Statistiques descriptives	117
IV Etude de cas	119
17 Analyser des données de datalogger de température	121
18 Obtenir le numéro WOS d'un article scientifique à partir de son numéro DOI	123

Chapter 1

Préambule

Ce livre est incomplet pour le moment et vous visualisez sa version préliminaire. Si vous avez des commentaires, des suggestions ou si vous identifiez des erreurs, n'hésitez pas à m'envoyer un mail (francois.rebaudo@ird.fr¹), ou si vous connaissez GitHub sur le site du projet (https://github.com/frareb/myRBook_FR). Ce livre est également disponible en espagnol (<http://myrbooksp.netlify.com/>).

Dernières modifications:

27/07/2018

- Etude de cas : Obtenir le numéro WOS d'un article scientifique à partir de son numéro DOI

25/07/2018

- Les fonctions (partie 1/3) : les fonctions les plus courantes

17/07/2018

- Les conteneurs de données (partie 4/5) : matrix
- Les conteneurs de données (partie 5/5) : array

13/07/2018

- mise en ligne du contenu en français sur la base du livre en espagnol
- Les conteneurs de données (partie 3/5) : data.frame

¹<mailto:francois.rebaudo@ird.fr>

Chapter 2

Remerciements

Je remercie tous les contributeurs qui ont participé à améliorer ce livre par leurs conseils, leurs suggestions de modifications et leurs corrections (par ordre alphabétique) :

Contributeurs :

Camila Benavides Frias (Bolivia)

Susi Loza Herrera (Bolivia)

Estefania Quenta Herrera (Bolivia)

Les versions gitbook, html et epub de ce livre utilisent les icônes open source de Font Awesome (<https://fontawesome.com>). La version PDF utilise les icônes issues du projet Tango disponibles depuis opencart (<https://opencart.org/>). Ce livre a été écrit avec le package R bookdown (<https://bookdown.org/>). Le code source est disponible sur GitHub (https://github.com/frareb/myRBook_FR). La compilation utilise Travis CI (<https://travis-ci.org>). La version en ligne est hébergée et mise à jour grâce à Netlify (<http://myrbookfr.netlify.com/>).

Chapter 3

Licence

Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 3.0 France (CC BY-NC-ND 3.0 FR ; <https://creativecommons.org/licenses/by-nc-nd/3.0/fr/>)

C'est un résumé (et non pas un substitut) de la licence.

Vous êtes autorisé à :

- Partager — copier, distribuer et communiquer le matériel par tous moyens et sous tous formats.
- L'Offrant ne peut retirer les autorisations concédées par la licence tant que vous appliquez les termes de cette licence.

Selon les conditions suivantes :

- Attribution — Vous devez créditer l'Œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Œuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Œuvre.
- Pas d'Utilisation Commerciale — Vous n'êtes pas autorisé à faire un usage commercial de cette Œuvre, tout ou partie du matériel la composant.
- Pas de modifications — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Œuvre originale, vous n'êtes pas autorisé à distribuer ou mettre à disposition l'Œuvre modifiée.
- Pas de restrictions complémentaires — Vous n'êtes pas autorisé à appliquer des conditions légales ou des mesures techniques qui restreindraient légalement autrui à utiliser l'Œuvre dans les conditions décrites par la licence.

Notes :

Vous n'êtes pas dans l'obligation de respecter la licence pour les éléments ou matériel appartenant au domaine public ou dans le cas où l'utilisation que vous souhaitez faire est couverte par une exception. Aucune garantie n'est donnée. Il se peut que la licence ne vous donne pas toutes les permissions nécessaires pour votre utilisation. Par exemple, certains droits comme les droits moraux, le droit des données personnelles et le droit à l'image sont susceptibles de limiter votre utilisation.

Chapter 4

Introduction

4.1 Pourquoi se former à R

manipuler ses données statistiques et nombreux packages disponibles communauté d'utilisateurs graphiques de qualité transparence scientifique et reproductibilité des résultats

4.2 Ce livre

L'objectif de ce livre est de fournir aux étudiants et aux personnes souhaitant s'initier à R une base solide pour ensuite mettre en oeuvre leurs propres projets scientifiques et la valorisation de leurs résultats. Il existe de nombreux livres dédiés à R, mais aucun ne couvre les éléments de base de ce langage dans un objectif de rendre les résultats scientifiques publiables et reproductibles. De manière générale ce livre s'adresse à toute la communauté scientifique et en particulier à celle intéressée par les sciences du vivant, et les nombreux exemples de ce livre s'appuieront sur des études en agronomie et en écologie.

initiation aux bases de R : pour tous ! perfectionnement en français contrairement à d'autres livres, importance de la lisibilité du code, des standards, orientation recherche "outil pour le scientifique"

4.3 Lectures complémentaires en français

- R pour les débutants, Emmanuel Paradis (https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf)
- Introduction à la programmation avec R, Vincent Goulet (https://cran.r-project.org/doc/contrib/Goulet_introduction_programmation_R.pdf)

Part I

Concepts de base

Chapter 5

Premiers pas

5.1 Installation de R

Le programme permettant l'installation du logiciel R peut être téléchargé depuis le site web de R : <https://www.r-project.org/>. Sur le site de R il faut au préalable choisir un miroir CRAN (serveur depuis lequel télécharger R ; sauf cas particulier le plus proche de sa localisation géographique), puis télécharger le fichier *base*. Les utilisateurs de Linux pourront préférer un `sudo apt-get install r-base`.



Le logiciel R peut être téléchargé depuis de nombreux serveurs du CRAN (Comprehensive R Archive Network) à travers le monde. Ces serveurs s'appellent des miroirs. Le choix du miroir est manuel. Les informations complémentaires comme cette note seront toujours représentées avec ce pictogramme *information*.

5.2 R comme calculatrice

Une fois le programme lancé, une fenêtre apparaît dont l'aspect peut varier en fonction de votre système d'exploitation (Figure 5.1). Cette fenêtre est dénommée la *console*.

La console correspond à l'interface où va être interprété le code, c'est à dire à l'endroit où le code va être transformé en langage machine, exécuter par l'ordinateur, puis retransmis sous une forme lisible par des humains. Cela correspond à l'écran d'affichage d'une calculatrice (Figure 5.2). C'est de cette manière que R va être utilisé dans la suite de cette section.



Tout au long de ce livre, les exemples de code R apparaîtront sur fond en gris. Ils peuvent être copiés et collés directement dans la console, bien qu'il soit préférable de reproduire soit même les exemples dans la console (ou plus tard dans les scripts). Le résultat de ce qui est envoyé dans la console apparaîtra également sur fond en gris avec `##` devant le code afin de bien faire la distinction entre le code et le résultat du code.

5.2.1 Les opérateurs arithmétiques

```
5 + 5
```

```
## [1] 10
```

Si nous écrivons `5 + 5` dans la console puis **Entrée**, le résultat apparaît précédé du chiffre `[1]` entre crochets. Ce chiffre correspond au numéro du résultat (dans notre cas, il n'y a qu'un seul résultat ; nous reviendrons sur cet aspect plus tard). Nous

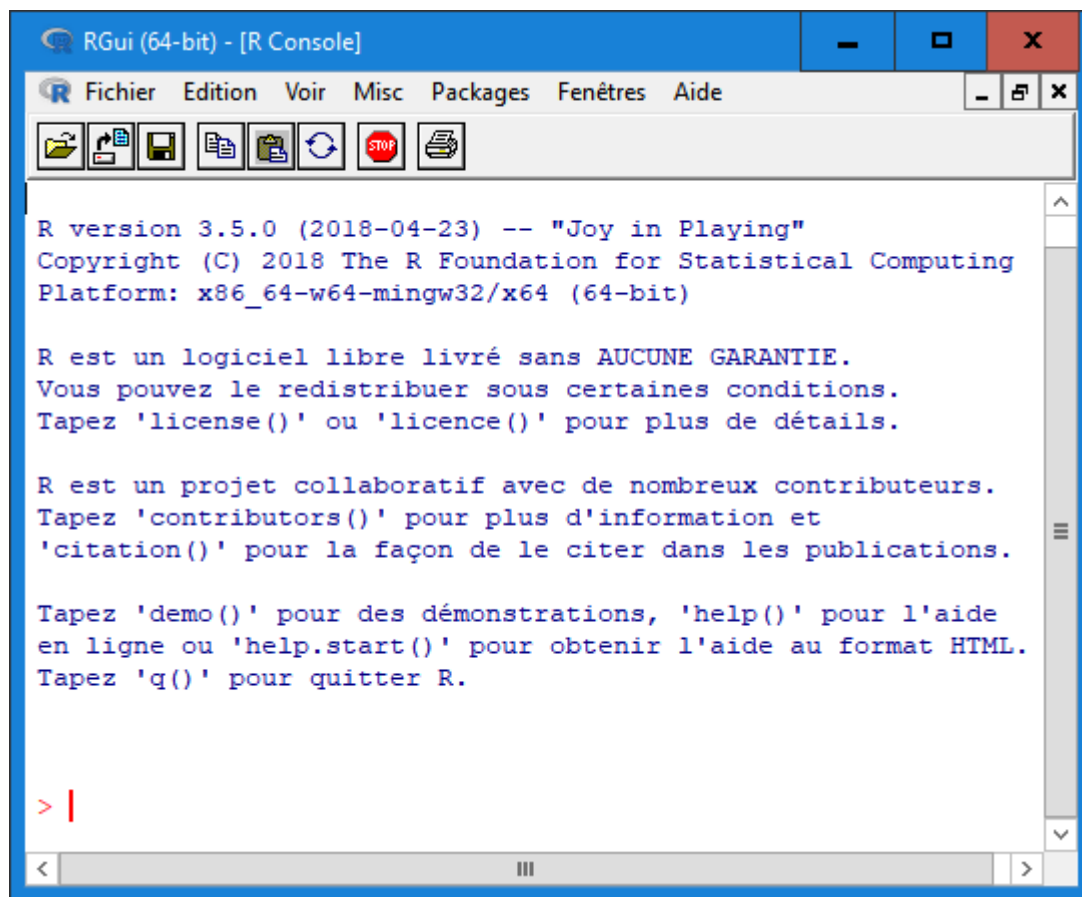


Figure 5.1: Capture d'écran de la console R sous Windows.

Table 5.1: Opérateurs arithmétiques.

Label	Opérateur
Addition	+
Soustraction	-
Multiplication	*
Division	/
Puissance	^
Modulo	%%
Quotien Décimal	%/%

pouvons également noter dans cet exemple l'utilisation d'espaces avant et après le signe +. Ces espaces ne sont pas nécessaires mais permettent au code d'être plus lisible par les humains (i.e., plus agréable à lire pour nous comme pour les personnes avec qui nous serons amenés à partager notre code). Les opérateurs arithmétiques disponibles sous R sont résumés dans la table 5.1.

Classiquement, les multiplications et les divisions sont prioritaires sur les additions et les soustractions. Au besoin nous pouvons utiliser des parenthèses.

```
5 + 5 * 2
```

```
## [1] 15
```




Figure 5.2: Capture d'écran de la console R sous Windows avec la calculatrice Windows.

```
(5 + 5) * 2
```

```
## [1] 20
```

L'opérateur modulo correspond au reste de la division euclidienne. Il est souvent utilisé en informatique par exemple pour savoir si un nombre est pair ou impair (un nombre modulo 2 va renvoyer 1 si il est impair et 0 si il est pair).

```
451 %% 2
```

```
## [1] 1
```

```
288 %% 2
```

```
## [1] 0
```

```
(5 + 5 * 2) %% 2
```

```
## [1] 1
```

```
((5 + 5) * 2) %% 2
```

```
## [1] 0
```

R intègre également certaines constantes dont π . Par ailleurs le signe infini est représenté par `Inf`.

Table 5.2: Opérateurs de comparaison.

Label	Operator
plus petit que	<
plus grand que	>
plus petit ou égal à	<=
plus grand ou égal à	>=
égal à	==
différent de	!=

```
pi
```

```
## [1] 3.141593
```

```
pi * 5^2
```

```
## [1] 78.53982
```

```
1/0
```

```
## [1] Inf
```



le *style* du code est important car le code est destiné à être lisible par nous plus tard et par d'autres personnes de manière générale. Pour avoir un style lisible il est recommandé de mettre des espaces avant et après les opérateurs arithmétiques. Les informations concernant le *style* seront toujours représentées avec ce pictogramme afin qu'elles soient facilement identifiables.

5.2.2 Les opérateurs de comparaison

R est cependant bien plus qu'une simple calculatrice puisque'il permet un autre type d'opérateurs : les opérateurs de comparaison. Ils servent comme leur nom l'indique à *comparer* des valeurs entre elles (Table 5.2).

Par exemple si nous voulons savoir si un chiffre est plus grand qu'un autre, nous pouvons écrire :

```
5 > 3
```

```
## [1] TRUE
```

R renvoie la valeur TRUE si la comparasion est vraie et FALSE si la comparaison est fausse.

```
5 > 3
```

```
## [1] TRUE
```

```
2 < 1.5
```

```
## [1] FALSE
```

```
2 <= 2
```

```
## [1] TRUE
```

```
3.2 >= 1.5
```

```
## [1] TRUE
```

Nous pouvons combiner les opérateurs arithmétiques avec les opérateurs de comparaison.

```
(5 + 8) > (3 * 45/2)
```

```
## [1] FALSE
```



Dans la comparaison $(5 + 8) > (3 * 45/2)$ les parenthèses ne sont pas nécessaires mais elles permettent au code d'être plus facile à lire.

Un opérateur de comparaison particulier est *égal à*. Nous verrons dans la section suivante que le signe $=$ est réservé à un autre usage : il permet d'affecter une valeur à un objet. L'opérateur de comparaison *égal à* doit donc être différent, c'est pour cela que R utilise $==$.

```
42 == 53
```

```
## [1] FALSE
```

```
58 == 58
```

```
## [1] TRUE
```

Un autre opérateur particulier est *différent de*. Il est utilisé avec *un point d'interrogation* suivi de *égal*, $!=$. Cet opérateur permet de d'obtenir la réponse inverse à $==$.

```
42 == 53
```

```
## [1] FALSE
```

```
42 != 53
```

```
## [1] TRUE
```

```
(3 + 2) != 5
```

```
## [1] FALSE
```

```
10/2 == 5
```

```
## [1] TRUE
```

R utilise TRUE et FALSE qui sont aussi des valeurs qui peuvent être testées avec les opérateurs de comparaison. Mais R attribue également une valeur à TRUE et FALSE :

```
TRUE == TRUE
```

```
## [1] TRUE
```

```
TRUE > FALSE
```

```
## [1] TRUE
```

```
1 == TRUE
```

```
## [1] TRUE
```

```
0 == FALSE
```

```
## [1] TRUE
```

```
TRUE + 1
```

```
## [1] 2
```

```
FALSE + 1
```

```
## [1] 1
```

```
(FALSE + 1) == TRUE
```

```
## [1] TRUE
```

La valeur de TRUE est de 1 et la valeur de FALSE est de 0. Nous verrons plus tard comment utiliser cette information dans les prochains chapitres.

R est aussi un langage relativement permissif, cela veut dire qu'il admet une certaine flexibilité dans la manière de rédiger le code. Débattre du bien fondé de cette flexibilité sort du cadre de ce livre mais nous pourrions trouver dans du code R sur Internet ou dans d'autres ouvrages le raccourcis T pour TRUE et F pour FALSE.

```
T == TRUE
```

```
## [1] TRUE
```

```
F == FALSE
```

```
## [1] TRUE
```

```
T == 1
```

```
## [1] TRUE
```

```
F == 0
```

```
## [1] TRUE
```

```
(F + 1) == TRUE
```

```
## [1] TRUE
```

Bien que cette façon de se référer à TRUE et FALSE par T et F soit assez répandue, dans ce livre nous utiliserons toujours TRUE et FALSE afin que le code soit plus facile à lire. Encore une fois l'objectif d'un code est de non seulement être fonctionnel mais aussi d'être facile à lire et à relire.

Table 5.3: Opérateurs logiques.

Label	Operador
n'est pas	!
et	&
ou	
ou exclusif	xor()

5.2.3 Les opérateurs logiques

Il existe un dernier type d'opérateur, les opérateurs logiques. Ils sont utiles pour combiner des opérateurs de comparaison (Table 5.3).

```
!TRUE
```

```
## [1] FALSE
```

```
!FALSE
```

```
## [1] TRUE
```

```
((3 + 2) == 5) & ((3 + 3) == 5)
```

```
## [1] FALSE
```

```
((3 + 2) == 5) & ((3 + 3) == 6)
```

```
## [1] TRUE
```

```
(3 < 5) & (5 < 5)
```

```
## [1] FALSE
```

```
(3 < 5) & (5 <= 5)
```

```
## [1] TRUE
```

L'opérateur logique `xor()` correspond à un *ou exclusif*. C'est à dire que l'un des deux **arguments** de la **fonction** `xor()` doit être vrai, mais pas les deux. Nous reviendrons plus tard sur les **fonctions** et leurs **arguments**, mais retenons que l'on identifie une fonction par ses parenthèses qui contiennent des arguments séparés par des virgules.

```
xor((3 + 2) == 5, (3 + 3) == 6)
```

```
## [1] FALSE
```

```
xor((3 + 2) == 5, (3 + 2) == 6)
```

```
## [1] TRUE
```

```
xor((3 + 3) == 5, (3 + 2) == 6)
```

```
## [1] FALSE
```

```
xor((3 + 3) == 5, (3 + 3) == 6)
```

```
## [1] TRUE
```



Il est recommandé que les virgules , soient suivies par un espace afin que le code soit plus agréable à lire.

5.2.4 Aide sur les opérateurs

Le fichier d'aide en anglais sur les opérateurs arithmétiques peut être obtenue avec la commande ? '+' celui sur les opérateurs de comparaison avec la commande ? '==' et celui sur les opérateurs logiques avec la commande ? '&'.

5.3 La notion d'objet

Un aspect important de la programmation avec R, mais aussi de la programmation en général est la notion d'objet. Comme indiqué sur la page web de wikipedia ([https://fr.wikipedia.org/wiki/Objet_\(informatique\)](https://fr.wikipedia.org/wiki/Objet_(informatique))), en informatique, un objet est un *conteneur*, c'est à dire quelque chose qui va contenir de l'information. L'information contenue dans un objet peut être très diverse, mais pour le moment nous allons contenir dans un objet le chiffre 5. Pour ce faire (et pour pouvoir le réutiliser par la suite), il nous faut donner un nom à notre objet. Avec R le nom des objets ne doit pas comprendre de caractères spéciaux comme `^$?|+(){}[]`, ne doit pas commencer par un chiffre ni contenir d'espaces. Le nom de l'objet doit être représentatif de ce qu'il contient, tout en étant ni trop court ni trop long. Imaginons que notre chiffre 5 corresponde au nombre de répétitions d'une expérience. Nous voudrions lui donner un nom faisant référence à *nombre* et à *répétition*, que nous pourrions réduire à *nbr* et *rep*, respectivement. Il existe plusieurs possibilités qui sont toutes assez répandues sous R :

- la séparation au moyen du caractère *tiret bas* : `nbr_rep`
- la séparation au moyen du caractère *point* : `nbr.rep`
- l'utilisation de lettres minuscules : `nbrrep`
- le style *lowerCamelCase* consistant en un premier mot en minuscules et des suivants avec une majuscule : `nbrRep`
- le style *UpperCamelCase* consistant à mettre une majuscule au début de chacun des mots : `NbrRep`

Toutes ces formes de nommer un objet sont équivalentes. Dans ce livre nous utiliserons le style *lowerCamelCase*. De manière générale il faut éviter les noms trop longs comme `leNombreDeRepetitions` ou trop courts comme `nR`, et les noms ne permettant pas d'identifier le contenu comme `maVariable` ou `monChiffre`, mais aussi `a` ou `b...`



Il existe différentes façons de définir un nom pour les objets que nous allons créer avec R. Dans ce livre il est utilisé le style *lowerCamelCase*. L'important n'est pas le choix du style mais la consistance dans son choix. L'objectif est d'avoir un code fonctionnel mais également un code facile et agréable à lire.

Maintenant que nous avons choisi un nom pour notre objet, il faut le créer et faire comprendre à R que notre objet doit contenir le chiffre 5. Il existe trois façons de créer un objet sous R :

- avec le signe `<-`
- avec le signe `=`
- avec le signe `->`

```
nbrRep <- 5
nbrRep = 5
5 -> nbrRep
```

Dans ce livre nous utiliserons toujours la forme `<-` par souci de consistance et aussi parce que c'est la forme la plus répandue.

```
nbrRep <- 5
```

Nous venons de créer un objet `nbrRep` et de lui affecter la valeur 5. Cet objet est désormais disponible dans notre environnement de calcul et peut donc être utilisé. Voici quelques exemples :

```
nbrRep + 2
```

```
## [1] 7
```

```
nbrRep * 5 - 45/56
```

```
## [1] 24.19643
```

```
pi * nbrRep^2
```

```
## [1] 78.53982
```

La valeur associée à notre objet `nbrRep` peut être modifiée de la même manière que lors de sa création :

```
nbrRep <- 5
nbrRep + 2
```

```
## [1] 7
```

```
nbrRep <- 10
nbrRep + 2
```

```
## [1] 12
```

```
nbrRep <- 5 * 2 + 7/3
nbrRep + 2
```

```
## [1] 14.33333
```

L'utilisation des objets prend tout son sens lorsque nous avons des opérations complexes à réaliser et rend le code plus agréable à lire et à comprendre.

```
(5 + 9^2 - 1/18) / (32 * 45/8 + 3)
```

```
## [1] 0.4696418
```

```
terme01 <- 5 + 9^2 - 1/18
terme02 <- 32 * 45/8 + 3
terme01 / terme02
```

```
## [1] 0.4696418
```

5.4 Les scripts

R est un langage de programmation souvent dénommé *langage de script*. Cela fait référence au fait que la plupart des utilisateurs vont écrire des petits bouts de code plutôt que des programmes entiers. R peut être utilisé comme une simple calculatrice, et dans

ce cas il ne sera pas nécessaire de conserver un historique des opérations qui ont été réalisées. Mais si les opérations à réaliser sont longues et complexes, il peut devenir nécessaire de pouvoir sauvegarder ce qui a été fait à un moment donné pour pouvoir poursuivre plus tard. Le fichier dans lequel seront conservées les opérations constitue ce que l'on appelle communément le script. Un script est donc un fichier contenant une succession d'informations compréhensibles par R et qu'il est possible d'exécuter.

