

Frédéric Bertrand
Emmanuelle Claeys
Myriam Maumy-Bertrand

Modélisation statistique par la pratique avec R

DUNOD

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements

d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour

les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, 2019

11 rue Paul Bert, 92240 Malakoff

www.dunod.com

ISBN 978-2-10-079352-5

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^o et 3^o a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Nous souhaitons remercier Aline, Anaëlle, David, Guillaume, Guy et Marie pour leur soutien, leurs encouragements, leur patience et la douceur dont ils faisaient preuve chaque jour de ce bel été 2019. Sans tous ces ingrédients, le livre n'aura jamais vu le jour !

AVANT-PROPOS

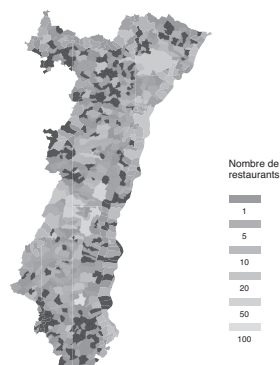
Chères étudiantes, chers étudiants,

Le livre que vous tenez entre vos mains est un subtil mélange entre analyse statistique et pratique de \mathbb{R} . Dans cet ouvrage, nous présentons des méthodes statistiques qui nécessitent déjà une bonne maîtrise des notions essentielles de la statistique. Nous rappelons que les auteurs ont déjà écrit un livre sur ce sujet : *Initiation à la statistique avec \mathbb{R}* , 2018, Dunod. Nous pourrions donc dire que ce livre est la suite logique du livre précédemment cité. Pourquoi une suite ? Parce que les étudiant(e)s, les élèves d'école d'ingénieurs, les doctorant(e)s auxquels nous nous adressons quotidiennement nous l'ont demandé ! Car les notions fondamentales de statistique ne suffisent plus à répondre aux questions qui sont désormais posées ou induites par les bases de données que vous avez à traiter.

Ce livre est donc le concentré de cours et de travaux dirigés que nous avons dispensés, animés et proposés aux apprentis-statisticiens. Il ne se veut pas être un livre de théorie de la statistique mais tout simplement un recueil d'aides, de conseils, de méthodes pour pouvoir explorer, analyser et modéliser la masse de données qui ne fait que croître quotidiennement.

À l'heure du *Big Data* et de la *Data Science*, il est inconcevable, quel que soit le parcours professionnel que vous suivez et quelle que soit l'université ou l'école d'ingénieurs qui vous dispense les cours, de ne pas savoir ce qu'est un coefficient de corrélation linéaire, de ne pas savoir appliquer une analyse en composantes principales ou une classification ascendante hiérarchique ou tout simplement de ne pas savoir représenter graphiquement la répartition des restaurants de votre région en utilisant les données *opendata*.

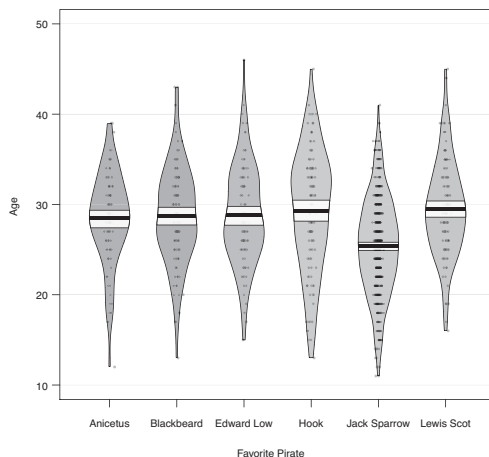
Répartition des restaurants en Alsace
Nombre de restaurants par district



Data: INSEE | Creation: d'après Yan Holtz | r-graph-gallery.com

Avant-propos

Il est également inenvisageable pour un praticien de la statistique de ne pas savoir ce qu'est un test de permutation, un intervalle de confiance bootstrap. Certes, la théorie mathématique de ces notions a déjà été établie il y a quelques années mais savoir bien s'en servir et dans les bonnes conditions reste l'enjeu majeur de tout étudiant ! C'est avec ce constat que nous avons écrit ce livre, vous montrez dans quel cas pratique il faut avoir recours à ces solutions relativement techniques mais faciles à mettre en oeuvre à condition d'être bien guidés.