5.4.1 Créer un script et le documenter

Pour ouvrir un nouveau script il suffit de créer un fichier texte vide qui sera édité par un éditeur de texte comme le bloc note sous Windows ou Mac OS, ou encore Gedit ou même nano sous Linux. Par convention ce fichier prend l'extension ".r" ou plus souvent ".R". C'est cette dernière convention qui sera utilisée dans ce livre. Depuis l'interface graphique de R il est possible de créer un nouveau script sous Mac OS et Windows via *fichier* puis *nouveau script* et *enregistrer sous*. Tout comme le nom des objets, le nom du script est important pour que nous puissions facilement identifier son contenu. Par exemple nous pourrions créer un fichier `formRConceptsBase.R` contenant les objets que nous venons de créer et les calculs effectués. Mais même avec des noms de variables et un nom de fichier bien définis, il sera difficile de se rappeler le sens de ce fichier sans une documentation accompagnant ce script. Pour documenter un script nous allons utiliser des *commentaires*. Les commentaires sont des éléments qui seront identifiés par R comme tel et qui ne seront pas exécutés. Pour spécifier à R que nous allons faire un commentaire, il faut utiliser le caractère octothorpe (croisillon) `#`. Les commentaires peuvent être insérés sur une nouvelle ligne ou en fin de ligne.

```
# creation objet nombre de repetitions
nbrRep <- 5 # commentaire de fin de ligne
```

Les commentaires peuvent aussi être utilisés pour qu'une ligne ne soit plus exécutée.

```
nbrRep <- 5
# nbrRep + 5
```

Pour en revenir à la documentation du script, il est recommandé de commencer chacun de ses scripts par une brève description de son contenu, puis lorsque le script devient long, de le structurer en différentes parties pour faciliter sa lecture.

```
# -----
# Voici un script pour acquérir les concepts de base
# avec R
# date de création : 25/06/2018
# auteur : François Rebaudo
# -----

# [1] création de l'objet nombre de répétitions
# -----

nbrRep <- 5

# [2] calculs simples
# -----

pi * nbrRep^2

## [1] 78.53982
```



Pour aller plus loin sur le style de code, un guide complet de recommandations est disponible en ligne (en anglais ; <http://style.tidyverse.org/>).

5.4.2 Exécuter un script

Depuis que nous avons un script, nous ne travaillons plus directement dans la console. Or seule la console est capable d'interpréter le code R et de nous renvoyer les résultats que nous souhaitons obtenir. Pour l'instant la technique la plus simple consiste à copier-coller les lignes que nous souhaitons exécuter depuis notre script vers la console. A partir de maintenant nous n'allons plus utiliser les éditeurs de texte comme le bloc note mais des éditeurs spécialisés pour la confection de scripts R. C'est l'objet du chapitre suivant.

Chapter 6

Choisir un environnement de développement

6.1 Editeurs de texte et environnement de développement

Il existe de très nombreux éditeurs de texte, le chapitre précédent a permis d'en introduire quelque uns parmi les plus simples comme le bloc note de Windows. Rapidement les limites de ces éditeurs ont rendu la tâche d'écrire un script fastidieuse. En effet, même en structurant son script avec des commentaires, il reste difficile de se repérer dans celui-ci. C'est là qu'interviennent les éditeurs de texte spécialisés qui vont permettre une écriture et une lecture agréable et simplifiée. L'éditeur de texte pour R certainement le plus répandu est Rstudio, mais il en existe bien d'autres. Faire une liste exhaustive de toutes les solutions disponibles sort du cadre de ce livre, ainsi nous nous focaliserons sur les trois solutions que j'utilise au quotidien que sont **Notepad++**, **Rstudio**, et **Geany**.

6.2 RStudio

6.2.1 Installer RStudio

Le programme pour installer Rstudio se retrouve dans la partie *Products* du site web de Rstudio (<https://www.rstudio.com/>). Nous allons installer RStudio pour un usage local (sur notre ordinateur), donc la version qui nous intéresse est *Desktop*. Nous allons utiliser la version *Open Source* qui est gratuite. Ensuite il nous suffit de sélectionner la version qui correspond à notre système d'exploitation, de télécharger le fichier correspondant et de l'exécuter pour lancer l'installation. Nous pouvons conserver les options par défaut tout au long de l'installation.

6.2.2 Un script avec RStudio

Nous pouvons alors ouvrir RStudio. Lors de la première ouverture, l'interface est divisée en deux avec à gauche la console R que nous avons vu au chapitre précédent (Figure 6.2). Pour ouvrir un nouveau script, nous allons dans le menu *File*, *New File*, *R script*. Par défaut ce fichier a comme nom *Untitled1*. Nous avons vu au chapitre précédent l'importance de donner un nom pertinent à nos scripts, c'est pourquoi nous allons le renommer *selecEnvDev.R*, dans le menu *File*, avec l'option *Save As....* Nous



Figure 6.1: Logo RStudio.

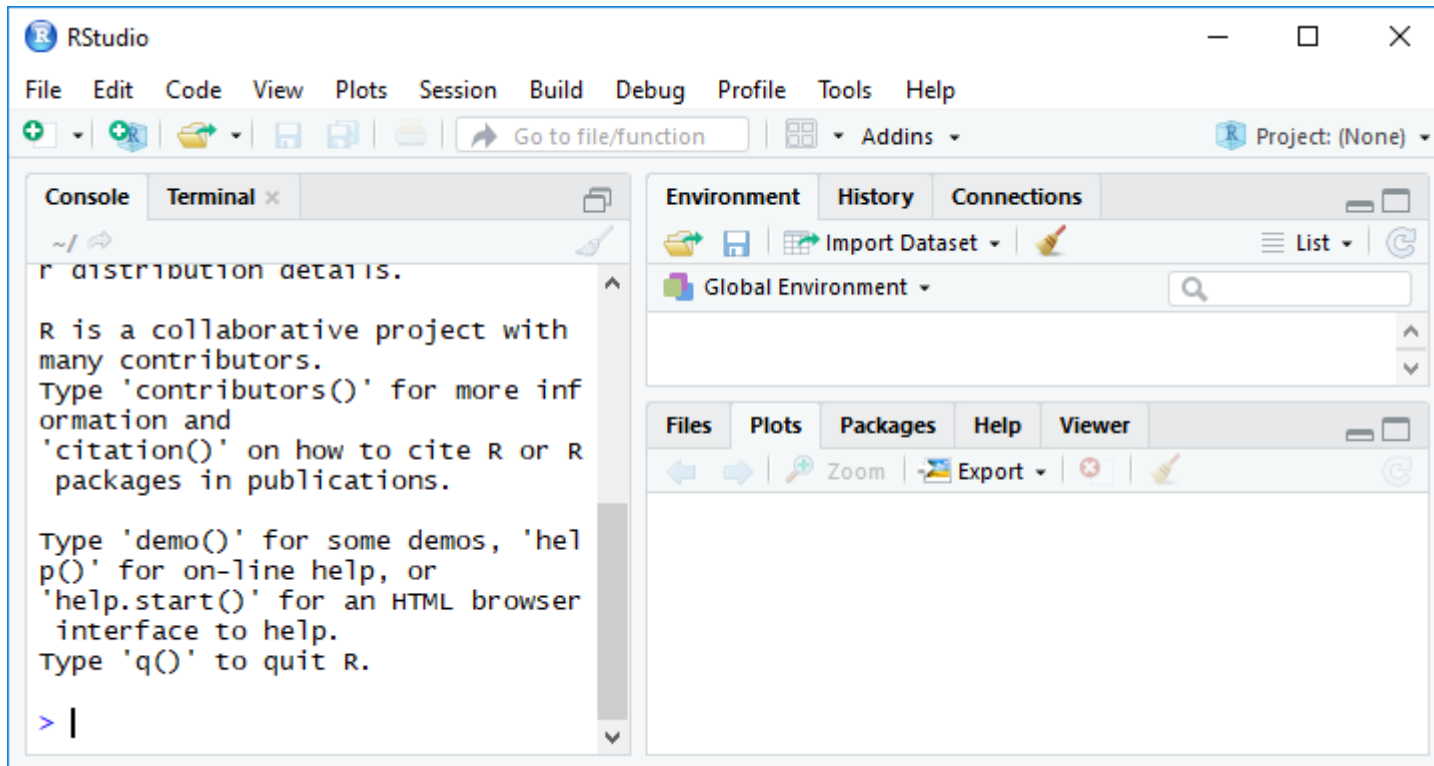


Figure 6.2: Capture d'écran de RStudio sous Windows : fenêtre par défaut.

avons pu noter que la partie gauche de RStudio est désormais séparée en deux, avec en bas de l'écran la console et en haut de l'écran le script.

Nous pouvons alors commencer l'écriture de notre script avec les commentaires décrivant ce que nous allons y trouver, et y ajouter un calcul simple. Une fois que nous avons recopié le code suivant, nous pouvons sauvegarder notre script avec la commande `CTRL + S` ou en se rendant dans *File*, puis *Save*.

```
# -----
# Un script para seleccionar su entorno de desarrollo
# fecha de creación : 27/06/2018
# autor : François Rebaudo
# -----

# [1] cálculos simples
# -----
nbrRep <- 5
pi * nbrRep^2

## [1] 78.53982
```

Pour exécuter notre script, il suffit de sélectionner les lignes que nous souhaitons exécuter et d'utiliser la combinaison de touches `CTRL + ENTER`. Le résultat apparaît dans la console (Figure 6.3).

Nous pouvons voir que par défaut dans la partie du script les commentaires apparaissent en vert, les chiffres en bleu, et le reste du code en noir. Dans la partie de la console ce qui a été exécuté apparaît en bleu et les résultats de l'exécution en noir. Nous pouvons également noter que dans la partie du code chaque ligne comporte un numéro correspondant au numéro de ligne à gauche sur fond gris. Il s'agit de la coloration syntaxique par défaut avec RStudio. Cette coloration syntaxique peut être modifiée en se rendant dans le menu *Tools, Global Options..., Appearance*, puis en choisissant un autre thème dans la liste *Editor theme*:

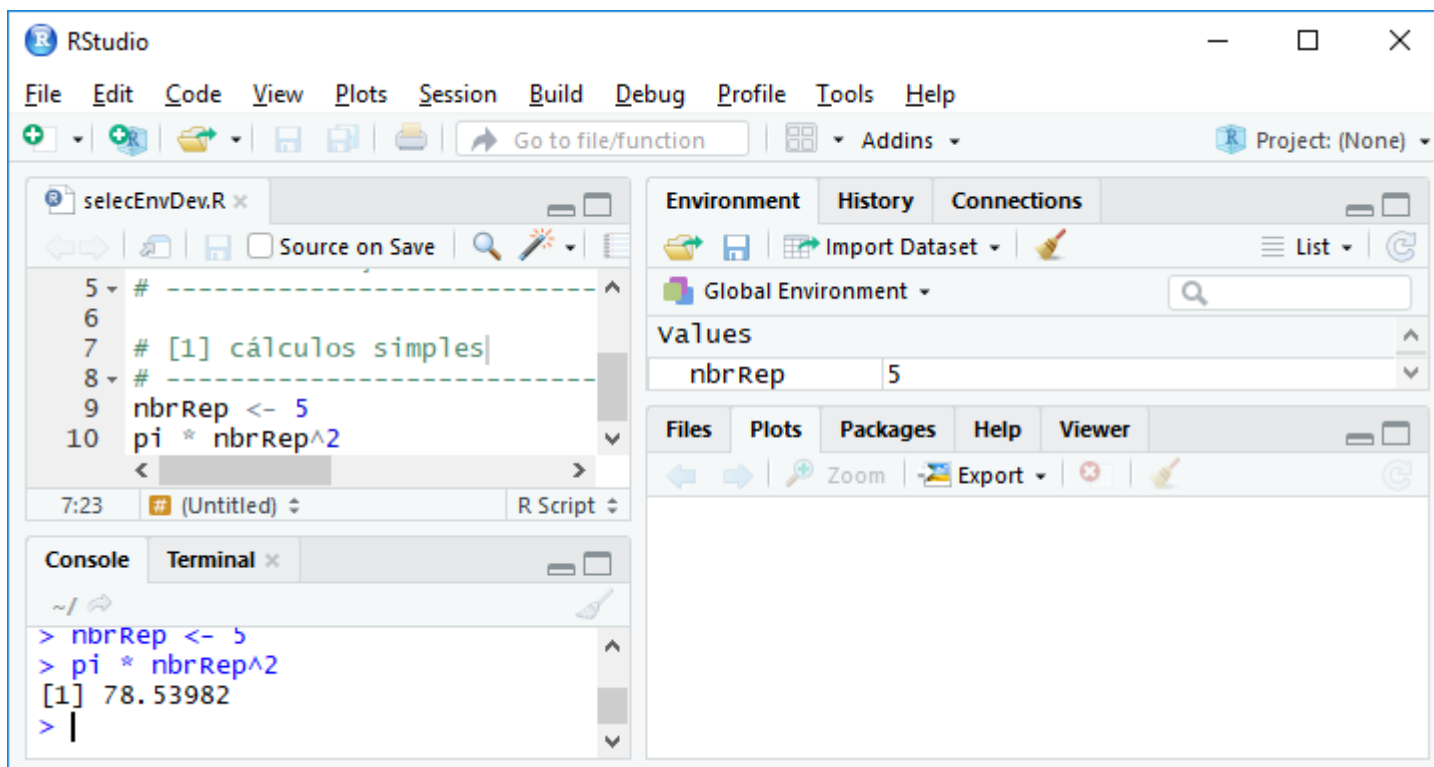


Figure 6.3: Capture d'écran de RStudio sous Windows : exécuter un script avec CTRL + ENTER.

Nous allons choisir le thème *Cobalt*, puis OK (Figure 6.4).

Nous savons comment créer un nouveau script, le sauvegarder, exécuter son contenu, et changer l'apparence de RStudio. Nous verrons les nombreux autres avantages de RStudio tout au long de ce livre car c'est l'environnement de développement qui sera utilisé. Nous serons néanmoins particulièrement vigilants à ce que tous les scripts développés tout au long de ce livre s'exécutent de la même façon quel que soit l'environnement de développement utilisé.

6.3 Notepad++ avec Npp2R

6.3.1 Installer Notepad++ (pour Windows uniquement)

Le programme pour installer Notepad++ se trouve dans l'onglet *Downloads* (<https://notepad-plus-plus.org/download/>). Vous pouvez choisir entre la version 32-bit et 64-bit (64-bit si vous ne savez pas quelle version choisir). Notepad++ seul est suffisant pour écrire un script, mais il est encore plus puissant avec *Notepad to R* (*Npp2R*) qui permet d'exécuter automatiquement nos scripts dans une console en local sur notre ordinateur ou à distance sur un serveur.

6.3.2 Installer Npp2R

Le programme pour installer Npp2R est hébergé sur le site de Sourceforge (<https://sourceforge.net/projects/npp2r/>). Npp2R doit être installé après Notepad++.

6.3.3 Un script avec Notepad++

Lors de la première ouverture Notepad++ affiche un fichier vide *new 1* (Figure 6.6).



Figure 6.4: Capture d'écran de RStudio sous Windows : changer les paramètres de coloration syntaxique.



Figure 6.5: Logo Notepad++

Puisque nous avons déjà créé un script pour le tester avec RStudio, nous allons l'ouvrir à nouveau avec Notepad++. Dans *Fichier*, sélectionnons *Ouvrir...* puis choisir le script *selecEnvDev.R* créé précédemment. Une fois le script ouvert, allons dans *Langage*, puis *R*, et encore une fois *R*. La coloration syntaxique apparaît (Figure 6.7).

L'exécution du script ne peut se faire que si Npp2R est en cours d'exécution. Pour se faire il est nécessaire de lancer le programme Npp2R depuis l'invite de Windows. Un icône devrait apparaître en bas de votre écran. L'exécution automatique du code depuis Notepad++ se fait en sélectionnant le code à exécuter puis en utilisant la commande F8. Si la commande ne fonctionne pas et que vous venez d'installer Notepad++, il est peut être nécessaire de redémarrer votre ordinateur. Si la commande fonctionne, une nouvelle fenêtre va s'ouvrir avec une console exécutant les lignes souhaitées (Figure 6.8).

Comme pour RStudio, la coloration syntaxique peut être modifiée depuis le menu *Paramètres*, et un nouveau thème peut être sélectionné (par exemple *Solarized* dans la Figure 6.9)

Par rapport aux autres éditeurs de texte, Notepad++ a l'avantage d'être très léger et offre une vaste gamme d'options pour personnaliser l'écriture du code.



Figure 6.6: Capture d'écran de Notepad++ sous Windows : fenêtre par défaut.



Figure 6.7: Capture d'écran de Notepad++ sous Windows : exécuter un script avec F8.

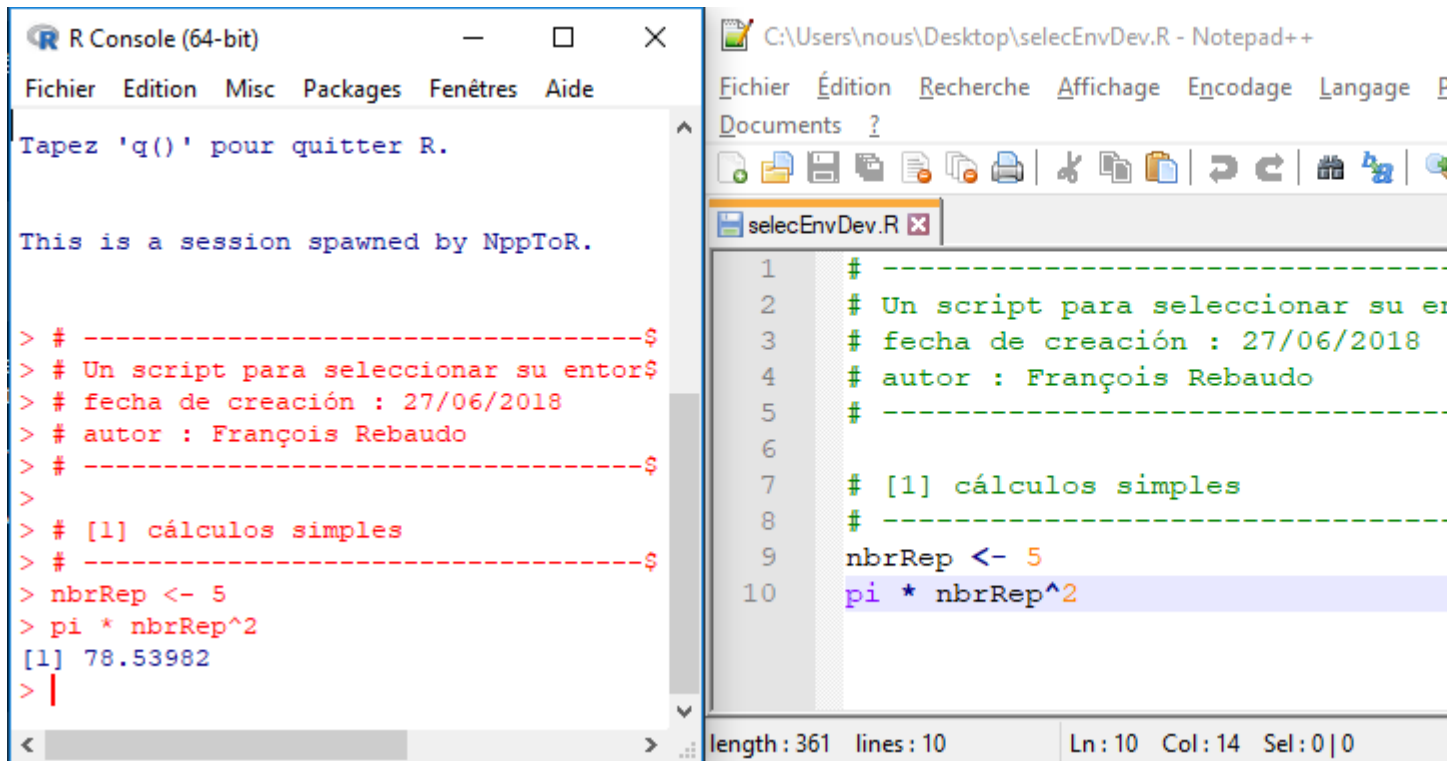


Figure 6.8: Capture d'écran de Notepad++ sous Windows : la console avec F8.



Figure 6.9: Capture d'écran de Notepad++ sous Windows : coloration syntaxique avec le thème Solarized.



Figure 6.10: Logo Geany

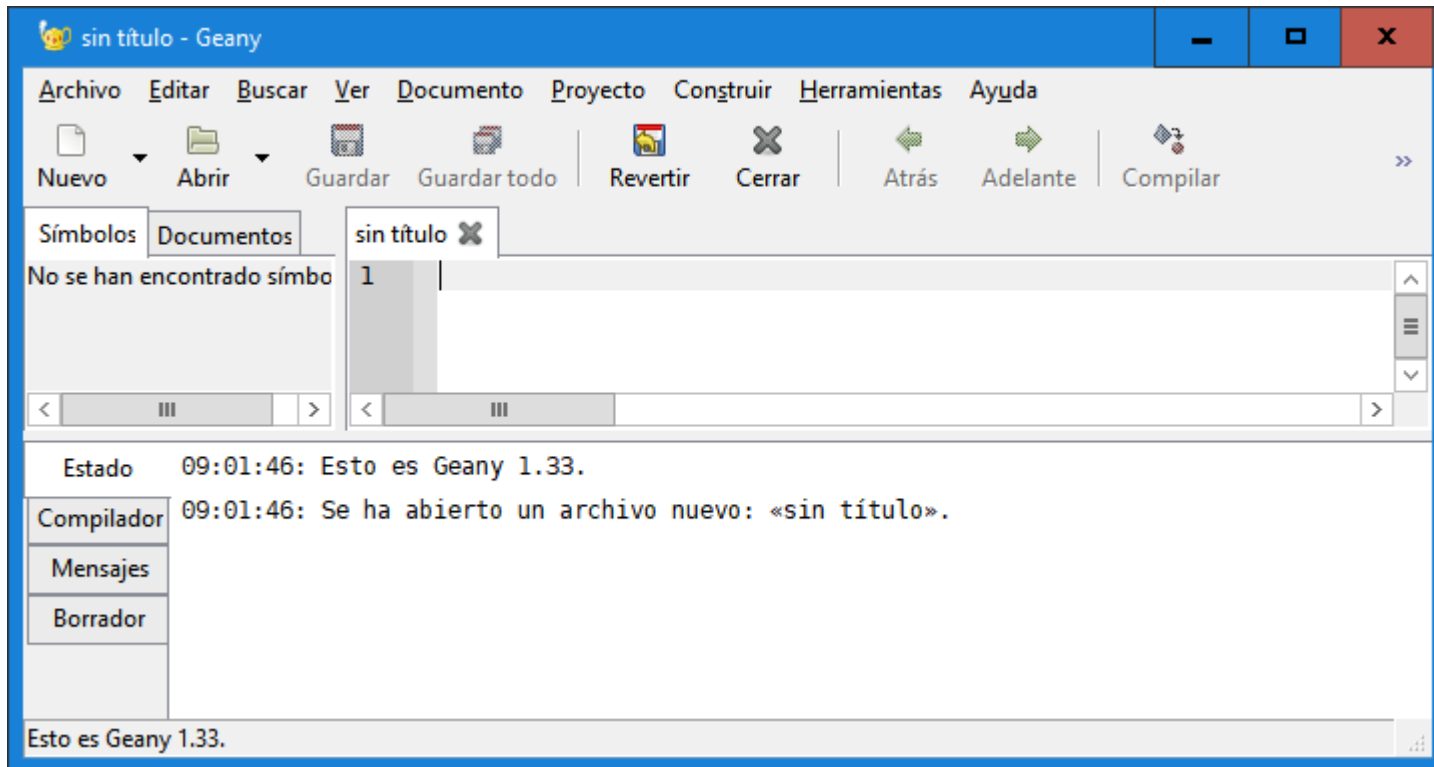


Figure 6.11: Capture d'écran de Geany sous Windows : fenêtre par défaut.

6.4 Geany (pour Linux, Mac OSX et Windows)

6.4.1 Installer Geany

Le programme pour l'installation de Geany se trouve sous l'onglet *Downloads* dans le menu de gauche *Releases* de la page web (<https://www.geany.org/>). Ensuite il suffit de télécharger l'exécutable pour Windows ou le dmg pour Mac OSX. Les utilisateurs de Linux préféreront un `sudo apt-get install geany`.

6.4.2 Un script avec Geany

Lors de la première ouverture, comme pour RStudio et Notepad++, un fichier vide est créé (Figure 6.11).

Nous pouvons ouvrir notre script avec *Fichier, Ouvrir* (Figure 6.12).

Pour exécuter notre script, la version de Geany pour Windows ne dispose pas d'un terminal intégré, ce qui rend son utilisation limitée sous ce système d'exploitation. L'exécution d'un script peut se faire en ouvrant R dans une fenêtre à part et en copiant et collant les lignes à exécuter. Sous Linux et Mac OSX, il suffit d'ouvrir R dans le terminal situé dans la partie basse de la fenêtre de Geany avec la commande `R`. Nous pouvons ensuite paramétrer Geany pour qu'une combinaison de touches permette

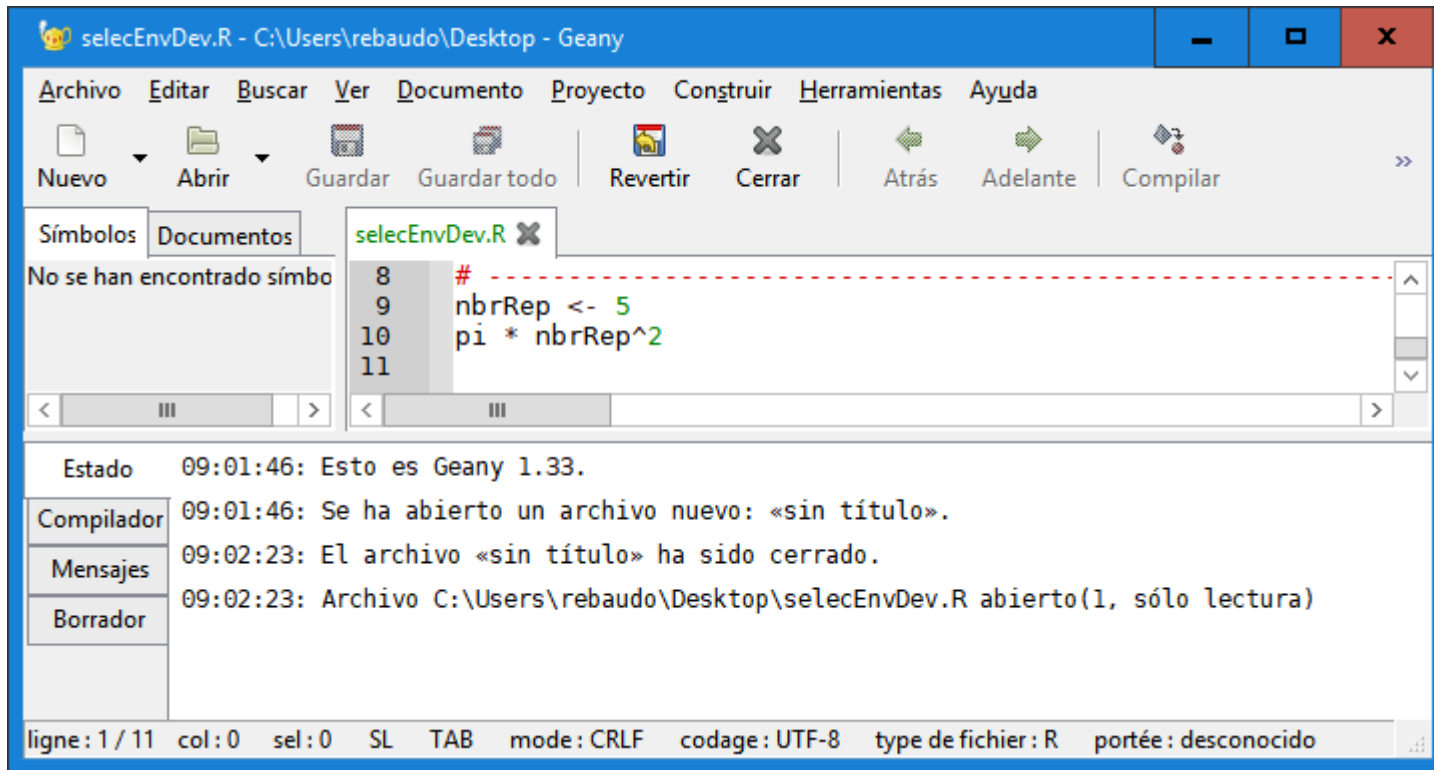


Figure 6.12: Capture d'écran de Geany sous Windows : ouvrir un script.

d'exécuter le code sélectionné (par exemple CTRL + R). Pour cela il faut tout d'abord autoriser l'envoi de sélection vers le terminal (`send_selection_unsafe=true`) dans le fichier `geany.conf` puis choisir la commande d'envoi vers le terminal (dans *Editar*, *Preferencias*, *Combinaciones*). Pour changer le thème de Geany, il existe une collection de thèmes accessibles sur GitHub (<https://github.com/geany/geany-themes/>). Le thème peut ensuite être changé via le menu *Ver*, *Cambiar esquema de color...* (un exemple avec le thème *Solarized*, Figure 6.13).

6.5 Autres solutions

Il existe beaucoup d'autres solutions, certaines spécialisées pour R comme **Tinn-R** (<https://sourceforge.net/projects/tinn-r/>), et d'autres plus généralistes pour la programmation comme **Atom** (<https://atom.io/>), **Sublime Text** (<https://www.sublimetext.com/>), **Vim** (<https://www.vim.org/>), **Gedit** (<https://wiki.gnome.org/Apps/Gedit>), **GNU Emacs** (<https://www.gnu.org/software/emacs/>), ou encore **Brackets** (<http://brackets.io/>) et **Eclipse** (<http://www.eclipse.org/>).

6.6 Conclusion

Félicitations, nous sommes arrivés au bout de ce chapitre sur environnements de développement pour utiliser R. Nous savons désormais :

- Installer RStudio, Geany ou Notepad++
- Reconnaître et choisir notre environnement préféré

A partir d'ici nous allons pouvoir nous concentrer sur le langage de programmation R dans un environnement facilitant le travail de lecture et d'écriture du code. C'est un grand pas en avant pour maîtriser R.

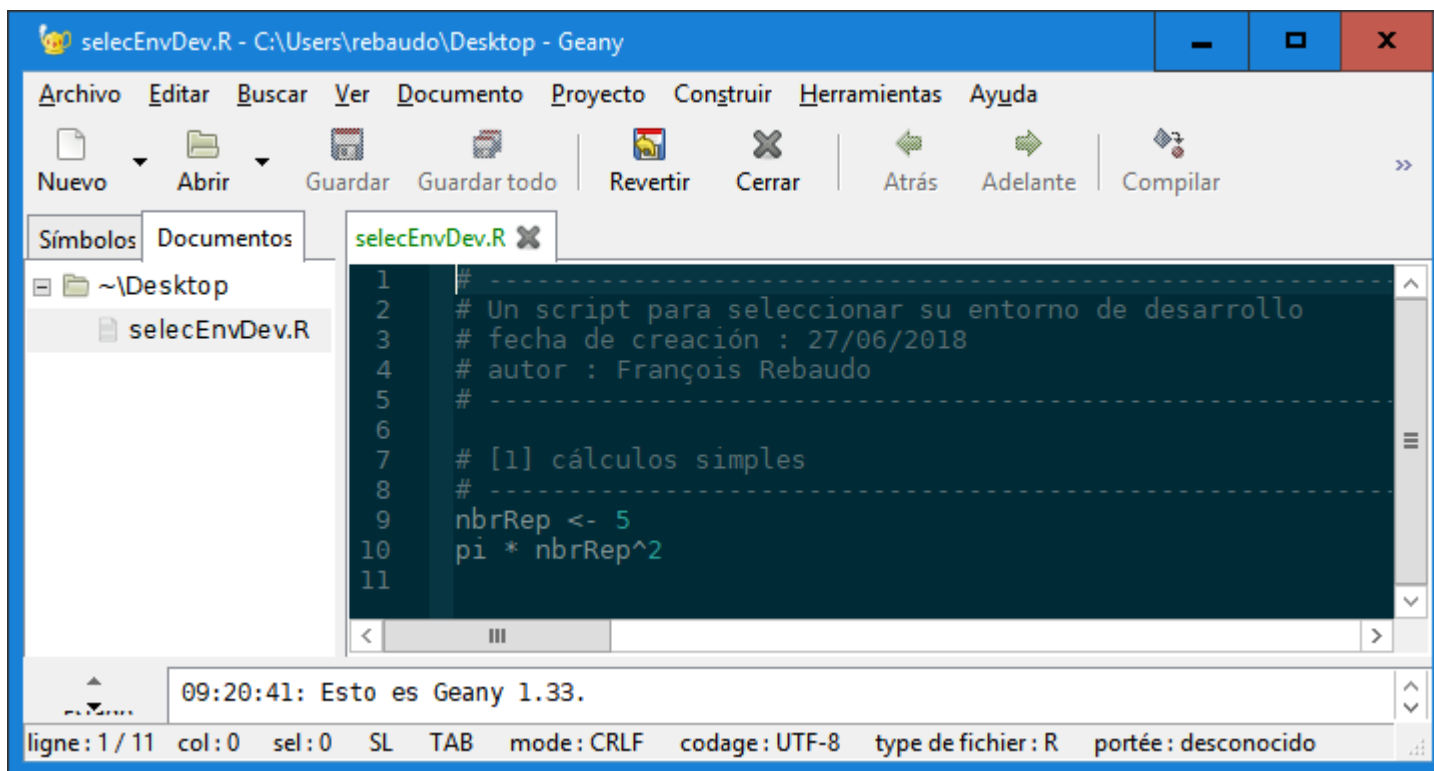


Figure 6.13: Capture d'écran de Geany sous Windows : changer les paramètres de coloration syntaxique.

Chapter 7

Les types de données

Nous avons vu précédemment comment créer un objet. Un objet est comme une boîte dans laquelle nous allons *stocker* de l'information. Jusqu'à présent nous n'avons stocké que des nombres mais dans ce chapitre nous allons voir qu'il est possible de stocker d'autres informations et nous allons nous attarder sur les types les plus courants. Dans ce chapitre nous allons utiliser des **fonctions** sur lesquelles nous reviendrons plus tard.

7.1 Le type `numeric`

Le type `numeric` correspond à ce que nous avons fait jusqu'à présent, stocker des nombres. Il existe deux principaux types de nombres avec R: les nombres entiers (*integers*), et les nombres à virgule (*double*). Par défaut R considère tous les nombres comme des nombres à virgule et attribue le type `double`. Pour vérifier le type de données nous allons utiliser la fonction `typeof()` qui prend comme argument un objet (ou directement l'information que nous souhaitons tester). Nous pouvons également utiliser la fonction `is.double()` qui va renvoyer `TRUE` si le nombre est au format `double` et `FALSE` dans le cas contraire. La fonction générique `is.numeric()` va quant à elle renvoyer `TRUE` si l'objet est au format `numeric` et `FALSE` dans le cas contraire.

```
nbrRep <- 5
typeof(nbrRep)
```

```
## [1] "double"
```

```
typeof(5.32)
```

```
## [1] "double"
```

```
is.numeric(5)
```

```
## [1] TRUE
```

```
is.double(5)
```

```
## [1] TRUE
```

Si nous voulons spécifier à R que nous allons travailler avec un nombre entier, alors il nous faut transformer notre nombre à virgule en nombre entier avec la fonction `as.integer()`. Nous pouvons également utiliser la fonction `is.integer()` qui va renvoyer `TRUE` si le nombre est au format `integer` et `FALSE` dans le cas contraire.

```
nbrRep <- as.integer(5)
typeof(nbrRep)
```

```
## [1] "integer"
```

```
typeof(5.32)
```

```
## [1] "double"
```

```
typeof(as.integer(5.32))
```

```
## [1] "integer"
```

```
as.integer(5.32)
```

```
## [1] 5
```

```
as.integer(5.99)
```

```
## [1] 5
```

```
is.numeric(nbrRep)
```

```
## [1] TRUE
```

Nous voyons ici que transformer un nombre comme 5.99 au format `integer` va renvoyer uniquement la partie entière, soit 5.

```
is.integer(5)
```

```
## [1] FALSE
```

```
is.numeric(5)
```

```
## [1] TRUE
```

```
is.integer(as.integer(5))
```

```
## [1] TRUE
```

```
is.numeric(as.integer(5))
```

```
## [1] TRUE
```

La somme d'un nombre entier et d'un nombre à virgule renvoie un nombre à virgule.

```
sumIntDou <- as.integer(5) + 5.2
typeof(sumIntDou)
```

```
## [1] "double"
```

```
sumIntInt <- as.integer(5) + as.integer(5)
typeof(sumIntInt)
```

```
## [1] "integer"
```

Pour résumer, le type `numeric` contient deux sous-types, les types `integer` pour les nombres entiers et le type `double` pour les nombres à virgule. Par défaut R attribue le type `double` aux nombres.

7.2 Le type character

Le type `character` correspond au texte. En effet, R permet de travailler avec du texte. Pour spécifier à R que l'information contenue dans un objet est au format texte (ou de manière générale pour tous les textes), il faut utiliser les guillemets doubles (`"`), ou simples (`'`).

```
myText <- "azerty"
myText2 <- 'azerty'
myText3 <- 'azerty uiop qsdg hklm'
typeof(myText3)
```

```
## [1] "character"
```

Les guillemets doubles ou simples sont utiles si l'on souhaite mettre des guillemets dans notre texte. Nous pouvons également *échapper* un caractère spécial comme un guillemet grâce au signe `backslash \`.

```
myText <- "a 'ze' 'rt' y"
myText2 <- 'a "zert" y'
myText3 <- 'azerty uiop qsdg hklm'
myText4 <- "qwerty \" azerty "
myText5 <- "qwerty \\ azerty "
```

Par défaut lorsque nous créons un objet, son contenu n'est pas renvoyé par la console. Sur Internet ou dans de nombreux ouvrages nous pouvons retrouver le nom de l'objet sur une ligne pour renvoyer son contenu :

```
myText <- "a 'ze' 'rt' y"
myText
```

```
## [1] "a 'ze' 'rt' y"
```

Dans ce livre nous n'utiliserons jamais cette façon de faire et préférons l'utilisation de la fonction `print()`, qui permet d'afficher dans la console le contenu d'un objet. Le résultat est le même mais le code est alors plus facile à lire et plus explicite sur ce qui est fait.

```
myText <- "a 'ze' 'rt' y"
print(myText)
```

```
## [1] "a 'ze' 'rt' y"
```

```
nbrRep <- 5
print(nbrRep)
```

```
## [1] 5
```

Nous pouvons également mettre des chiffres au format texte, mais il ne faut pas oublier de mettre des guillemets pour spécifier le type `character` ou utiliser la fonction `as.character()`. Une opération entre du texte et un nombre renvoie une erreur. Par exemple si l'on ajoute 10 à "5", R nous signale qu'un **argument** de la **fonction** `+` n'est pas de type `numeric` et que donc l'opération n'est pas possible. Nous ne pouvons pas non plus ajouter du texte à du texte, mais verrons plus tard comment *concaténer* deux chaînes de texte.

```
myText <- "qwerty"
typeof(myText)
```

```
## [1] "character"
```

```
myText2 <- 5
typeof(myText2)
```

```
## [1] "double"
```

```
myText3 <- "5"
typeof(myText3)
```

```
## [1] "character"
```

```
myText2 + 10
```

```
## [1] 15
```

```
as.character(5)
```

```
## [1] "5"
```

```
# myText3 + 10 # Error in myText3 + 10 : non-numeric argument to binary operator
# "a" + "b" # Error in "a" + "b" : non-numeric argument to binary operator
```

Pour résumer, le type `character` permet la saisie de texte, nous pouvons le reconnaître grâce aux guillemets simples ou doubles.

7.3 Le type factor

Le type `factor` correspond aux facteurs. Les facteurs sont un choix parmi une liste finie de possibilités. Par exemple les pays sont des facteurs car il y a une liste finie de pays dans le monde à un temps donné. Un facteur peut être défini avec la fonction `factor()` ou transformé en utilisant la fonction `as.factor()`. Comme pour les autres types de donnée nous pouvons utiliser la fonction `is.factor()` pour vérifier le type de donnée. Pour avoir la liste de toutes les possibilités, il existe la fonction `levels()` (cette fonction prendra plus de sens quand nous aurons abordé les types de conteneur de l'information).

```
factor01 <- factor("aaa")
print(factor01)
```

```
## [1] aaa
```

```
## Levels: aaa
```



```
typeof(factor01)
```

```
## [1] "integer"
```

```
is.factor(factor01)
```

```
## [1] TRUE
```

```
levels(factor01)
```

```
## [1] "aaa"
```

Un facteur peut être transformé en texte avec la fonction `as.character()` mais également en nombre avec `as.numeric()`. Lors de la transformation en nombre chaque facteur prend la valeur de sa position dans la liste des possibilités. Dans notre cas il n'y a qu'une seule possibilité donc la fonction `as.numeric()` va renvoyer 1:

```
factor01 <- factor("aaa")  
as.character(factor01)
```

```
## [1] "aaa"
```

```
as.numeric(factor01)
```

```
## [1] 1
```

7.4 Le type logical

Le type `logical` correspond aux valeurs `TRUE` et `FALSE` (et `NA`) que nous avons déjà vu avec les opérateurs de comparaison.

```
aLogic <- TRUE  
print(aLogic)
```

```
## [1] TRUE
```

```
typeof(aLogic)
```

```
## [1] "logical"
```

```
is.logical(aLogic)
```

```
## [1] TRUE
```

```
aLogic + 1
```

```
## [1] 2
```

```
as.numeric(aLogic)
```

```
## [1] 1
```

```
as.character(aLogic)
```

```
## [1] "TRUE"
```

7.5 A propos de NA

La valeur NA peut être utilisée pour spécifier l'absence de données ou les données manquantes. Par défaut NA est de type `logical` mais il peut être utilisé pour du texte, ou des nombres.

```
print(NA)
```

```
## [1] NA
```

```
typeof(NA)
```

```
## [1] "logical"
```

```
typeof(as.integer(NA))
```

```
## [1] "integer"
```

```
typeof(as.character(NA))
```

```
## [1] "character"
```

```
NA == TRUE
```

```
## [1] NA
```

```
NA == FALSE
```

```
## [1] NA
```

```
NA > 1
```

```
## [1] NA
```

```
NA + 1
```

```
## [1] NA
```

7.6 Conclusion

Félicitations, nous sommes arrivés au bout de ce chapitre sur les type de données. Nous savons désormais :

- Reconnaître et faire des objets dans les principaux types de données
- Transformer les types de données d'un type à un autre

Ce chapitre un peu fastidieux est la base pour aborder le prochain chapitre sur les conteneurs des données.