Nous commencerons donc la modélisation en étudiant les mesures de liaison (chapitre 2). Certes ce chapitre est dense car il nous permet de vous présenter les commandes de \mathbb{R} qui vous permettront de réaliser ces alternatives (tests de permutation approché et exact, techniques bootstrap). Nous continuerons ensuite avec l'analyse exploratoire des données (chapitre 3) car rappelons-le encore une fois, un bon graphique vaut mieux qu'un long discours. Par exemple, si vous avez été conquis par les boîtes à moustaches, vous allez très vite adopter les *pirates plots*. Ces *pirates plot* sont un exemple de graphiques complets, appelés *RDI* pour *Raw data, Descriptive and Inferential statistics*. Le dernier chapitre présente les modèles qui tournent autour de la régression.

Pour chacun des chapitres, nous avons essayé de trouver des jeux de données de différents domaines. Nous avons écrit le livre avec la même idée : présenter rapidement les notions de statistique et les mettre en pratique avec \mathbb{R} . Le livre dispose d'un package compagnon nommé *ModStatR* en téléchargement libre depuis le CRAN, <https://cran.r-project.org/package=ModStatR>, ou Github, <https://github.com/fbertran/ModStatR>.

Voilà, nous espérons que cet ouvrage vous plaira et que vous le lirez avec le même plaisir que nous avons eu à l'écrire. Enfin, nous vous signalons, que les corrections des exercices sont disponibles dans les compléments en ligne sur le site dunod.com, à la page de présentation de l'ouvrage.

Bonne lecture !

Les auteurs.

Toutes vos remarques, vos commentaires, vos critiques, et même vos encouragements, seront accueillis avec plaisir.

frederic.bertrand1@utt.fr
emmanuelle.claeys@unistra.fr
mmaumy@math.unistra.fr

TABLE DES MATIÈRES

AVANT-PROPOS	iii
COMMENT UTILISER CE LIVRE ?	vi
CHAPITRE 1 • GÉNÉRALITÉS SUR LE LANGAGE R	1
1 Présentation du langage R	1
2 Installation du langage R	4
3 Travailler avec R	8
4 Écrire et compiler des scripts sous R	14
5 R sans les mains	16
CHAPITRE 2 • MESURES DE LIAISON	20
1 Coefficient de corrélation linéaire	21
2 Coefficient de corrélation multiple	79
3 Coefficient de corrélation partielle	92
4 Coefficient de corrélation de Spearman	107
5 Coefficient de corrélation de Kendall	118
CHAPITRE 3 • ANALYSE EXPLORATOIRE DES DONNÉES	129
1 Analyse en composantes principales	129
2 Analyse factorielle des correspondances	167
3 Analyse non symétrique des correspondances	182
4 Analyse des correspondances multiples	191
5 Analyse factorielle des données mixtes	197
6 Classification ascendante hiérarchique et méthode des K-moyennes	199
CHAPITRE 4 • ANALYSE DE RÉGRESSION	216
1 Les multifacettes de la régression	216
2 Modèle de régression pour réponse quantitative	217
3 Estimation et diagnostics	222
4 Tests d'hypothèses	233
5 Intervalles et régions de confiance	236
6 Choix de modèle par sélection de variables	237
7 Régression pénalisée	245
8 Utilisation du logiciel R	246
9 Valeurs manquantes	268
BIBLIOGRAPHIE	291
INDEX	293

Comment utiliser ce livre ?

Les rubriques



Des remarques pour aller plus loin



Les pièges et difficultés sont signalés



Des conseils méthodologiques sont donnés
tout au long de l'apprentissage





Voir compléments en ligne

GÉNÉRALITÉS SUR LE LANGAGE



1


« Les détails, comme chacun le sait, conduisent à la vertu et au bonheur ; les généralités sont, au point de vue intellectuel, des maux inévitables. » De Aldous Huxley, *Le Meilleur des mondes*, 1932.

INTRODUCTION

Ce chapitre contient la présentation du langage , l'installation de  et des packages (modules en français) sous les trois principaux systèmes d'exploitation. Il présente également les lignes de commande incontournables qu'il faut savoir maîtriser.







OBJECTIFS

- Présenter .
- Installer .
- Prise en main des premières lignes de commande incontournables.