Chapter 8

Les conteneurs de données

Jusqu'à présent nous avons fait des objets simples ne contenant qu'une seule valeur. Nous avons néanmoins pu voir qu'un objet avait différents attributs, comme sa valeur, mais aussi le type de donnée contenue. maintenant nous allons voir qu'il existe différents types de conteneurs permettant de stocker plusieurs données.

8.1 Le conteneur vector

Dans R, un `vector` est une combinaison de données avec la particularité que toutes les données contenues dans un `vector` sont du même type. Nous pouvons donc stocker plusieurs `numeric` ou `character` dans un `vector`, mais pas les deux. Le conteneur `vector` est important car c'est l'élément de base de R.

8.1.1 Créer un vector

Pour créer un `vector` nous allons utiliser la fonction `c()` qui permet de combiner des éléments en un `vector`. Les éléments à combiner doivent être séparés par des virgules.

```
miVec01 <- c(1, 2, 3, 4) # un vecteur de 4 éléments de type numeric ; double
print(miVec01)
```

```
## [1] 1 2 3 4
```

```
typeof(miVec01)
```

```
## [1] "double"
```

```
is.vector(miVec01)
```

```
## [1] TRUE
```

La fonction `is.vector()` permet de vérifier le type de conteneur.

```
miVec02 <- c("a", "b", "c")
print(miVec02)
```

```
## [1] "a" "b" "c"
```

```
typeof(miVec02)
```

```
## [1] "character"
```

```
is.vector(miVec02)
```

```
## [1] TRUE
```

```
miVec03 <- c(TRUE, FALSE, FALSE, TRUE)  
print(miVec03)
```

```
## [1] TRUE FALSE FALSE TRUE
```

```
typeof(miVec03)
```

```
## [1] "logical"
```

```
is.vector(miVec03)
```

```
## [1] TRUE
```

```
miVecNA <- c(1, NA, 3, NA, 5)  
print(miVecNA)
```

```
## [1] 1 NA 3 NA 5
```

```
typeof(miVecNA)
```

```
## [1] "double"
```

```
is.vector(miVecNA)
```

```
## [1] TRUE
```

```
miVec04 <- c(1, "a")  
print(miVec04)
```

```
## [1] "1" "a"
```

```
typeof(miVec04)
```

```
## [1] "character"
```

```
is.vector(miVec04)
```

```
## [1] TRUE
```

Si l'on combine différents types de données, par défaut R va chercher à transformer les éléments en un seul type. Si comme ici dans l'objet `miVec03` nous avons des `character` et des `numeric`, R va transformer tous les éléments en `character`.

```
miVec05 <- c(factor("abc"), "def")
print(miVec05)
```

```
## [1] "1"    "def"
```

```
typeof(miVec05)
```

```
## [1] "character"
```

```
miVec06 <- c(TRUE, "def")
print(miVec06)
```

```
## [1] "TRUE" "def"
```

```
typeof(miVec06)
```

```
## [1] "character"
```

```
miVec07 <- c(factor("abc"), 55)
print(miVec07)
```

```
## [1] 1 55
```

```
typeof(miVec07)
```

```
## [1] "double"
```

```
miVec08 <- c(TRUE, 55)
print(miVec08)
```

```
## [1] 1 55
```

```
typeof(miVec08)
```

```
## [1] "double"
```

Nous pouvons aussi combiner des objets existants au sein d'un vector.

```
miVec09 <- c(miVec02, "d", "e", "f")
print(miVec09)
```

```
## [1] "a" "b" "c" "d" "e" "f"
```

```
miVec10 <- c("aaa", "aa", miVec09, "d", "e", "f")
print(miVec10)
```

```
## [1] "aaa" "aa"  "a"   "b"   "c"   "d"   "e"   "f"   "d"   "e"   "f"
```

```
miVec11 <- c(789, miVec01, 564)
print(miVec11)
```

```
## [1] 789 1 2 3 4 564
```

8.1.2 Opérations sur un vector

Nous pouvons également effectuer des opérations sur un vector.

```
print(miVec01)
```

```
## [1] 1 2 3 4
```

```
miVec01 + 1
```

```
## [1] 2 3 4 5
```

```
miVec01 - 1
```

```
## [1] 0 1 2 3
```

```
miVec01 * 2
```

```
## [1] 2 4 6 8
```

```
miVec01 /10
```

```
## [1] 0.1 0.2 0.3 0.4
```

Les opérations d'un vector sur un autre sont aussi possibles, mais il faut veiller à ce que le nombre d'éléments d'un vector soit le même que l'autre, sinon R va effectuer le calcul en repartant du début. Voici un exemple pour illustrer ce que R fait:

```
miVec12 <- c(1, 1, 1, 1, 1, 1, 1, 1, 1)
print(miVec12)
```

```
## [1] 1 1 1 1 1 1 1 1 1
```

```
miVec13 <- c(10, 20, 30)
print(miVec13)
```

```
## [1] 10 20 30
```

```
miVec12 + miVec13 # vecteurs de tailles différentes : attention au résultat
```

```
## [1] 11 21 31 11 21 31 11 21 31
```

```
miVec14 <- c(10, 20, 30, 40, 50, 60, 70, 80, 90)
print(miVec14)
```

```
## [1] 10 20 30 40 50 60 70 80 90
```

```
miVec12 + miVec14 # les vecteurs sont de la même longueur
```

```
## [1] 11 21 31 41 51 61 71 81 91
```

```
miVec15 <- c(1, 1, 1, 1)
print(miVec15)
```

```
## [1] 1 1 1 1
```

```
miVec15 + miVec13 # vecteurs de tailles différentes et non multiples
```

```
## Warning in miVec15 + miVec13: la taille d'un objet plus long n'est pas
## multiple de la taille d'un objet plus court
```

```
## [1] 11 21 31 11
```

8.1.3 Accéder aux valeurs d'un vector

Il est souvent nécessaire de pouvoir accéder aux valeurs d'un vector, c'est à dire de récupérer une valeur ou un groupe de valeurs au sein d'un vector. Pour accéder à un élément dans un vector nous utilisons les crochets []. Entre les crochets, nous pouvons utiliser un numéro correspondant au numéro de l'élément dans le vector.

```
miVec20 <- c(10, 20, 30, 40, 50, 60, 70, 80, 90)
miVec21 <- c("a", "b", "c", "d", "e", "f", "g", "h", "i")
print(miVec20)
```

```
## [1] 10 20 30 40 50 60 70 80 90
```

```
print(miVec21)
```

```
## [1] "a" "b" "c" "d" "e" "f" "g" "h" "i"
```

```
print(miVec20[1])
```

```
## [1] 10
```

```
print(miVec21[3])
```

```
## [1] "c"
```

Nous pouvons aussi utiliser la combinaison de différents éléments (un autre vector).

```
print(miVec20[c(1, 5, 9)])
```

```
## [1] 10 50 90
```

```
print(miVec21[c(4, 3, 1)])
```

```
## [1] "d" "c" "a"
```

```
print(miVec21[c(4, 4, 3, 4, 3, 2, 5)])
```

```
## [1] "d" "d" "c" "d" "c" "b" "e"
```

Nous pouvons aussi sélectionner des éléments en utilisant un opérateur de comparaison ou un opérateur logique.

```
print(miVec20[miVec20 >= 50])
```

```
## [1] 50 60 70 80 90
```

```
print(miVec20[(miVec20 >= 50) & ((miVec20 < 80))])
```

```
## [1] 50 60 70
```

```
print(miVec20[miVec20 != 50])
```

```
## [1] 10 20 30 40 60 70 80 90
```

```
print(miVec20[miVec20 == 30])
```

```
## [1] 30
```

```
print(miVec20[(miVec20 == 30) | (miVec20 == 50)])
```

```
## [1] 30 50
```

```
print(miVec21[miVec21 == "a"])
```

```
## [1] "a"
```

Une autre fonctionnalité intéressante est de conditionner les éléments à sélectionner dans un `vector` en fonction d'un autre `vector`.

```
print(miVec21[miVec20 >= 50])
```

```
## [1] "e" "f" "g" "h" "i"
```

```
print(miVec21[(miVec20 >= 50) & ((miVec20 < 80))])
```

```
## [1] "e" "f" "g"
```

```
print(miVec21[miVec20 != 50])
```

```
## [1] "a" "b" "c" "d" "f" "g" "h" "i"
```

```
print(miVec21[miVec20 == 30])
```

```
## [1] "c"
```

```
print(miVec21[(miVec20 == 30) | (miVec20 == 50)])
```

```
## [1] "c" "e"
```



```
print(miVec21[(miVec20 == 30) | (miVec21 == "h")])
```

```
## [1] "c" "h"
```

Il est aussi possible d'exclure certains éléments plutôt que de les sélectionner.

```
print(miVec20[-1])
```

```
## [1] 20 30 40 50 60 70 80 90
```

```
print(miVec21[-5])
```

```
## [1] "a" "b" "c" "d" "f" "g" "h" "i"
```

```
print(miVec20[-c(1, 2, 5)])
```

```
## [1] 30 40 60 70 80 90
```

```
print(miVec21[-c(1, 2, 5)])
```

```
## [1] "c" "d" "f" "g" "h" "i"
```

Les éléments d'un vector peuvent aussi être sélectionné sur la base d'un vector de type logical. Dans ce cas seuls les éléments avec une valeur TRUE seront sélectionnés.

```
miVec22 <- c(TRUE, TRUE, FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE)
print(miVec21[miVec22])
```

```
## [1] "a" "b" "d" "f" "h" "i"
```

8.1.4 Nommer les éléments d'un vector

Les éléments d'un vector peuvent être nommé pour pouvoir s'y référer par la suite et opérer une sélection. La fonction `names()` permet de récupérer les noms des éléments d'un vecteur.

```
miVec23 <- c(aaa = 10, bbb = 20, ccc = 30, ddd = 40, eee = 50)
print(miVec23)
```

```
## aaa bbb ccc ddd eee
## 10 20 30 40 50
```

```
print(miVec23["bbb"])
```

```
## bbb
## 20
```

```
print(miVec23[c("bbb", "ccc", "bbb")])
```

```
## bbb ccc bbb
## 20 30 20
```

```
names(miVec23)
```

```
## [1] "aaa" "bbb" "ccc" "ddd" "eee"
```

8.1.5 Modifier les éléments d'un vector

Pour modifier un vecteur, nous opérons de la même façon que pour modifier un objet simple, avec le signe `<-` et l'élément ou les éléments à modifier entre crochets.

```
print(miVec21)
```

```
## [1] "a" "b" "c" "d" "e" "f" "g" "h" "i"
```

```
miVec21[3] <- "zzz"
print(miVec21)
```

```
## [1] "a" "b" "zzz" "d" "e" "f" "g" "h" "i"
```

```
miVec21[(miVec20 >= 50) & ((miVec20 < 80))] <- "qwerty"
print(miVec21)
```

```
## [1] "a" "b" "zzz" "d" "qwerty" "qwerty" "qwerty" "h"
## [9] "i"
```

```
print(miVec23)
```

```
## aaa bbb ccc ddd eee
## 10 20 30 40 50
```

```
miVec23["ccc"] <- miVec23["ccc"] + 100
print(miVec23)
```

```
## aaa bbb ccc ddd eee
## 10 20 130 40 50
```

Nous pouvons aussi changer les noms associés aux éléments d'un vector.

```
print(miVec23)
```

```
## aaa bbb ccc ddd eee
## 10 20 130 40 50
```

```
names(miVec23)[2] <- "bb_bb"
print(miVec23)
```

```
## aaa bb_bb ccc ddd eee
## 10 20 130 40 50
```

Nous pouvons faire bien plus avec un vector et reviendrons sur leur manipulations et les opérations lors du chapitre sur les fonctions.

8.2 Le conteneur list

Le deuxième type de conteneur que nous allons introduire est le conteneur `list`, qui est également le deuxième conteneur après le type `vector` de part son importance dans la programmation avec R. Le conteneur de type `list` permet de stocker une **liste** d'éléments. Contrairement à ce que nous avons vu précédemment avec le type `vector`, les éléments du type `list` peuvent être différents (par exemple un `vector` de type `numeric`, puis un vecteur de type `character`). Les éléments du type `list` peuvent aussi être des conteneurs différents (par exemple un `vector`, puis une `list`). Le type de conteneur `list` prendra tout son sens lorsque nous aurons étudié les **boucles** et les **fonctions** de la famille `apply`.

8.2.1 Créer une list

Pour créer une `list` nous allons utiliser la fonction `list()` qui prend comme argument des éléments (objets).

```
miList01 <- list()
print(miList01)
```

```
## list()
```

```
miList02 <- list(5, "qwerty", c(4, 5, 6), c("a", "b", "c"))
print(miList02)
```

```
## [[1]]
## [1] 5
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 4 5 6
##
## [[4]]
## [1] "a" "b" "c"
```

```
miList03 <- list(5, "qwerty", list(c(4, 5, 6), c("a", "b", "c")))
print(miList03)
```

```
## [[1]]
## [1] 5
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [[3]][[1]]
## [1] 4 5 6
##
## [[3]][[2]]
## [1] "a" "b" "c"
```

La fonction `is.list()` permet de tester si nous avons bien créer un objet de type `list`.

```
is.list(miList02)
```

```
## [1] TRUE
```

```
typeof(miList02)
```

```
## [1] "list"
```

8.2.2 Accéder aux valeurs d'une list

Les éléments du conteneur `list` sont identifiables grâce aux double crochets `[[]]`.

```
print(miList02)
```

```
## [[1]]
## [1] 5
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 4 5 6
##
## [[4]]
## [1] "a" "b" "c"
```

Dans l'objet `miList02` de type `list`, il y a quatre éléments identifiables avec `[[1]]`, `[[2]]`, `[[3]]`, et `[[4]]`. Chacun des éléments est de type `vector` de taille 1 et de type `double` pour le premier élément, de taille 1 et de type `character` pour le deuxième élément, de taille 3 et de type `double` pour le troisième élément, et de taille 3 et de type `character` pour le quatrième élément.

```
typeof(miList02)
```

```
## [1] "list"
```

```
print(miList02[[1]])
```

```
## [1] 5
```

```
typeof(miList02[[1]])
```

```
## [1] "double"
```

```
print(miList02[[2]])
```

```
## [1] "qwerty"
```

```
typeof(miList02[[2]])
```

```
## [1] "character"
```

```
print(miList02[[3]])
```

```
## [1] 4 5 6
```

```
typeof(miList02[[3]])
```

```
## [1] "double"
```

```
print(miList02[[4]])
```

```
## [1] "a" "b" "c"
```

```
typeof(miList02[[4]])
```

```
## [1] "character"
```

L'accès au deuxième élément du vector situé en quatrième position de la liste se fait donc avec `miList02[[4]][2]`. Nous utilisons un double crochet pour le quatrième élément de la liste, puis un simple crochet pour le deuxième élément du vector.

```
print(miList02[[4]][2])
```

```
## [1] "b"
```

Comme une liste peut contenir elle-même une ou plusieurs listes, nous pouvons accéder à l'information recherchée en combinant les doubles crochets. L'objet `miList04` est une liste de deux éléments, les listes `miList02` et `miList03`. L'objet `miList03` contient lui-même une liste comme élément en troisième position. Pour accéder au premier élément du vector en première position de l'élément en troisième position du deuxième élément de la liste `miList04`, nous pouvons utiliser `miList04[[2]][[3]][[1]][1]`. Il n'y a pas de limite quant à la profondeur des listes mais dans la pratique il n'y a que rarement besoin de faire des listes de listes de listes.

```
miList04 <- list(miList02, miList03)
print(miList04)
```

```
## [[1]]
## [[1]][[1]]
## [1] 5
##
## [[1]][[2]]
## [1] "qwerty"
##
## [[1]][[3]]
## [1] 4 5 6
##
## [[1]][[4]]
## [1] "a" "b" "c"
##
##
## [[2]]
## [[2]][[1]]
## [1] 5
##
## [[2]][[2]]
```

```
## [1] "qwerty"
##
## [[2]][[3]]
## [[2]][[3]][[1]]
## [1] 4 5 6
##
## [[2]][[3]][[2]]
## [1] "a" "b" "c"
```

```
print(miList04[[2]][[3]][[1]][1])
```

```
## [1] 4
```

Pour rendre concret l'exemple précédent, nous pouvons imaginer des espèces de foreurs de maïs (*Sesamia nonagrioides* et *Ostrinia nubilalis*), échantillonnées dans différents sites, avec différentes abondances à quatre dates. Ici nous allons donner des noms aux éléments des listes.

```
bddInsect <- list(Snonagrioides = list(site01 = c(12, 5, 8, 7), site02 = c(5, 23, 4, 41), site03 = c(12, 0, 0, 0)),
  Onubilalis = list(site01 = c(12, 1, 2, 3), site02 = c(0, 0, 0, 1), site03 = c(1, 1, 2, 3)))
print(bddInsect)
```

```
## $Snonagrioides
## $Snonagrioides$site01
## [1] 12 5 8 7
##
## $Snonagrioides$site02
## [1] 5 23 4 41
##
## $Snonagrioides$site03
## [1] 12 0 0 0
##
##
## $Onubilalis
## $Onubilalis$site01
## [1] 12 1 2 3
##
## $Onubilalis$site02
## [1] 0 0 0 1
##
## $Onubilalis$site03
## [1] 1 1 2 3
```



La lecture d'une ligne de code longue comme celle de la création de l'objet `bddInsect` est difficile à lire car la profondeur des éléments ne peut se déduire que grâce aux parenthèses. C'est pourquoi nous allons reorganiser le code pour lui donner plus de lisibilité grâce à l'**indentation**. L'indentation consiste à mettre l'information à des niveaux différents de telle manière que nous puissions rapidement identifier les différents niveaux d'un code. L'indentation se fait au moyen de la touche de tabulation du clavier. Nous reviendrons sur l'indentation avec plus de précisions lors du chapitre sur les **boucles**. Nous retiendrons pour le moment que si une ligne de code est trop longue, nous gagnons en lisibilité en passant à la ligne et que R va lire l'ensemble comme une seule ligne de code.

```
bddInsect <- list(
  Snonagrioides = list(
    site01 = c(12, 5, 8, 7),
    site02 = c(5, 23, 4, 41),
```

```

    site03 = c(12, 0, 0, 0)
  ),
  Onubilalis = list(
    site01 = c(12, 1, 2, 3),
    site02 = c(0, 0, 0, 1),
    site03 = c(1, 1, 2, 3)
  )
)

```

Nous pouvons sélectionner les données d'abondance du deuxième site de la première espèce comme précédemment `bddInsect[[1]][[2]]`, ou alternativement en utilisant les noms des éléments `bddInsect$Snonagrioides$site02`. Pour ce faire nous utilisons le signe `$`, ou alors le nom des éléments avec des guillemets simples ou doubles `bddInsect[['Snonagrioides']][[2]]`.

```
print(bddInsect[[1]][[2]])
```

```
## [1] 5 23 4 41
```

```
print(bddInsect$Snonagrioides$site02)
```

```
## [1] 5 23 4 41
```

```
print(bddInsect[['Snonagrioides']][['site02']])
```

```
## [1] 5 23 4 41
```

Comme pour les vecteurs nous pouvons récupérer les noms des éléments avec la fonction `names()`.

```
names(bddInsect)
```

```
## [1] "Snonagrioides" "Onubilalis"
```

```
names(bddInsect[[1]])
```

```
## [1] "site01" "site02" "site03"
```

Lorsque nous utilisons les doubles crochets `[[]]` ou le signe `$`, R renvoie le contenu de l'élément sélectionné. Dans notre exemple les données d'abondance sont contenues sous la forme d'un `vector`, donc R renvoie un élément de type `vector`. Si nous souhaitons sélectionner un élément d'une `list` mais en conservant le format `list`, alors nous pouvons utiliser les crochets simples `[]`.

```
print(bddInsect[[1]][[2]])
```

```
## [1] 5 23 4 41
```

```
typeof(bddInsect[[1]][[2]])
```

```
## [1] "double"
```

```
is.list(bddInsect[[1]][[2]])
```

```
## [1] FALSE
```

```
print(bddInsect[[1]][2])
```

```
## $site02
## [1]  5 23  4 41
```

```
typeof(bddInsect[[1]][2])
```

```
## [1] "list"
```

```
is.list(bddInsect[[1]][2])
```

```
## [1] TRUE
```

L'utilisation des crochets simples `[]` est utile lorsque nous souhaitons récupérer plusieurs éléments d'une `list`. Par exemple pour sélectionner les abondances d'insectes des deux premiers sites de la première espèce, nous utiliserons `bddInsect[[1]][c(1, 2)]` ou alternativement `bddInsect[[1]][c("site01", "site02")]`.

```
print(bddInsect[[1]][c(1, 2)])
```

```
## $site01
## [1] 12  5  8  7
##
## $site02
## [1]  5 23  4 41
```

```
print(bddInsect[[1]][c("site01", "site02")])
```

```
## $site01
## [1] 12  5  8  7
##
## $site02
## [1]  5 23  4 41
```

8.2.3 Modification d'une `list`

Une `list` peut être modifiée de la même façon que pour le conteneur `vector`, c'est à dire en se référant avec des crochets à l'élément que nous souhaitons modifier.

```
print(miList02)
```

```
## [[1]]
## [1] 5
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 4 5 6
##
## [[4]]
## [1] "a" "b" "c"
```



```
miList02[[1]] <- 12
print(miList02)
```

```
## [[1]]
## [1] 12
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 4 5 6
##
## [[4]]
## [1] "a" "b" "c"
```

```
miList02[[4]] <- c("d", "e", "f")
print(miList02)
```

```
## [[1]]
## [1] 12
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 4 5 6
##
## [[4]]
## [1] "d" "e" "f"
```

```
miList02[[4]] <- c("a", "b", "c", miList02[[4]], "g", "h", "i")
print(miList02)
```

```
## [[1]]
## [1] 12
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 4 5 6
##
## [[4]]
## [1] "a" "b" "c" "d" "e" "f" "g" "h" "i"
```

```
miList02[[4]][5] <- "eee"
print(miList02)
```

```
## [[1]]
## [1] 12
##
## [[2]]
```

```
## [1] "qwerty"
##
## [[3]]
## [1] 4 5 6
##
## [[4]]
## [1] "a" "b" "c" "d" "eee" "f" "g" "h" "i"
```

```
miList02[[3]] <- miList02[[3]] * 10 - 1
print(miList02)
```

```
## [[1]]
## [1] 12
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 39 49 59
##
## [[4]]
## [1] "a" "b" "c" "d" "eee" "f" "g" "h" "i"
```

```
miList02[[3]][2] <- miList02[[1]] * 100
print(miList02)
```

```
## [[1]]
## [1] 12
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 39 1200 59
##
## [[4]]
## [1] "a" "b" "c" "d" "eee" "f" "g" "h" "i"
```

```
print(bddInsect)
```

```
## $Snonagrioides
## $Snonagrioides$site01
## [1] 12 5 8 7
##
## $Snonagrioides$site02
## [1] 5 23 4 41
##
## $Snonagrioides$site03
## [1] 12 0 0 0
##
##
## $Onubilalis
## $Onubilalis$site01
```

```
## [1] 12  1  2  3
##
## $Onubilalis$site02
## [1] 0 0 0 1
##
## $Onubilalis$site03
## [1] 1 1 2 3

bddInsect[['Snonagrioides']][['site02']] <- c(2, 4, 6, 8)
print(bddInsect)
```

```
## $Snonagrioides
## $Snonagrioides$site01
## [1] 12  5  8  7
##
## $Snonagrioides$site02
## [1] 2 4 6 8
##
## $Snonagrioides$site03
## [1] 12  0  0  0
##
##
## $Onubilalis
## $Onubilalis$site01
## [1] 12  1  2  3
##
## $Onubilalis$site02
## [1] 0 0 0 1
##
## $Onubilalis$site03
## [1] 1 1 2 3
```

Pour combiner deux list, il suffit d'utiliser la fonction `c()` que nous avons utilisée pour créer un vector.

```
miList0203 <- c(miList02, miList03)
print(miList0203)

## [[1]]
## [1] 12
##
## [[2]]
## [1] "qwerty"
##
## [[3]]
## [1] 39 1200 59
##
## [[4]]
## [1] "a" "b" "c" "d" "eee" "f" "g" "h" "i"
##
## [[5]]
## [1] 5
##
## [[6]]
## [1] "qwerty"
```

```
##
## [[7]]
## [[7]][[1]]
## [1] 4 5 6
##
## [[7]][[2]]
## [1] "a" "b" "c"
```

Un objet de type `list` peut être transformé en `vector` avec la fonction `unlist()` si le format des éléments de la `list` le permet (un `vector` ne peut contenir que des éléments du même type).

```
miList05 <- list("a", c("b", "c"), "d")
print(miList05)
```

```
## [[1]]
## [1] "a"
##
## [[2]]
## [1] "b" "c"
##
## [[3]]
## [1] "d"
```

```
miVec24 <- unlist(miList05)
print(miVec24)
```

```
## [1] "a" "b" "c" "d"
```

```
miList06 <- list(c(1, 2, 3), c(4, 5, 6, 7), 8, 9, c(10, 11))
print(miList06)
```

```
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [1] 4 5 6 7
##
## [[3]]
## [1] 8
##
## [[4]]
## [1] 9
##
## [[5]]
## [1] 10 11
```

```
miVec25 <- unlist(miList06)
print(miVec25)
```

```
## [1] 1 2 3 4 5 6 7 8 9 10 11
```

Pour ajouter un élément à une `list`, nous pouvons utiliser la fonction `c()` ou alors les crochets doubles `[[]]`.

```
print(miList05)
```

```
## [[1]]
## [1] "a"
##
## [[2]]
## [1] "b" "c"
##
## [[3]]
## [1] "d"
```

```
miList05 <- c(miList05, "e")
print(miList05)
```

```
## [[1]]
## [1] "a"
##
## [[2]]
## [1] "b" "c"
##
## [[3]]
## [1] "d"
##
## [[4]]
## [1] "e"
```

```
miList05[[5]] <- c("fgh", "ijk")
print(miList05)
```

```
## [[1]]
## [1] "a"
##
## [[2]]
## [1] "b" "c"
##
## [[3]]
## [1] "d"
##
## [[4]]
## [1] "e"
##
## [[5]]
## [1] "fgh" "ijk"
```

Pour supprimer un élément à une `list`, la technique la plus rapide consiste à attribuer la valeur `NULL` à l'élément à supprimer.

```
print(miList05)
```

```
## [[1]]
## [1] "a"
##
## [[2]]
## [1] "b" "c"
```

```
##
## [[3]]
## [1] "d"
##
## [[4]]
## [1] "e"
##
## [[5]]
## [1] "fgh" "ijk"
```

```
miList05[[2]] <- NULL
print(miList05)
```

```
## [[1]]
## [1] "a"
##
## [[2]]
## [1] "d"
##
## [[3]]
## [1] "e"
##
## [[4]]
## [1] "fgh" "ijk"
```

8.3 Le conteneur `data.frame`

Le conteneur `data.frame` peut être assimilé à un *tableau*. Il s'agit en fait d'un cas particulier de `list` où tous les éléments de la `list` ont la même longueur.

8.3.1 Créer un `data.frame`

Pour créer un `data.frame` nous allons utiliser la fonction `data.frame()` qui prend comme arguments les éléments du tableau que nous souhaitons créer. Les éléments sont de type `vector` et font tous la même taille. Nous pouvons donner un nom à chaque *colonne* (`vector`) de notre *tableau* (`data.frame`).

```
# création d'un data.frame
miDf01 <- data.frame(
  numbers = c(1, 2, 3, 4),
  logicals = c(TRUE, TRUE, FALSE, TRUE),
  characters = c("a", "b", "c", "d")
)
print(miDf01)
```

```
##   numbers logicals characters
## 1      1      TRUE         a
## 2      2      TRUE         b
## 3      3     FALSE         c
## 4      4      TRUE         d
```

```
# création des vecteurs, puis du data.frame
numbers <- c(1, 2, 3, 4)
logicals <- c(TRUE, TRUE, FALSE, TRUE)
characters <- c("a", "b", "c", "d")
miDf01 <- data.frame(numbers, logicals, characters)
print(miDf01)
```

```
##   numbers logicals characters
## 1      1      TRUE         a
## 2      2      TRUE         b
## 3      3     FALSE         c
## 4      4      TRUE         d
```

8.3.2 Accéder aux valeurs d'un data.frame

L'accès aux différentes valeurs d'un data.frame peut se faire de la même façon que pour un conteneur de type list.

```
print(miDf01$numbers) # vector
```

```
## [1] 1 2 3 4
```

```
print(miDf01[[1]]) # vector
```

```
## [1] 1 2 3 4
```

```
print(miDf01[1]) # list
```

```
##   numbers
## 1      1
## 2      2
## 3      3
## 4      4
```

```
print(miDf01["numbers"]) # list
```

```
##   numbers
## 1      1
## 2      2
## 3      3
## 4      4
```

```
print(miDf01[["numbers"]]) # vector
```

```
## [1] 1 2 3 4
```

Nous pouvons aussi utiliser une autre forme qui consiste à spécifier le ou les lignes suivi d'une virgule (et donc d'un espace après la virgule), puis le ou les colonnes entre crochets simples. Si l'information ligne ou colonne est omise, R affichera toutes les lignes ou toutes les colonnes. Là encore nous pouvons utiliser le numéro correspondant à un élément ou alors le nom de l'élément que nous souhaitons sélectionner.

```
myRow <- 2
myCol <- 1
print(miDf01[myRow, myCol])
```

```
## [1] 2
```

```
print(miDf01[myRow, ])
```

```
##  numbers logicals characters
## 2      2      TRUE      b
```

```
print(miDf01[, myCol])
```

```
## [1] 1 2 3 4
```

```
myCol <- "numbers"
print(miDf01[, myCol])
```

```
## [1] 1 2 3 4
```

Il est possible de sélectionner plusieurs lignes ou plusieurs colonnes.

```
print(miDf01[, c(1, 2)])
```

```
##  numbers logicals
## 1      1      TRUE
## 2      2      TRUE
## 3      3     FALSE
## 4      4      TRUE
```

```
print(miDf01[c(2, 1), ])
```

```
##  numbers logicals characters
## 2      2      TRUE      b
## 1      1      TRUE      a
```

Puisque chaque colonne est au format vector, nous pouvons également faire une sélection qui dépendra du contenu avec les opérateurs de comparaison et les opérateurs logiques.

```
miDfSub01 <- miDf01[miDf01$numbers > 2, ]
print(miDfSub01)
```

```
##  numbers logicals characters
## 3      3     FALSE      c
## 4      4      TRUE      d
```

```
miDfSub02 <- miDf01[(miDf01$logicals == TRUE) & (miDf01$numbers < 2), ]
print(miDfSub02)
```

```
##  numbers logicals characters
## 1      1      TRUE      a
```



```
miDfSub03 <- miDf01[(miDf01$numbers %% 2) == 0, ]
print(miDfSub03)
```

```
##  numbers logicals characters
## 2      2      TRUE         b
## 4      4      TRUE         d
```

```
miDfSub04 <- miDf01[((miDf01$numbers %% 2) == 0) | (miDf01$logicals == TRUE), ]
print(miDfSub04)
```

```
##  numbers logicals characters
## 1      1      TRUE         a
## 2      2      TRUE         b
## 4      4      TRUE         d
```

8.3.3 Modifier un data.frame

Pour ajouter un élément à un `data.frame`, nous allons procéder comme pour un conteneur de type `list`. Il faudra veiller à ce que le nouvel élément soit de la même taille que les autres éléments de notre `data.frame`. Par défaut un nouvel élément dans un `data.frame` prend comme nom la lettre `V` suivie du numéro de la colonne. Nous pouvons changer les noms de colonne avec la fonction `colnames()`. Nous avons la possibilité de donner un nom aux lignes avec la fonction `rownames()`.

```
newVec <- c(4, 5, 6, 7)
miDf01[[4]] <- newVec
print(miDf01)
```

```
##  numbers logicals characters V4
## 1      1      TRUE         a  4
## 2      2      TRUE         b  5
## 3      3     FALSE         c  6
## 4      4      TRUE         d  7
```

```
print(colnames(miDf01))
```

```
## [1] "numbers" "logicals" "characters" "V4"
```

```
colnames(miDf01)[4] <- "newVec"
print(miDf01)
```

```
##  numbers logicals characters newVec
## 1      1      TRUE         a      4
## 2      2      TRUE         b      5
## 3      3     FALSE         c      6
## 4      4      TRUE         d      7
```

```
print(rownames(miDf01))
```

```
## [1] "1" "2" "3" "4"
```

```
rownames(miDf01) <- c("row1", "row2", "row3", "row4")
print(miDf01)
```

```
##      numbers logicals characters newVec
## row1      1      TRUE          a      4
## row2      2      TRUE          b      5
## row3      3     FALSE          c      6
## row4      4      TRUE          d      7
```

```
newVec2 <- c(40, 50, 60, 70)
miDf01$newVec2 <- newVec2
print(miDf01)
```

```
##      numbers logicals characters newVec newVec2
## row1      1      TRUE          a      4      40
## row2      2      TRUE          b      5      50
## row3      3     FALSE          c      6      60
## row4      4      TRUE          d      7      70
```

Comme le type de conteneur `data.frame` est un cas particulier de `list`, la sélection et la modification se fait comme pour un conteneur de type `list`. Comme les éléments d'un `data.frame` sont de type `vector`, la sélection et la modification des éléments d'un `data.frame` se fait comme pour un conteneur de type `vector`.

```
miDf01$newVec2 <- miDf01$newVec2 * 2
print(miDf01)
```

```
##      numbers logicals characters newVec newVec2
## row1      1      TRUE          a      4      80
## row2      2      TRUE          b      5     100
## row3      3     FALSE          c      6     120
## row4      4      TRUE          d      7     140
```

```
miDf01$newVec2 + miDf01$newVec
```

```
## [1]  84 105 126 147
```

```
miDf01$newVec2[2] <- 0
print(miDf01)
```

```
##      numbers logicals characters newVec newVec2
## row1      1      TRUE          a      4      80
## row2      2      TRUE          b      5       0
## row3      3     FALSE          c      6     120
## row4      4      TRUE          d      7     140
```

Un `vector` peut être transformé en `data.frame` avec la fonction `as.data.frame()`.

```
print(newVec2)
```

```
## [1] 40 50 60 70
```

```
print(as.data.frame(newVec2))
```

```
##      newVec2
## 1         40
## 2         50
## 3         60
## 4         70
```

```
is.data.frame(newVec2)
```

```
## [1] FALSE
```

```
is.data.frame(as.data.frame(newVec2))
```

```
## [1] TRUE
```

8.4 Le conteneur matrix

Le conteneur `matrix` peut être vu comme un `vector` à deux dimensions, les lignes et les colonnes. Il correspond à une matrice en mathématique, et ne peut contenir qu'un seul type de données (`logical`, `numeric`, `character`, ...).

8.4.1 Créer une matrix

Pour créer une `matrix` nous allons tout d'abord créer un `vector`, puis spécifier le nombre souhaité de lignes et de colonnes dans la fonction `matrix()`.

```
vecForMatrix <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
miMat <- matrix(vecForMatrix, nrow = 3, ncol = 4)
print(miMat)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

Il n'est pas nécessaire de spécifier le nombre de lignes `nrow` et le nombre de colonnes `ncol`. Si nous utilisons l'un ou l'autre de ces arguments, R va automatiquement calculer le nombre correspondant.

```
miMat <- matrix(vecForMatrix, nrow = 3)
print(miMat)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

```
miMat <- matrix(vecForMatrix, ncol = 4)
print(miMat)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

Nous observons que les différents éléments du vector initial sont renseignés par colonne. Si nous souhaitons renseigner la matrix par lignes alors il faut donner la valeur TRUE à l'argument byrow.

```
miMat <- matrix(vecForMatrix, nrow = 3, byrow = TRUE)
print(miMat)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    6    7    8
## [3,]    9   10   11   12
```

Nous pouvons également donner un nom aux lignes et aux colonnes de notre matrix lors de sa création avec l'argument dimnames qui prend comme valeur une list de deux éléments : le nom des lignes puis le nom des colonnes. Nous pouvons aussi changer le nom des lignes et des colonnes a posteriori avec les fonctions rownames() et colnames().

```
miMat <- matrix(
  vecForMatrix,
  nrow = 3,
  byrow = TRUE,
  dimnames = list(c("r1", "r2", "r3"), c("c1", "c2", "c3", "c4"))
)
print(miMat)
```

```
##      c1 c2 c3 c4
## r1   1  2  3  4
## r2   5  6  7  8
## r3   9 10 11 12
```

```
colnames(miMat) <- c("col1", "col2", "col3", "col4")
rownames(miMat) <- c("row1", "row2", "row3")
print(miMat)
```

```
##      col1 col2 col3 col4
## row1    1    2    3    4
## row2    5    6    7    8
## row3    9   10   11   12
```

Il est possible de créer une matrix à partir d'un data.frame avec la fonction as.matrix() sous réserve que le data.frame ne contienne que des élément de même type (par exemple que des éléments de type numeric).

```
vecForMat01 <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
vecForMat02 <- vecForMat01 * 10
vecForMat03 <- vecForMat01 / 10
dfForMat <- data.frame(vecForMat01, vecForMat02, vecForMat03)
print(dfForMat)
```

```
##      vecForMat01 vecForMat02 vecForMat03
## 1              1           10         0.1
## 2              2           20         0.2
```

```
## 3      3      30      0.3
## 4      4      40      0.4
## 5      5      50      0.5
## 6      6      60      0.6
## 7      7      70      0.7
## 8      8      80      0.8
## 9      9      90      0.9
## 10     10     100     1.0
## 11     11     110     1.1
## 12     12     120     1.2
```

```
is.matrix(dfForMat)
```

```
## [1] FALSE
```

```
as.matrix(dfForMat)
```

```
##      vecForMat01 vecForMat02 vecForMat03
## [1,]          1          10          0.1
## [2,]          2          20          0.2
## [3,]          3          30          0.3
## [4,]          4          40          0.4
## [5,]          5          50          0.5
## [6,]          6          60          0.6
## [7,]          7          70          0.7
## [8,]          8          80          0.8
## [9,]          9          90          0.9
## [10,]         10         100          1.0
## [11,]         11         110          1.1
## [12,]         12         120          1.2
```

```
is.matrix(as.matrix(dfForMat))
```

```
## [1] TRUE
```

Nous pouvons aussi créer une `matrix` à partir d'un `vector` avec la fonction `as.matrix()` (matrice de une seule colonne).

```
as.matrix(vecForMat01)
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
## [4,]    4
## [5,]    5
## [6,]    6
## [7,]    7
## [8,]    8
## [9,]    9
## [10,]  10
## [11,]  11
## [12,]  12
```

8.4.2 Manipuler et faire des opérations sur une matrix

Toutes les opérations terme à terme sont possibles sur les matrix.

```
# opérations terme à terme
miMat01 <- matrix(vecForMat01, ncol = 3)
miVecOp <- c(1, 10, 100, 1000)
miMat01 * miVecOp
```

```
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]   20   60  100
## [3,]  300  700 1100
## [4,] 4000 8000 12000
```

```
miMat01 + miVecOp
```

```
##      [,1] [,2] [,3]
## [1,]    2    6   10
## [2,]   12   16   20
## [3,]  103  107  111
## [4,] 1004 1008 1012
```

```
miMat01 / miMat01
```

```
##      [,1] [,2] [,3]
## [1,]    1    1    1
## [2,]    1    1    1
## [3,]    1    1    1
## [4,]    1    1    1
```

```
miMat01 - 10
```

```
##      [,1] [,2] [,3]
## [1,]   -9   -5   -1
## [2,]   -8   -4    0
## [3,]   -7   -3    1
## [4,]   -6   -2    2
```

Pour effectuer des opérations algébriques nous pouvons utiliser la fonction `%*%`.

```
# opérations algébriques
miVecConf <- c(1, 10, 100)
miMat01 %*% miVecConf
```

```
##      [,1]
## [1,]  951
## [2,] 1062
## [3,] 1173
## [4,] 1284
```

```
miMat02 <- matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), ncol = 3)
print(miMat02)
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
miMat02 %*% miMat02
```

```
##      [,1] [,2] [,3]
## [1,]   30   66  102
## [2,]   36   81  126
## [3,]   42   96  150
```

La diagonale d'une `matrix` peut être obtenue avec la fonction `diag()`, et le déterminant d'une `matrix` avec la fonction `det()`.

```
print(miMat02)
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
diag(miMat02)
```

```
## [1] 1 5 9
```

```
det(miMat02)
```

```
## [1] 0
```

Il est souvent utile de pouvoir transposer une `matrix` (colonnes en lignes ou lignes en colonnes). Pour cela il existe les fonctions `aperm()` ou `t()`. la fonction `t()` est plus générique et fonctionne aussi sur les `data.frame`.

```
aperm(miMat01)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    6    7    8
## [3,]    9   10   11   12
```

```
t(miMat01)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    6    7    8
## [3,]    9   10   11   12
```

8.4.3 Accéder aux éléments d'une `matrix`

Comme pour un `data.frame`, nous pouvons accéder aux éléments d'une matrice en spécifiant un numéro de ligne et un numéro de colonne entre crochets simples `[]`, et séparés par une virgule. Si `i` est le numéro de ligne et `j` le numéro de colonne, alors `miMat01[i, j]` renvoie l'élément situé à la ligne `i` et à la colonne `j`. `miMat01[i,]` renvoie tous les éléments de la ligne `i`, et `miMat01[, j]` tous les éléments de la colonne `j`. Les sélections multiples sont possibles. Nous pouvons également

accéder à un élément en fonction de sa position dans la matrice entre crochets simples `[]` en comptant par colonne puis par ligne. Dans notre exemple la valeur du dixième élément est 10.

```
i <- 2
j <- 1
print(miMat01[i, j])
```

```
## [1] 2
```

```
print(miMat01[i, ])
```

```
## [1] 2 6 10
```

```
print(miMat01[, j])
```

```
## [1] 1 2 3 4
```

```
print(miMat01[c(1, 2), c(2, 3)])
```

```
##      [,1] [,2]
## [1,]    5    9
## [2,]    6   10
```

```
print(miMat01[10])
```

```
## [1] 10
```

8.5 Le conteneur array

Le conteneur de type `array` est une généralisation du conteneur de type `matrix`. Là où le type `matrix` a deux dimensions (les lignes et les colonnes), le type `array` a un nombre indéfini de dimensions. Nous pouvons connaître le nombre de dimensions d'un `array` (et donc d'une `matrix`) avec la fonction `dim()`.

```
dim(miMat01)
```

```
## [1] 4 3
```

8.5.1 Créer un array

La création d'un `array` est similaire à celle d'une `matrix` avec une dimension supplémentaire.

```
miVecArr <- c(1, 2, 3, 4, 5, 6, 7, 8, 9)
miArray <- array(miVecArr, dim = c(3, 3, 2))
print(miArray)
```

```
## , , 1
##
##      [,1] [,2] [,3]
## [1,]    1    4    7
```



```
## [2,] 2 5 8
## [3,] 3 6 9
##
## , , 2
##
##      [,1] [,2] [,3]
## [1,] 1 4 7
## [2,] 2 5 8
## [3,] 3 6 9
```

```
dim(miArray)
```

```
## [1] 3 3 2
```

```
is.array(miArray)
```

```
## [1] TRUE
```

```
miVecArr02 <- 10 * miVecArr
miArray02 <- array(c(miVecArr, miVecArr02), dim = c(3, 3, 2))
print(miArray02)
```

```
## , , 1
##
##      [,1] [,2] [,3]
## [1,] 1 4 7
## [2,] 2 5 8
## [3,] 3 6 9
##
## , , 2
##
##      [,1] [,2] [,3]
## [1,] 10 40 70
## [2,] 20 50 80
## [3,] 30 60 90
```

```
dim(miArray02)
```

```
## [1] 3 3 2
```

```
is.array(miArray02)
```

```
## [1] TRUE
```

Nous pouvons donner des noms aux lignes et aux colonnes, mais aussi aux éléments.

```
miArray02 <- array(
  c(miVecArr, miVecArr02),
  dim = c(3, 3, 2),
  dimnames = list(
    c("r1", "r2", "r3"),
    c("c1", "c2", "c3"),
```

```

    c("matrix1", "matrix2")
  )
)
print(miArray02)

```

```

## , , matrix1
##
##      c1 c2 c3
## r1   1  4  7
## r2   2  5  8
## r3   3  6  9
##
## , , matrix2
##
##      c1 c2 c3
## r1  10 40 70
## r2  20 50 80
## r3  30 60 90

```

8.5.2 Manipuler un array

La manipulation d'un array se fait de la même façon que pour une matrix. pour accéder aux différents éléments d'un array, il suffit de spécifier la ligne i, la colonne j, et la matrix k.

```

i <- 2
j <- 1
k <- 1
print(miArray02[i, j, k])

```

```
## [1] 2
```

```
print(miArray02[, j, k])
```

```
## r1 r2 r3
##  1  2  3

```

```
print(miArray02[i, , k])
```

```
## c1 c2 c3
##  2  5  8

```

```
print(miArray02[i, j, ])

```

```
## matrix1 matrix2
##      2      20

```

8.6 Conclusion

Félicitations ! Nous connaissons à présent les principaux types d'objets que nous allons utiliser avec R. Un objet se caractérise par ses attributs :

- le type de conteneur (`vector`, `data.frame`, `matrix`, `array`)
- le type de contenu de chacun des éléments (`numeric`, `logical`, `character`, ...)
- la valeur de chacun des éléments (5, "qwerty", TRUE, ...)

Tous ces objets sont stockés temporairement dans l'environnement global de R (dans la RAM de notre ordinateur). Le prochain chapitre va traiter des fonctions, et mettra en lumière un des aspects qui rend R si puissant pour analyser et gérer nos données.

Chapter 9

Les fonctions

9.1 Qu'est-ce qu'une fonction

Avec ce chapitre nous allons avoir un premier aperçu de la puissance de R grâce aux fonctions. Une fonction est un ensemble de lignes de code permettant d'exécuter une tâche particulière. Nous avons vu de nombreuses fonctions lors des précédents chapitres, la plus simple étant la fonction `+` permettant d'ajouter deux nombres entre eux, ou d'autres plus complexes comme `c()` ou `data.frame()` permettant de créer un `vector` ou un `data.frame`. Dans tous les cas une fonction se reconnaît grâce aux parenthèses qui la suivent dans laquelle nous allons renseigner des **arguments**. Les arguments correspondent aux informations que nous souhaitons transmettre à notre fonction pour qu'elle exécute la tâche que nous souhaitons réaliser.



Pour les fonctions les plus simples comme `+`, les parenthèses ont été supprimées pour que le code soit plus facile à lire, mais il s'agit bien d'une fonction qui peut s'utiliser avec des parenthèses si nous utilisons le signe `+` entre guillemets. Les arguments sont les chiffres que nous souhaitons ajouter.

```
5 + 2
```

```
## [1] 7
```

```
'+'(5, 2)
```

```
## [1] 7
```

Dans ce chapitre nous allons nous focaliser sur les fonctions les plus courantes de façon à ce que ce chapitre soit consultable comme un dictionnaire. Il ne s'agit donc pas de tout apprendre par coeur mais bien de savoir que ces fonctions existent et de pouvoir consulter ce chapitre plus tard comme référence. Avec le temps et la pratique nous finirons par les connaître par coeur ! Il y a plus de 1000 fonctions à ce jour dans la version de base de R, et plus de 10000 packages complémentaires pouvant être installés, chacun contenant plusieurs dizaines de fonctions. Avant de nous lancer dans l'écriture d'une nouvelle fonction, il faudra toujours vérifier qu'elle n'existe pas déjà.

9.2 Les fonctions les plus courantes

Pour travailler avec les fonctions nous allons utiliser le jeu de données `iris` qui est inclus avec la version de base de R et qui correspond à la longueur et à la largeur des sépales et des pétales de différentes espèces d'iris. Le jeu de données est sous la forme d'un `data.frame` de 5 colonnes et de 150 lignes. Pour plus d'information sur le jeu de données `iris` nous pouvons consulter la documentation de R avec la fonction `help(iris)`. L'accès à la documentation est l'objet de la section ci-dessous.

9.2.1 L'accès à la documentation

9.2.1.1 `help()`

La fonction indispensable de R est celle permettant d'accéder à la documentation. Toutes les fonctions et tous les jeux de données de R possèdent une documentation. Nous pouvons accéder à la documentation avec la fonction `help()` ou en utilisant le raccourci `?`.

```
help(matrix) # équivalent à ?matrix
```

La documentation est toujours structurée de la même manière. Tout d'abord nous avons le nom de la fonction recherchée `matrix`, suivie entre accolades par le nom du package R dont la fonction dépend. Nous verrons comment installer des packages additionnels plus tard. Pour l'instant nous disposons de ceux fournis avec la version de base de R. Ici nous pouvons voir que la fonction `matrix()` dépend du package `base`.

Ensuite nous pouvons voir le libellé de la fonction (Matrices), suivi des paragraphes *Description*, *Usage*, et *Arguments*. Parfois vient s'ajouter les paragraphes *Details*, *Note*, *References*, et *See also*. Le dernier paragraphe est *Examples*. La dernière ligne de la documentation permet de revenir à l'index du package dont dépend la fonction consultée.

En copiant collant `help(matrix)` dans notre console R, nous pouvons voir que le paragraphe *Description* indique ce que fait la fonction. Dans le cas de `help(matrix)`, il y a trois fonctions qui sont présentées : `matrix()`, `as.matrix()`, et `is.matrix()`.

```
# Description
# matrix creates a matrix from the given set of values.
# as.matrix attempts to turn its argument into a matrix.
# is.matrix tests if its argument is a (strict) matrix.
```

Le paragraphe *Usage* explique comment utiliser la fonction et quels sont les valeurs par défaut éventuelles pour chacun des paramètres.

```
# Usage
# matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE,
#         dimnames = NULL)
```

La fonction `matrix()` peut prendre 5 arguments : `data`, `nrow`, `ncol`, `byrow`, et `dimnames`. Nous pouvons voir que par défaut une `matrix` sera composée d'une seule ligne et d'une seule colonne, et que les informations seront renseignées par colonne.

Le paragraphe *Arguments* détaille les valeurs et le type de conteneur de chacun des arguments de notre fonction. Par exemple nous pouvons voir que l'argument `dimnames` doit être de type `list`. C'est pourquoi nous avons utilisé ce format lors de la section sur les `matrix`.

```
# Arguments
# data      an optional data vector (including a list or expression vector).
#           Non-atomic classed R objects are coerced by as.vector and all
#           attributes discarded.
# nrow      the desired number of rows.
# ncol      the desired number of columns.
# byrow     logical. If FALSE (the default) the matrix is filled by columns,
#           otherwise the matrix is filled by rows.
# dimnames  A dimnames attribute for the matrix: NULL or a list of length 2
#           giving the row and column names respectively. An empty list is
#           treated as NULL, and a list of length one as row names. The
```

```
#          list can be named, and the list names will be used as names for
#          the dimensions.
```

Le paragraphe Details apporte des éléments complémentaires sur la fonction. Le paragraphe Examples procure des exemples reproductibles dans la console.

```
## Example of setting row and column names
mdat <- matrix(c(1,2,3, 11,12,13), nrow = 2, ncol = 3, byrow = TRUE,
               dimnames = list(c("row1", "row2"),
                               c("C.1", "C.2", "C.3")))
mdat
```

```
##      C.1 C.2 C.3
## row1   1   2   3
## row2  11  12  13
```



Le nom des arguments n'est pas nécessaire pour qu'une fonction soit correctement interprétée par R. Néanmoins par soucis de clarté il est préférable d'utiliser le nom des arguments suivi du signe = pour que le code soit plus lisible.

```
# bon exemple
mdat <- matrix(c(1,2,3, 11,12,13), nrow = 2, ncol = 3, byrow = TRUE)
# mauvais exemple
mdat <- matrix(c(1,2,3, 11,12,13), 2, 3, TRUE)
```

9.2.1.2 help.search()

La fonction `help.search()` ou `??` permet de rechercher une expression dans l'ensemble de la documentation. Elle est utile lorsque l'on cherche une fonctionnalité sans connaître le nom de la fonction sous R.

```
help.search("average")
```

La fonction `help.search()` renvoie vers une page contenant la liste des pages où l'expression a été retrouvée sous la forme `nom-du-package::nom-de-la-fonction`.

9.2.2 Visualiser les données

9.2.2.1 str()

La fonction `str()` permet de visualiser la structure interne d'un objet, comme indiqué dans la documentation que nous pouvons consulter avec `help(str)`.

```
str(iris)
```

```
## 'data.frame':   150 obs. of  5 variables:
## $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
## $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
## $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
## $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
## $ Species     : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

La fonction `str()` renvoie le type d'objet (`data.frame`), le nombre d'observations (150), le nombre de variables (5), le nom de chacune des variables (`Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, et `Species`), le type de chacune des variables (`num`, `Factor`), et les premières valeurs de chacune des variables. C'est une fonction utile pour avoir un aperçu d'un jeu de données, mais aussi pour contrôler que les données sont du type voulu avant de procéder à des analyses statistiques.

9.2.2.2 `head()` et `tail()`

La fonction `head()` renvoie les premières valeurs d'un objet, et la fonction `tail()` les dernières valeurs d'un objet. Par défaut six valeurs sont retournées, l'argument `n` contrôle le nombre de valeurs à retourner.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1         3.5         1.4         0.2   setosa
## 2         4.9         3.0         1.4         0.2   setosa
## 3         4.7         3.2         1.3         0.2   setosa
## 4         4.6         3.1         1.5         0.2   setosa
## 5         5.0         3.6         1.4         0.2   setosa
## 6         5.4         3.9         1.7         0.4   setosa
```

```
tail(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 145         6.7         3.3         5.7         2.5 virginica
## 146         6.7         3.0         5.2         2.3 virginica
## 147         6.3         2.5         5.0         1.9 virginica
## 148         6.5         3.0         5.2         2.0 virginica
## 149         6.2         3.4         5.4         2.3 virginica
## 150         5.9         3.0         5.1         1.8 virginica
```

```
head(iris, n = 2)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1         3.5         1.4         0.2   setosa
## 2         4.9         3.0         1.4         0.2   setosa
```

9.2.2.3 `names()`

Nous avons déjà vu la fonction `names()` qui permet à la fois de connaître le nom des éléments d'un objet, mais aussi d'assigner des noms aux éléments d'un objet.

```
names(iris)
```

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

```
irisCopy <- iris
names(irisCopy) <- c("a", "b", "c", "d", "e")
names(irisCopy)
```

```
## [1] "a" "b" "c" "d" "e"
```


9.2.2.4 cat() et print()

La fonction `cat()` permet d'afficher le contenu d'un objet alors que la fonction `print()` retourne la valeur d'un objet avec la possibilité d'effectuer des conversions.

```
cat(names(iris))

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species

print(names(iris))

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

cat(iris[1, 1])

## 5.1

print(iris[1, 1])

## [1] 5.1

print(iris[1, 1], digits = 0)

## [1] 5
```

9.2.3 Manipuler les données**9.2.3.1 rank()**

La fonction `rank()` renvoie pour un ensemble d'éléments le numéro de la position de chacun des éléments. En cas d'éléments de même valeur, l'argument `ties.method` permet de faire un choix sur le classement. Comme pour toutes les fonctions, les détails sont présents dans la documentation.

```
vecManip <- c(10, 20, 30, 70, 60, 50, 40)
rank(vecManip)

## [1] 1 2 3 7 6 5 4

vecManip2 <- c(10, 20, 30, 10, 50, 10, 40)
rank(vecManip2)

## [1] 2 4 5 2 7 2 6

rank(vecManip2, ties.method = "first")

## [1] 1 4 5 2 7 3 6

rank(vecManip2, ties.method = "min")
```

```
## [1] 1 4 5 1 7 1 6
```

```
print(iris[, 1])
```

```
## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4
## [18] 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5
## [35] 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0
## [52] 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8
## [69] 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0 5.4
## [86] 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8
## [103] 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7
## [120] 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7
## [137] 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5.9
```

```
rank(iris[, 1], ties.method = "average")
```

```
## [1] 37.0 19.5 10.5 7.5 27.5 49.5 7.5 27.5 3.0 19.5 49.5
## [12] 14.0 14.0 1.0 77.0 69.5 49.5 37.0 69.5 37.0 49.5 37.0
## [23] 7.5 37.0 14.0 27.5 27.5 43.5 43.5 10.5 14.0 49.5 43.5
## [34] 56.0 19.5 27.5 56.0 19.5 3.0 37.0 27.5 5.0 3.0 27.5
## [45] 37.0 14.0 37.0 7.5 46.0 27.5 138.0 112.0 135.5 56.0 118.0
## [56] 69.5 104.0 19.5 121.5 43.5 27.5 82.0 86.5 92.5 62.5 126.5
## [67] 62.5 77.0 97.5 62.5 82.0 92.5 104.0 92.5 112.0 121.5 132.0
## [78] 126.5 86.5 69.5 56.0 56.0 77.0 86.5 49.5 86.5 126.5 104.0
## [89] 62.5 56.0 56.0 92.5 77.0 27.5 62.5 69.5 69.5 97.5 37.0
## [100] 69.5 104.0 77.0 139.0 104.0 118.0 145.0 19.5 143.0 126.5 141.0
## [111] 118.0 112.0 132.0 69.5 77.0 112.0 118.0 147.5 147.5 86.5 135.5
## [122] 62.5 147.5 104.0 126.5 141.0 97.5 92.5 112.0 141.0 144.0 150.0
## [133] 112.0 104.0 92.5 147.5 104.0 112.0 86.5 135.5 126.5 135.5 77.0
## [144] 132.0 126.5 126.5 104.0 118.0 97.5 82.0
```

```
# help(rank)
# ...
# Usage
# rank(x, na.last = TRUE,
#      ties.method = c("average", "first", "last",
#                      "random", "max", "min"))
```

9.2.3.2 order()

La fonction `order()` retourne le numéro du réarrangement des éléments en fonction de leur position. Elle est très utile par exemple pour trier un `data.frame` en fonction d'une colonne.

```
print(vecManip2)
```

```
## [1] 10 20 30 10 50 10 40
```

```
rank(vecManip2)
```

```
## [1] 2 4 5 2 7 2 6
```

```
order(vecManip2)
```

```
## [1] 1 4 6 2 3 7 5
```

```
print(iris[, 1])
```

```
## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4
## [18] 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5
## [35] 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0
## [52] 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8
## [69] 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0 5.4
## [86] 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8
## [103] 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7
## [120] 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7
## [137] 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5.9
```

```
rank(iris[, 1])
```

```
## [1] 37.0 19.5 10.5 7.5 27.5 49.5 7.5 27.5 3.0 19.5 49.5
## [12] 14.0 14.0 1.0 77.0 69.5 49.5 37.0 69.5 37.0 49.5 37.0
## [23] 7.5 37.0 14.0 27.5 27.5 43.5 43.5 10.5 14.0 49.5 43.5
## [34] 56.0 19.5 27.5 56.0 19.5 3.0 37.0 27.5 5.0 3.0 27.5
## [45] 37.0 14.0 37.0 7.5 46.0 27.5 138.0 112.0 135.5 56.0 118.0
## [56] 69.5 104.0 19.5 121.5 43.5 27.5 82.0 86.5 92.5 62.5 126.5
## [67] 62.5 77.0 97.5 62.5 82.0 92.5 104.0 92.5 112.0 121.5 132.0
## [78] 126.5 86.5 69.5 56.0 56.0 77.0 86.5 49.5 86.5 126.5 104.0
## [89] 62.5 56.0 56.0 92.5 77.0 27.5 62.5 69.5 69.5 97.5 37.0
## [100] 69.5 104.0 77.0 139.0 104.0 118.0 145.0 19.5 143.0 126.5 141.0
## [111] 118.0 112.0 132.0 69.5 77.0 112.0 118.0 147.5 147.5 86.5 135.5
## [122] 62.5 147.5 104.0 126.5 141.0 97.5 92.5 112.0 141.0 144.0 150.0
## [133] 112.0 104.0 92.5 147.5 104.0 112.0 86.5 135.5 126.5 135.5 77.0
## [144] 132.0 126.5 126.5 104.0 118.0 97.5 82.0
```

```
order(iris[, 1])
```

```
## [1] 14 9 39 43 42 4 7 23 48 3 30 12 13 25 31 46 2
## [18] 10 35 38 58 107 5 8 26 27 36 41 44 50 61 94 1 18
## [35] 20 22 24 40 45 47 99 28 29 33 60 49 6 11 17 21 32
## [52] 85 34 37 54 81 82 90 91 65 67 70 89 95 122 16 19 56
## [69] 80 96 97 100 114 15 68 83 93 102 115 143 62 71 150 63 79
## [86] 84 86 120 139 64 72 74 92 128 135 69 98 127 149 57 73 88
## [103] 101 104 124 134 137 147 52 75 112 116 129 133 138 55 105 111 117
## [120] 148 59 76 66 78 87 109 125 141 145 146 77 113 144 53 121 140
## [137] 142 51 103 110 126 130 108 131 106 118 119 123 136 132
```

```
head(iris[order(iris[, 1]),], n = 10)
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 14          4.3          3.0          1.1          0.1 setosa
## 9           4.4          2.9          1.4          0.2 setosa
## 39          4.4          3.0          1.3          0.2 setosa
## 43          4.4          3.2          1.3          0.2 setosa
```

```
## 42      4.5      2.3      1.3      0.3 setosa
## 4       4.6      3.1      1.5      0.2 setosa
## 7       4.6      3.4      1.4      0.3 setosa
## 23      4.6      3.6      1.0      0.2 setosa
## 48      4.6      3.2      1.4      0.2 setosa
## 3       4.7      3.2      1.3      0.2 setosa
```

9.2.3.3 sort()

La fonction `sort()` permet de trier les éléments d'un objet. Elle ne permet pas de trier selon plusieurs variables comme c'est le cas avec `order()`.

```
print(vecManip2)
```

```
## [1] 10 20 30 10 50 10 40
```

```
sort(vecManip2)
```

```
## [1] 10 10 10 20 30 40 50
```

```
print(iris[, 1])
```

```
## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4
## [18] 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5
## [35] 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0
## [52] 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8
## [69] 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0 5.4
## [86] 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8
## [103] 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7
## [120] 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7
## [137] 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5.9
```

```
sort(iris[, 1])
```

```
## [1] 4.3 4.4 4.4 4.4 4.5 4.6 4.6 4.6 4.6 4.7 4.7 4.8 4.8 4.8 4.8 4.8 4.9
## [18] 4.9 4.9 4.9 4.9 4.9 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.1 5.1
## [35] 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.2 5.2 5.2 5.2 5.3 5.4 5.4 5.4 5.4 5.4
## [52] 5.4 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.6 5.6 5.6 5.6 5.6 5.6 5.7 5.7 5.7
## [69] 5.7 5.7 5.7 5.7 5.7 5.8 5.8 5.8 5.8 5.8 5.8 5.8 5.9 5.9 5.9 6.0 6.0
## [86] 6.0 6.0 6.0 6.0 6.1 6.1 6.1 6.1 6.1 6.1 6.1 6.2 6.2 6.2 6.2 6.3 6.3
## [103] 6.3 6.3 6.3 6.3 6.3 6.3 6.4 6.4 6.4 6.4 6.4 6.4 6.4 6.5 6.5 6.5 6.5
## [120] 6.5 6.6 6.6 6.7 6.7 6.7 6.7 6.7 6.7 6.7 6.7 6.8 6.8 6.8 6.9 6.9 6.9
## [137] 6.9 7.0 7.1 7.2 7.2 7.2 7.3 7.4 7.6 7.7 7.7 7.7 7.7 7.7 7.9
```

9.2.3.4 append()

Cette fonction permet d'ajouter un élément à un vector à une position déterminée par l'argument `after`. Cette fonction est aussi plus rapide que son alternative consistant à utiliser la fonction `c()`.

```
print(vecManip2)
```

```
## [1] 10 20 30 10 50 10 40
```

```
append(vecManip2, 5)
```

```
## [1] 10 20 30 10 50 10 40 5
```

```
append(vecManip2, 5, after = 2)
```

```
## [1] 10 20 5 30 10 50 10 40
```

9.2.3.5 cbind() et rbind()

Les fonctions `cbind()` et `rbind()` permettent de combiner des éléments par colonne ou par ligne.

```
cbind(vecManip2, vecManip2)
```

```
##      vecManip2 vecManip2
## [1,]         10         10
## [2,]         20         20
## [3,]         30         30
## [4,]         10         10
## [5,]         50         50
## [6,]         10         10
## [7,]         40         40
```

```
rbind(vecManip2, vecManip2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## vecManip2 10  20  30  10  50  10  40
## vecManip2 10  20  30  10  50  10  40
```

9.2.3.6 paste() et paste0()

Voilà deux fonctions que nous allons beaucoup utiliser par la suite. Les fonctions `paste()` et `paste0()` permettent de concaténer des chaînes de caractère. La fonction `paste0()` est équivalente à `paste()` sans proposer de séparateur entre les éléments à concaténer. Elle est aussi plus rapide.

```
paste(1, "a")
```

```
## [1] "1 a"
```

```
paste0(1, "a")
```

```
## [1] "1a"
```

```
paste(1, "a", sep = "_")
```

```
## [1] "1_a"
```

```
paste0("prefix_", vecManip2, "_suffix")
```

```
## [1] "prefix_10_suffix" "prefix_20_suffix" "prefix_30_suffix"
## [4] "prefix_10_suffix" "prefix_50_suffix" "prefix_10_suffix"
## [7] "prefix_40_suffix"
```

```
paste(vecManip2, rank(vecManip2), sep = "_")
```

```
## [1] "10_2" "20_4" "30_5" "10_2" "50_7" "10_2" "40_6"
```

9.2.3.7 rev()

La fonction `rev()` renvoie les éléments d'un objet dans l'ordre inverse.

```
print(vecManip2)
```

```
## [1] 10 20 30 10 50 10 40
```

```
rev(vecManip2)
```

```
## [1] 40 10 50 10 30 20 10
```

9.2.3.8 %in%()

La fonction `%in%()` peut être assimilée à un opérateur de comparaison. Cette fonction prend deux objets comme arguments et renvoie TRUE ou FALSE pour chacun des éléments du premier objet en fonction de leur présence ou absence dans le second objet. Pour accéder à la documentation de la fonction, il faut utiliser des guillemets `help('%in%')`.

```
print(vecManip)
```

```
## [1] 10 20 30 70 60 50 40
```

```
print(vecManip2)
```

```
## [1] 10 20 30 10 50 10 40
```

```
vecManip %in% vecManip2
```

```
## [1] TRUE TRUE TRUE FALSE FALSE TRUE TRUE
```

```
vecManip2 %in% vecManip
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

9.2.4 Fonctions mathématiques

Nous avons déjà vu les fonctions `+`, `-`, `*`, `/`, `^`, `%%` et autres opérateurs arithmétiques. R possède également les fonctions mathématiques de base comme exponentielle `exp()`, racine carrée `sqrt()`, valeur absolue `abs()`, sinus `sin()`, cosinus

`cos()`, tangente `tan()`, logarithme népérien `log()`, logarithme décimal `log10()`, arc cosinus `acos()`, arc sinus `asin()`, et arc tangente `atan()`.

```
print(vecManip2)
```

```
## [1] 10 20 30 10 50 10 40
```

```
exp(vecManip2)
```

```
## [1] 2.202647e+04 4.851652e+08 1.068647e+13 2.202647e+04 5.184706e+21
## [6] 2.202647e+04 2.353853e+17
```

```
sqrt(vecManip2)
```

```
## [1] 3.162278 4.472136 5.477226 3.162278 7.071068 3.162278 6.324555
```

```
abs(-vecManip2)
```

```
## [1] 10 20 30 10 50 10 40
```

```
sin(vecManip2)
```

```
## [1] -0.5440211 0.9129453 -0.9880316 -0.5440211 -0.2623749 -0.5440211
## [7] 0.7451132
```

```
cos(vecManip2)
```

```
## [1] -0.8390715 0.4080821 0.1542514 -0.8390715 0.9649660 -0.8390715
## [7] -0.6669381
```

```
tan(vecManip2)
```

```
## [1] 0.6483608 2.2371609 -6.4053312 0.6483608 -0.2719006 0.6483608
## [7] -1.1172149
```

```
log(vecManip2)
```

```
## [1] 2.302585 2.995732 3.401197 2.302585 3.912023 2.302585 3.688879
```

```
log10(vecManip2)
```

```
## [1] 1.000000 1.301030 1.477121 1.000000 1.698970 1.000000 1.602060
```

```
acos(vecManip2/100)
```

```
## [1] 1.470629 1.369438 1.266104 1.470629 1.047198 1.470629 1.159279
```

```
asin(vecManip2/100)
```

```
## [1] 0.1001674 0.2013579 0.3046927 0.1001674 0.5235988 0.1001674 0.4115168
```

```
atan(vecManip2/100)
```

```
## [1] 0.09966865 0.19739556 0.29145679 0.09966865 0.46364761 0.09966865
## [7] 0.38050638
```

9.2.5 Statistiques descriptives

Nous pouvons également effectuer des statistiques descriptives très simplement à partir d'un jeu de données.

9.2.5.1 mean()

La fonction `mean()` renvoie la moyenne. Pour ignorer les valeurs manquantes `NA`, il faut donner la valeur `TRUE` à l'argument `na.rm()`.

```
mean(iris[, 1])
```

```
## [1] 5.843333
```

```
vecManip3 <- c(1, 5, 6, 8, NA, 45, NA, 14)
mean(vecManip3)
```

```
## [1] NA
```

```
mean(vecManip3, na.rm = TRUE)
```

```
## [1] 13.16667
```

9.2.5.2 sd()

La fonction `sd()` renvoie l'écart type.

```
sd(iris[, 1])
```

```
## [1] 0.8280661
```

```
print(vecManip3)
```

```
## [1] 1 5 6 8 NA 45 NA 14
```

```
sd(vecManip3)
```

```
## [1] NA
```

```
sd(vecManip3, na.rm = TRUE)
```

```
## [1] 16.16684
```


9.2.5.3 max()

La fonction `max()` renvoie la valeur maximale et `min()` la valeur minimale.

```
max(iris[, 1])
```

```
## [1] 7.9
```

```
print(vecManip3)
```

```
## [1] 1 5 6 8 NA 45 NA 14
```

```
max(vecManip3)
```

```
## [1] NA
```

```
max(vecManip3, na.rm = TRUE)
```

```
## [1] 45
```

```
min(iris[, 1])
```

```
## [1] 4.3
```

```
min(vecManip3)
```

```
## [1] NA
```

```
min(vecManip3, na.rm = TRUE)
```

```
## [1] 1
```

9.2.5.4 quantile()

La fonction `quantile()` renvoie le quantile défini par l'argument `probs`.

```
quantile(iris[, 1])
```

```
## 0% 25% 50% 75% 100%
## 4.3 5.1 5.8 6.4 7.9
```

```
quantile(iris[, 1], probs = c(0, 0.25, 0.5, 0.75, 1))
```

```
## 0% 25% 50% 75% 100%
## 4.3 5.1 5.8 6.4 7.9
```

```
quantile(iris[, 1], probs = c(0, 0.1, 0.5, 0.9, 1))
```

```
## 0% 10% 50% 90% 100%
## 4.3 4.8 5.8 6.9 7.9
```

9.2.5.5 summary()

La fonction `summary()` renvoie un résumé composé du minimum, premier quartile, médiane, moyenne, troisième quartile et maximum.

```
summary(iris[, 1])
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.300   5.100   5.800   5.843   6.400   7.900
```

9.2.5.6 median()

La fonction `median()` renvoie la médiane.

```
median(iris[, 1])
```

```
## [1] 5.8
```

```
print(vecManip3)
```

```
## [1]  1  5  6  8 NA 45 NA 14
```

```
median(vecManip3)
```

```
## [1] NA
```

```
median(vecManip3, na.rm = TRUE)
```

```
## [1] 7
```

9.2.5.7 length()

La fonction `length()` renvoie la taille d'un objet (nombre d'éléments).

```
length(iris[, 1])
```

```
## [1] 150
```

```
print(vecManip3)
```

```
## [1]  1  5  6  8 NA 45 NA 14
```

```
length(vecManip3)
```

```
## [1] 8
```

9.2.5.8 `nrow()` et `ncol()`

La fonction `nrow()` renvoie le nombre de lignes et la fonction `ncol()` le nombre de colonnes d'un objet.

```
nrow(iris)
```

```
## [1] 150
```

```
ncol(iris)
```

```
## [1] 5
```

9.2.5.9 `round()`, `ceiling()`, `floor()`, et `trunc()`

La fonction `round()` permet de sélectionner un certain nombre de décimales (0 par défaut)

```
round(5.56874258564)
```

```
## [1] 6
```

```
round(5.56874258564, digits = 2)
```

```
## [1] 5.57
```

La fonction `ceiling()` renvoie le plus petit nombre entier qui ne soit pas inférieure à la valeur renseignée.

```
ceiling(5.9999)
```

```
## [1] 6
```

```
ceiling(5.0001)
```

```
## [1] 6
```

La fonction `floor()` renvoie le plus grand nombre entier qui ne soit pas supérieure à la valeur renseignée.

```
floor(5.9999)
```

```
## [1] 5
```

```
floor(5.0001)
```

```
## [1] 5
```

La fonction `trunc()` renvoie la partie entière de la valeur renseignée.

```
trunc(5.9999)
```

```
## [1] 5
```

```
trunc(5.0001)
```

```
## [1] 5
```

9.2.5.10 rowSums() et colSums()

Les fonctions `rowSums()` et `colSums()` calculent la somme des lignes et des colonnes.

```
rowSums(iris[, c(1, 2, 3, 4)])
```

```
##      [1] 10.2  9.5  9.4  9.4 10.2 11.4  9.7 10.1  8.9  9.6 10.8 10.0  9.3  8.5
##     [15] 11.2 12.0 11.0 10.3 11.5 10.7 10.7 10.7  9.4 10.6 10.3  9.8 10.4 10.4
##     [29] 10.2  9.7  9.7 10.7 10.9 11.3  9.7  9.6 10.5 10.0  8.9 10.2 10.1  8.4
##     [43]  9.1 10.7 11.2  9.5 10.7  9.4 10.7  9.9 16.3 15.6 16.4 13.1 15.4 14.3
##     [57] 15.9 11.6 15.4 13.2 11.5 14.6 13.2 15.1 13.4 15.6 14.6 13.6 14.4 13.1
##     [71] 15.7 14.2 15.2 14.8 14.9 15.4 15.8 16.4 14.9 12.8 12.8 12.6 13.6 15.4
##     [85] 14.4 15.5 16.0 14.3 14.0 13.3 13.7 15.1 13.6 11.6 13.8 14.1 14.1 14.7
##     [99] 11.7 13.9 18.1 15.5 18.1 16.6 17.5 19.3 13.6 18.3 16.8 19.4 16.8 16.3
##    [113] 17.4 15.2 16.1 17.2 16.8 20.4 19.5 14.7 18.1 15.3 19.2 15.7 17.8 18.2
##    [127] 15.6 15.8 16.9 17.6 18.2 20.1 17.0 15.7 15.7 19.1 17.7 16.8 15.6 17.5
##    [141] 17.8 17.4 15.5 18.2 18.2 17.2 15.7 16.7 17.3 15.8
```

```
colSums(iris[, c(1, 2, 3, 4)])
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
##           876.5           458.6           563.7           179.9
```

9.2.5.11 rowMeans() et colMeans()

Les fonctions `rowMeans()` et `colMeans()` calculent la moyenne des lignes et des colonnes.

```
rowMeans(iris[, c(1, 2, 3, 4)])
```

```
##      [1] 2.550 2.375 2.350 2.350 2.550 2.850 2.425 2.525 2.225 2.400 2.700
##     [12] 2.500 2.325 2.125 2.800 3.000 2.750 2.575 2.875 2.675 2.675 2.675
##     [23] 2.350 2.650 2.575 2.450 2.600 2.600 2.550 2.425 2.425 2.675 2.725
##     [34] 2.825 2.425 2.400 2.625 2.500 2.225 2.550 2.525 2.100 2.275 2.675
##     [45] 2.800 2.375 2.675 2.350 2.675 2.475 4.075 3.900 4.100 3.275 3.850
##     [56] 3.575 3.975 2.900 3.850 3.300 2.875 3.650 3.300 3.775 3.350 3.900
##     [67] 3.650 3.400 3.600 3.275 3.925 3.550 3.800 3.700 3.725 3.850 3.950
##     [78] 4.100 3.725 3.200 3.200 3.150 3.400 3.850 3.600 3.875 4.000 3.575
##     [89] 3.500 3.325 3.425 3.775 3.400 2.900 3.450 3.525 3.525 3.675 2.925
##    [100] 3.475 4.525 3.875 4.525 4.150 4.375 4.825 3.400 4.575 4.200 4.850
##    [111] 4.200 4.075 4.350 3.800 4.025 4.300 4.200 5.100 4.875 3.675 4.525
##    [122] 3.825 4.800 3.925 4.450 4.550 3.900 3.950 4.225 4.400 4.550 5.025
##    [133] 4.250 3.925 3.925 4.775 4.425 4.200 3.900 4.375 4.450 4.350 3.875
##    [144] 4.550 4.550 4.300 3.925 4.175 4.325 3.950
```

```
colMeans(iris[, c(1, 2, 3, 4)])
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
##      5.843333      3.057333      3.758000      1.199333
```

9.2.5.12 aggregate()

La fonction `aggregate()` permet de grouper les éléments d'un objet en fonction d'une valeur. L'argument `by` définit l'élément sur lequel est effectué le regroupement. Il doit être de type `list`.

```
aggregate(iris[, c(1, 2, 3, 4)], by = list(iris$Species), FUN = mean)
```

```
##      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1      setosa      5.006      3.428      1.462      0.246
## 2 versicolor      5.936      2.770      4.260      1.326
## 3 virginica      6.588      2.974      5.552      2.026
```

```
aggregate(iris[, c(1, 2)], by = list(iris$Species), FUN = summary)
```

```
##      Group.1 Sepal.Length.Min. Sepal.Length.1st Qu. Sepal.Length.Median
## 1      setosa      4.300      4.800      5.000
## 2 versicolor      4.900      5.600      5.900
## 3 virginica      4.900      6.225      6.500
##      Sepal.Length.Mean Sepal.Length.3rd Qu. Sepal.Length.Max.
## 1      5.006      5.200      5.800
## 2      5.936      6.300      7.000
## 3      6.588      6.900      7.900
##      Sepal.Width.Min. Sepal.Width.1st Qu. Sepal.Width.Median Sepal.Width.Mean
## 1      2.300      3.200      3.400      3.428
## 2      2.000      2.525      2.800      2.770
## 3      2.200      2.800      3.000      2.974
##      Sepal.Width.3rd Qu. Sepal.Width.Max.
## 1      3.675      4.400
## 2      3.000      3.400
## 3      3.175      3.800
```

9.2.5.13 range()

La fonction `range()` renvoie le minimum et le maximum.

```
range(iris[, 1])
```

```
## [1] 4.3 7.9
```

```
print(vecManip3)
```

```
## [1] 1 5 6 8 NA 45 NA 14
```

```
range(vecManip3)
```

```
## [1] NA NA
```

```
range(vecManip3, na.rm = TRUE)
```

```
## [1] 1 45
```

9.2.5.14 unique()

La fonction `unique()` renvoie les valeurs uniques d'un objet (sans les doublons).

```
unique(iris[, 1])
```

```
## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.4 4.8 4.3 5.8 5.7 5.2 5.5 4.5 5.3 7.0 6.4
## [18] 6.9 6.5 6.3 6.6 5.9 6.0 6.1 5.6 6.7 6.2 6.8 7.1 7.6 7.3 7.2 7.7 7.4
## [35] 7.9
```

```
print(vecManip3)
```

```
## [1] 1 5 6 8 NA 45 NA 14
```

```
unique(vecManip3)
```

```
## [1] 1 5 6 8 NA 45 14
```

9.3 Autres fonctions utiles

Nous ne pouvons aborder toutes les fonctions utiles, ici nous ne ferons qu'aborder certaines fonctions. Tout au long de ce livre de nouvelles fonctions seront utilisées. Lorsqu'une nouvelle fonction est utilisée, notre réflexe doit être toujours le même : consulter la documentation avec la fonction `help()`.

9.3.1 seq_along()

La fonction `seq_along()` permet de créer un vector de la taille de l'objet renseigné et prenant comme valeurs les chiffres de 1 à N (N correspondant aux nombres d'éléments de l'objet). Cette fonction nous servira beaucoup lors du chapitre sur les boucles.

```
print(vecManip3)
```

```
## [1] 1 5 6 8 NA 45 NA 14
```

```
seq_along(vecManip3)
```

```
## [1] 1 2 3 4 5 6 7 8
```

9.3.2 :

La fonction `:` permet de créer une séquence de a à b par pas de 1, avec a et b le début et la fin de la séquence souhaitée. Il a été difficile d'écrire les chapitres précédents sans y avoir recours tant cette fonction est utile. Voici quelques exemples.

```
5:10
```

```
## [1] 5 6 7 8 9 10
```

```
head(iris[, c(1, 2, 3, 4)])
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1          3.5          1.4          0.2
## 2          4.9          3.0          1.4          0.2
## 3          4.7          3.2          1.3          0.2
## 4          4.6          3.1          1.5          0.2
## 5          5.0          3.6          1.4          0.2
## 6          5.4          3.9          1.7          0.4
```

```
head(iris[, 1:4]) # ; -)
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1          3.5          1.4          0.2
## 2          4.9          3.0          1.4          0.2
## 3          4.7          3.2          1.3          0.2
## 4          4.6          3.1          1.5          0.2
## 5          5.0          3.6          1.4          0.2
## 6          5.4          3.9          1.7          0.4
```

```
miVec01 <- c(1, 2, 3, 4)
miVec01 <- 1:4 # ; -)
-10:12
```

```
## [1] -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
## [18] 7 8 9 10 11 12
```

```
5:-5
```

```
## [1] 5 4 3 2 1 0 -1 -2 -3 -4 -5
```

```
paste("X", 1:10, sep = "_")
```

```
## [1] "X_1" "X_2" "X_3" "X_4" "X_5" "X_6" "X_7" "X_8" "X_9" "X_10"
```

9.3.3 rep()

la fonction `rep()` permet de répéter des éléments.

```
miVec12 <- c(1, 1, 1, 1, 1, 1, 1, 1, 1)
miVec12 <- rep(1, times = 9) # ; -)
rep("Hola", times = 3)
```

```
## [1] "Hola" "Hola" "Hola"
```

```
rep(1:3, time = 3)
```

```
## [1] 1 2 3 1 2 3 1 2 3
```

```
rep(1:3, length.out = 10)
```

```
## [1] 1 2 3 1 2 3 1 2 3 1
```

```
rep(1:3, each = 3)
```

```
## [1] 1 1 1 2 2 2 3 3 3
```

9.3.4 seq()

la fonction `seq()` permet de créer une séquence personnalisée.

```
seq(from = 0, to = 1, by = 0.2)
```

```
## [1] 0.0 0.2 0.4 0.6 0.8 1.0
```

```
seq(from = 20, to = 10, length.out = 10)
```

```
## [1] 20.00000 18.88889 17.77778 16.66667 15.55556 14.44444 13.33333
## [8] 12.22222 11.11111 10.00000
```

```
letters[seq(from = 1, to = 26, by = 2)]
```

```
## [1] "a" "c" "e" "g" "i" "k" "m" "o" "q" "s" "u" "w" "y"
```

```
rep(seq(from = 1, to = 2, by = 0.5), times = 3)
```

```
## [1] 1.0 1.5 2.0 1.0 1.5 2.0 1.0 1.5 2.0
```

9.3.5 getwd()

La fonction `getwd()` définit le répertoire de travail. Cela correspond à l'endroit relatif à partir duquel R se positionne pour identifier les fichiers. Ce concept prendra son sens lorsque nous verrons comment importer et exporter des données.

```
getwd()
```

```
## [1] "C:/Users/nous/Documents/Francois/TRAVAIL/GitHub/myRBook_FR"
```

9.3.6 setwd()

La fonction `setwd()` permet de définir un nouveau répertoire de travail.

```
oldWd <- getwd()
print(oldWd)
```

```
## [1] "C:/Users/nous/Documents/Francois/TRAVAIL/GitHub/myRBook_FR"
```

```
setwd("../")
getwd()
```

```
## [1] "C:/Users/nous/Documents/Francois/TRAVAIL/GitHub"
```



```
setwd(oldWd)
getwd()
```

```
## [1] "C:/Users/nous/Documents/Francois/TRAVAIL/GitHub/myRBook_FR"
```

9.3.7 list.files()

La fonction `list.files()` permet de faire la liste de tous les fichiers présents dans le répertoire de travail.

```
list.files(pattern = "(pdf)$") # pdf
```

```
## character(0)
```

9.3.8 ls()

Tout comme la fonction `list.files()` permet de faire la liste de tous les fichiers présents dans le répertoire de travail, la fonction `ls()` permet de faire la liste de tous les objets présents dans l'environnement de travail de R.

```
ls()
```

```
## [1] "aLogic"      "bddInsect"   "characters"  "contrib"
## [5] "dfForMat"    "factor01"    "i"           "irisCopy"
## [9] "j"           "k"           "logicals"    "mdat"
## [13] "miArray"     "miArray02"   "miDf01"      "miDfSub01"
## [17] "miDfSub02"   "miDfSub03"   "miDfSub04"   "miList01"
## [21] "miList02"    "miList0203"  "miList03"    "miList04"
## [25] "miList05"    "miList06"    "miMat"       "miMat01"
## [29] "miMat02"     "miVec01"     "miVec02"     "miVec03"
## [33] "miVec04"     "miVec05"     "miVec06"     "miVec07"
## [37] "miVec08"     "miVec09"     "miVec10"     "miVec11"
## [41] "miVec12"     "miVec13"     "miVec14"     "miVec15"
## [45] "miVec20"     "miVec21"     "miVec22"     "miVec23"
## [49] "miVec24"     "miVec25"     "miVecArr"    "miVecArr02"
## [53] "miVecConf"   "miVecNA"     "miVecOp"     "msg"
## [57] "myCol"       "myRow"       "myText"      "myText2"
## [61] "myText3"     "myText4"     "myText5"     "nbrRep"
## [65] "newVec"      "newVec2"     "numbers"     "oldWd"
## [69] "opAriDf"     "sumIntDou"   "sumIntInt"   "terme01"
## [73] "terme02"     "vecForMat01" "vecForMat02" "vecForMat03"
## [77] "vecForMatrix" "vecManip"    "vecManip2"   "vecManip3"
```

```
zzz <- "a new object"
ls()
```

```
## [1] "aLogic"      "bddInsect"   "characters"  "contrib"
## [5] "dfForMat"    "factor01"    "i"           "irisCopy"
## [9] "j"           "k"           "logicals"    "mdat"
## [13] "miArray"     "miArray02"   "miDf01"      "miDfSub01"
## [17] "miDfSub02"   "miDfSub03"   "miDfSub04"   "miList01"
## [21] "miList02"    "miList0203"  "miList03"    "miList04"
## [25] "miList05"    "miList06"    "miMat"       "miMat01"
```

```
## [29] "miMat02"      "miVec01"      "miVec02"      "miVec03"
## [33] "miVec04"      "miVec05"      "miVec06"      "miVec07"
## [37] "miVec08"      "miVec09"      "miVec10"      "miVec11"
## [41] "miVec12"      "miVec13"      "miVec14"      "miVec15"
## [45] "miVec20"      "miVec21"      "miVec22"      "miVec23"
## [49] "miVec24"      "miVec25"      "miVecArr"     "miVecArr02"
## [53] "miVecConf"    "miVecNA"      "miVecOp"      "msg"
## [57] "myCol"        "myRow"        "myText"       "myText2"
## [61] "myText3"      "myText4"      "myText5"      "nbrRep"
## [65] "newVec"       "newVec2"      "numbers"      "oldWd"
## [69] "opAriDf"      "sumIntDou"    "sumIntInt"    "terme01"
## [73] "terme02"      "vecForMat01"  "vecForMat02"  "vecForMat03"
## [77] "vecForMatrix" "vecManip"     "vecManip2"    "vecManip3"
## [81] "zzz"
```

9.3.9 rm()

La fonction `rm()` permet de supprimer un objet présent dans l'environnement de travail de R.

```
rm(zzz)
ls()
```

```
## [1] "aLogic"      "bddInsect"    "characters"   "contrib"
## [5] "dfForMat"    "factor01"     "i"            "irisCopy"
## [9] "j"           "k"            "logicals"     "mdat"
## [13] "miArray"     "miArray02"    "miDf01"       "miDfSub01"
## [17] "miDfSub02"   "miDfSub03"    "miDfSub04"    "miList01"
## [21] "miList02"    "miList0203"   "miList03"     "miList04"
## [25] "miList05"    "miList06"     "miMat"        "miMat01"
## [29] "miMat02"     "miVec01"      "miVec02"      "miVec03"
## [33] "miVec04"     "miVec05"      "miVec06"      "miVec07"
## [37] "miVec08"     "miVec09"      "miVec10"      "miVec11"
## [41] "miVec12"     "miVec13"      "miVec14"      "miVec15"
## [45] "miVec20"     "miVec21"      "miVec22"      "miVec23"
## [49] "miVec24"     "miVec25"      "miVecArr"     "miVecArr02"
## [53] "miVecConf"   "miVecNA"      "miVecOp"      "msg"
## [57] "myCol"       "myRow"        "myText"       "myText2"
## [61] "myText3"     "myText4"      "myText5"      "nbrRep"
## [65] "newVec"      "newVec2"      "numbers"      "oldWd"
## [69] "opAriDf"     "sumIntDou"    "sumIntInt"    "terme01"
## [73] "terme02"     "vecForMat01"  "vecForMat02"  "vecForMat03"
## [77] "vecForMatrix" "vecManip"     "vecManip2"    "vecManip3"
```

9.4 Quelques exercices

Voici quelques exercices pour se perfectionner dans l'usage des fonctions et en apprendre de nouvelles grâce à la documentation. Certains exercices sont difficiles, nous pourrons y revenir plus tard.

9.4.1 Séquences

Reproduisons les séquences suivantes :

-3 -4 -5 -6 -7 -8 -9 -10 -11

-3 -1 1 3 5 7 9 11

3.0 3.2 3.4 3.6 3.8 4.0

20 18 16 14 12 10 8 6

"a" "f" "k" "p" "u" "z"

"a" "a" "a" "a" "a" "f" "f" "f" "f" "f" "k" "k" "k" "k" "k" "k" "p" "p" "p" "p" "p" "p" "u" "u" "u" "u" "u" "u" "z" "z" "z" "z" "z"

Solutions possibles (car il y a toujours plusieurs solutions) :

```
-3:-11
```

```
## [1] -3 -4 -5 -6 -7 -8 -9 -10 -11
```

```
seq(from = -3, to = 11, by = 2)
```

```
## [1] -3 -1 1 3 5 7 9 11
```

```
seq(from = 3.0, to = 4.0, by = 0.2)
```

```
## [1] 3.0 3.2 3.4 3.6 3.8 4.0
```

```
letters[seq(from = 1, to = 26, by = 5)]
```

```
## [1] "a" "f" "k" "p" "u" "z"
```

```
letters[rep(seq(from = 1, to = 26, by = 5), each = 5)]
```

```
## [1] "a" "a" "a" "a" "a" "f" "f" "f" "f" "f" "k" "k" "k" "k" "k" "p" "p"
## [18] "p" "p" "p" "u" "u" "u" "u" "u" "u" "z" "z" "z" "z" "z"
```

9.4.2 Statistiques descriptives

Dans le jeu de données `iris`, combien de valeurs de largeur de sépales sont supérieures à 3 ? Entre 2.8 et 3.2 ?

Comment peut-on visualiser la distribution des données (fonction `table`) ?

Quelles sont les 10 valeurs les plus petites ?

Comment calculer un intervalle contenant 90% des valeurs ?

Si la distribution des données était Normale, quelle serait la valeur théorique de cet intervalle de 90% (fonction `qnorm`) ?

Solutions :

```
length(iris$Sepal.Width[iris$Sepal.Width > 3])
```

```
## [1] 67
```

```
length(iris$Sepal.Width[iris$Sepal.Width > 2.8 &
  iris$Sepal.Width < 3.2])
```

```
## [1] 47
```

```
table(iris$Sepal.Width)
```

```
##
##      2 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9      3 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8
##      1   3   4   3   8   5   9  14  10  26  11  13   6  12   6   4   3   6
## 3.9   4 4.1 4.2 4.4
##      2   1   1   1   1
```

```
table(round(iris$Sepal.Width))
```

```
##
##      2   3   4
##     19 106  25
```

```
irisSepWCopy <- iris$Sepal.Width
irisSepWCopy <- irisSepWCopy[order(irisSepWCopy)]
head(irisSepWCopy, n = 10)
```

```
## [1] 2.0 2.2 2.2 2.2 2.3 2.3 2.3 2.3 2.4 2.4
```

```
quantile(irisSepWCopy, probs = c(0.05, 0.95))
```

```
##      5%   95%
## 2.345 3.800
```

```
qnorm(
  p = c(0.05, 0.95),
  mean = mean(irisSepWCopy),
  sd = sd(irisSepWCopy)
)
```

```
## [1] 2.340397 3.774270
```

9.5 Ecrire une fonction

XXX

9.6 Conclusion

Chapter 10

Importer et exporter des données

10.1 Lire des données depuis un fichier tableur

XXX

10.2 Sauver des données pour R

save load

10.3 Exporter des données

write XXX

Chapter 11

Les boucles

11.1 Pourquoi faire des boucles

XXX un peu d'algo

11.2 La boucle if

XXX

11.3 La boucle switch

XXX

11.4 La boucle for

XXX

11.5 La boucle while

XXX

11.6 repeat, next, break, stop

XXX

11.7 Les boucles de la famille apply

11.7.1 apply

xxx

11.7.2 sapply

xxx

11.7.3 lapply

xxx

11.7.4 tapply

xxx

11.7.5 mapply

xxx

Part II

Les graphiques

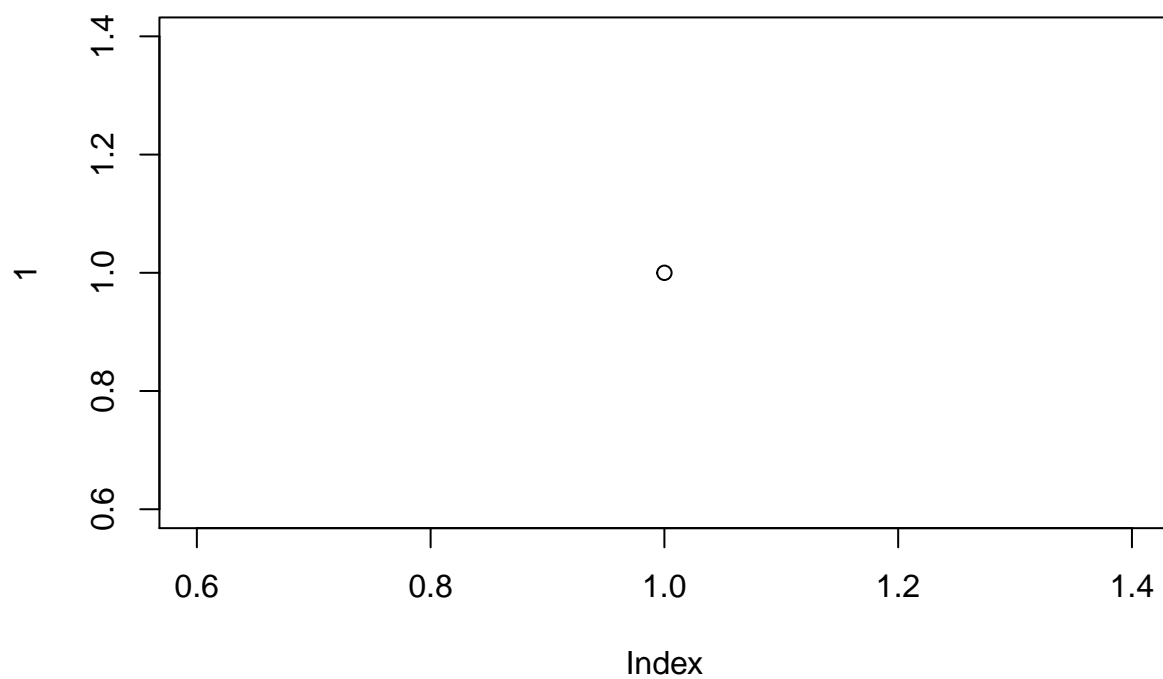
Chapter 12

Graphiques simples

12.1 plot

xxx

```
plot(1)
```



12.2 hist**12.3 barplot****12.4 boxplot****12.5 image et contour**

Chapter 13

La gestion des couleurs

13.1 colors()

13.2 RGB

13.3 Paletas

Chapter 14

Graphiques composés

14.1 mfrow

14.2 layout

Chapter 15

Manipuler les graphiques

15.1 Inkscape

15.2 The Gimp

Part III

Statistiques

Chapter 16

Statistiques descriptives

Part IV

Etude de cas

Chapter 17

Analyser des données de datalogger de température

Chapter 18

Obtenir le numéro WOS d'un article scientifique à partir de son numéro DOI

Il peut être intéressant d'obtenir le numéro WOS d'un article scientifique. Ce numéro est cependant fastidieux à obtenir, d'autant plus si nous souhaitons le récupérer pour une liste d'articles ! Par chance *The Kitchen Research Group* dans leur blog de juin 2015 (<http://kitchingroup.cheme.cmu.edu/blog/2015/06/08/Getting-a-WOS-Accession-number-from-a-DOI/>) propose une méthode pour récupérer le numéro WOS à partir du numéro DOI. C'est cette méthode que nous allons utiliser avec R et le package `httr`. En bref, cette méthode consiste à interroger le site web du WOS à partir du numéro DOI. Le site web du WOS va répondre en spécifiant qu'il faut se connecter pour accéder à l'article. Il y a pour cela une redirection vers une page web dont l'URL contient le numéro WOS. Il suffit alors d'extraire le numéro WOS de l'URL de la page web. **Si nous utilisons un proxy pour accéder au site web du WOS, il faut donc le désactiver pour que la méthode fonctionne.**

Tout d'abord si cela n'est pas déjà fait il faut installer le package `httr` avec `install.packages("httr")`, puis le charger avec `library("httr")`. Une autre solution consiste à utiliser la fonction suivante qui va vérifier si le package est installé puis le charger (il existe de nombreuses déclinaisons de cette fonction sur internet, il s'agit ici d'un mélange de multiples sources).

```
pkgCheck <- function(packages){
  for(x in packages){
    try(if (!require(x, character.only = TRUE)){
      install.packages(x, dependencies = TRUE)
      if(!require(x, character.only = TRUE)) {
        stop()
      }
    })
  }
}
pkgCheck("httr")
```

La liste des numéros DOI est la suivante (contenu dans un vector) :

```
myDOIs <- c("10.1111/2041-210X.12935", "10.1007/s13355-017-0480-5")
print(myDOIs)
```

```
## [1] "10.1111/2041-210X.12935" "10.1007/s13355-017-0480-5"
```

Pour chaque DOI, nous allons interroger le site web du WOS, récupérer l'URL de redirection, puis récupérer le numéro WOS. Nous allons donc faire une boucle sur notre vector contenant les DOI. Nous utilisons la fonction `seq_along()` qui va prendre comme valeurs les éléments d'une séquence de 1 à la taille de l'objet `myDOIs`, soit : `i = 1`, puis `i = 2`.

```
for(i in seq_along(myDOIs)){
  # ...
}
```

Dans la boucle, le DOI que nous allons traiter est donc `myDOIs[i]` que nous allons appeler `myDOI`. La page d'interrogation du WOS correspond à la concaténation de l'URL du WOS avec le numéro WOS (l'URL est présentée sur plusieurs lignes pour respecter la largeur de page de ce livre).

```
for(i in seq_along(myDOIs)){
  myDOI <- myDOIs[i]
  myWebPage <- paste0(
    "http://ws.isiknowledge.com/",
    "cps/openurl/service?url_ver=Z39.88-2004",
    "&rft_id=info:doi/", myDOI)
}
```

Maintenant nous allons utiliser la fonction `GET()` du package `httr` pour récupérer l'URL.

```
for(i in seq_along(myDOIs)){
  myDOI <- myDOIs[i]
  myWebPage <- paste0(
    "http://ws.isiknowledge.com/",
    "cps/openurl/service?url_ver=Z39.88-2004",
    "&rft_id=info:doi/", myDOI)
  r <- GET(myWebPage)
  urlWOS <- r[[1]]
}
```

Il se peut que pour un article, il n'y ait pas de numéro WOS correspondant. Pour que notre script ne soit pas arrêté en cas d'erreur il faut donc **gérer cette exception**. Nous allons créer un objet `tryExtract` qui va prendre comme valeur une chaîne de caractères vide `""`. Ensuite nous allons essayer avec la fonction `try()` d'extraire le numéro WOS avec la fonction `substr()`. Le numéro WOS se situe depuis le caractère numéro 117 jusqu'au caractère numéro 131 de l'URL.

```
for(i in seq_along(myDOIs)){
  myDOI <- myDOIs[i]
  myWebPage <- paste0(
    "http://ws.isiknowledge.com/",
    "cps/openurl/service?url_ver=Z39.88-2004",
    "&rft_id=info:doi/", myDOI)
  r <- GET(myWebPage)
  urlWOS <- r[[1]]
  tryExtract <- ""
  try(tryExtract <- substr(x = urlWOS, start = 117, stop = 131), silent = TRUE)
}
```

Nous pouvons ensuite vérifier que l'extraction correspond bien à un numéro en utilisant une **expression régulière**. Ici nous allons simplement vérifier que l'extraction ne contient que des chiffres. Dans le cas contraire `tryExtract` reprendra sa valeur initiale `""`.

```
for(i in seq_along(myDOIs)){
  myDOI <- myDOIs[i]
  myWebPage <- paste0(
```

```

      "http://ws.isiknowledge.com/",
      "cps/openurl/service?url_ver=Z39.88-2004",
      "&rft_id=info:doi/", myDOI)
r <- GET(myWebPage)
urlWOS <- r[[1]]
tryExtract <- ""
try(tryExtract <- substr(x = urlWOS, start = 117, stop = 131), silent = TRUE)
if(!grepl(pattern = '^([0-9]*)$', x = tryExtract)){tryExtract <- ""}
}

```

Le résultat est ensuite stocké dans un vector créé au préalable et appelé vecWOS.

```

vecWOS <- vector()
for(i in seq_along(myDOIs)){
  myDOI <- myDOIs[i]
  myWebPage <- paste0(
    "http://ws.isiknowledge.com/",
    "cps/openurl/service?url_ver=Z39.88-2004",
    "&rft_id=info:doi/", myDOI)
  r <- GET(myWebPage)
  urlWOS <- r[[1]]
  tryExtract <- ""
  try(tryExtract <- substr(x = urlWOS, start = 117, stop = 131), silent = TRUE)
  if(!grepl(pattern = '^([0-9]*)$', x = tryExtract)){tryExtract <- ""}
  vecWOS <- append(vecWOS, tryExtract)
}

```

Nous pouvons alors créer un objet de type `data.frame` qui va contenir les numéros DOI et les numéros WOS, et éventuellement l'exporter dans un fichier CSV.

```

dfDOIWOS <- data.frame(DOI = myDOIs, WOS = vecWOS)
write.csv(dfDOIWOS, file = "dfDOIWOS.csv", row.names = FALSE)

```

Le résultat est le suivant (non exécuté car la procédure d'interrogation avec la fonction `GET()` est très lente : **si nous souhaitons travailler sur une liste de plusieurs dizaines ou centaines d'articles, plusieurs heures seront nécessaires** avant d'obtenir le résultat).

```

#                DOI                WOS
# 1  10.1111/2041-210X.12935 000429421800031
# 2  10.1007/s13355-017-0480-5 000400381400016

```

Voici le code complet :

```

pkgCheck <- function(packages){
  for(x in packages){
    try(if (!require(x, character.only = TRUE)){
      install.packages(x, dependencies = TRUE)
      if(!require(x, character.only = TRUE)) {
        stop()
      }
    })
  }
}

```

```

}
pkgCheck("httr")

myDOIs <- c("10.1111/2041-210X.12935", "10.1007/s13355-017-0480-5")

vecWOS <- vector()
for(i in seq_along(myDOIs)){
  myDOI <- myDOIs[i]
  myWebPage <- paste0(
    "http://ws.isiknowledge.com/",
    "cps/openurl/service?url_ver=Z39.88-2004",
    "&rft_id=info:doi/", myDOI)
  r <- GET(myWebPage)
  urlWOS <- r[[1]]
  tryExtract <- ""
  try(tryExtract <- substr(x = urlWOS, start = 117, stop = 131), silent = TRUE)
  if(!grepl(pattern = '^[0-9]*$', x = tryExtract)){tryExtract <- ""}
  vecWOS <- append(vecWOS, tryExtract)
}

dfDOIWOS <- data.frame(DOI = myDOIs, WOS = vecWOS)
write.csv(dfDOIWOS, file = "dfDOIWOS.csv", row.names = FALSE)

```

La boucle `for()` pourrait être remplacée par une boucle `sapply()`, ou pour gagner en temps d'exécution par une parallélisation. Voici un exemple de temps d'exécution renvoyé par la fonction `system.time()` :

```

#   user  system elapsed
# 0.10    0.01   35.00

```

Nous venons de faire un script qui permet à partir d'une liste de numéros DOI d'obtenir automatiquement les numéros WOS.