Les informations sur  sont disponibles sur le site internet dédié au projet :
<https://www.r-project.org/>

1 PRÉSENTATION DU LANGAGE

1.1 Qu'est-ce-que le langage ?

- Le langage  est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.
-  un langage de programmation complet. Cette caractéristique rend par conséquent le langage  différent des autres logiciels de statistique.
-  est disponible pour Microsoft Windows, Macintosh et de nombreux systèmes de type Unix.
- Actuellement, quatre à cinq nouvelles versions de  apparaissent par an.
-  est distribué gratuitement sous les termes de la « GNU », *General Public Licence Version 2*, Juin 1991.

- R est écrit en C, C++, FORTRAN et Java. De plus, R est plus orienté programmation objet que la plupart des autres logiciels ou langage de programmation statistique.
- R est un clone gratuit du langage S-Plus, actuellement commercialisé par Tibco Software Inc., créé autour du langage S qui a été développé par John Chambers des laboratoires Bell.
- R a été créé en 1993 par Ross Ihaka et Robert Gentleman à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Development Core Team. L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Pour tenter de conclure, Ross Ihaka déclare, dans l'article qui est consacré à R dans le *New-York Times* en Janvier 2009, : *“R is a real demonstration of the power of collaboration, and I don't think you could construct something like this any other way. We could have chosen to be commercial, and we would have sold five copies of the software.”*¹

1.2 Comment se procurer le langage R ?

- L'adresse <https://www.r-project.org/> est le premier résultat pour la recherche de la lettre R avec le moteur de recherche Google™ et la meilleure source d'informations sur le langage R. Vous y trouverez les différentes distributions du langage, de nombreuses bibliothèques de fonctions et des documents d'aide.
- Le langage R est gratuit et se télécharge directement depuis internet. Il évolue très rapidement et à peu près tous les six mois une nouvelle version du langage est proposée au public. Elle est accessible via la page officielle consacrée au projet <https://www.r-project.org/>.
- Pour faire face au très grand nombre de téléchargements du langage, un système de miroirs, le Comprehensive R Archive Network <https://cran.r-project.org/>, a été mis en place. Les mêmes fichiers sont ainsi disponibles simultanément sur différents serveurs situés à plusieurs endroits dans le monde². Cette organisation présente au moins deux avantages majeurs pour vous : pouvoir choisir un miroir proche de chez vous où que vous soyez dans le monde ou un miroir de secours lorsque le miroir que vous avez l'habitude d'utiliser est indisponible.

En juin 2019, il existait six miroirs en France dont les adresses sont les suivantes :

1. Laboratoire de Biométrie et Biologie Évolutive (LBBE), UMR CNRS 5558, Lyon
<https://pbil.univ-lyon1.fr/CRAN/>




1. Vous pouvez consulter l'article ici : <https://archive.nytimes.com/www.nytimes.com/2009/01/07/technology/business-computing/07program.html>.


2. La liste des serveurs est disponible à l'adresse <https://cran.r-project.org/mirrors.html> et leur état à l'adresse https://cran.r-project.org/mirmon_report.html





2. Institut de Biologie et Chimie des Protéines, CNRS et Université de Lyon 1
<https://mirror.ibcp.fr/pub/CRAN/>
<http://mirror.ibcp.fr/pub/CRAN/>
3. Institut de Biologie du Développement de Marseille (IBDM), CNRS et Université Aix-Marseille
<https://cran.biotoools.fr/>
<http://cran.biotoools.fr/>
4. Institut de Génétique Humaine (IGH), Montpellier
<https://ftp.igh.cnrs.fr/pub/CRAN/>
<http://ftp.igh.cnrs.fr/pub/CRAN/>
5. Institut de Radioprotection et de Sûreté Nucléaire (IRSN), Paris
<http://cran.irsn.fr/>
6. SAMM, Université Paris 1 Panthéon-Sorbonne
<https://cran.univ-paris1.fr/>
<http://cran.univ-paris1.fr/>


Certains miroirs proposent un accès sécurisé en `https`. Il est recommandé de le privilégier.

1.3 Remarques sur le langage avant l'installation

-  fonctionne avec une grande variété de systèmes d'exploitation et en particulier avec Microsoft Windows, macOS X et de nombreux systèmes de type Unix.  est soit disponible sous la forme de fichiers prêts à être installés soit sous la forme de fichiers sources à compiler soi-même.
- Il existe une version française du langage  même si le site officiel est rédigé en langue anglaise.
- La très grande majorité des fonctions du langage ne diffère pas d'un système d'exploitation à l'autre bien que les interfaces graphiques ne sont pas similaires.


Il existe plusieurs interfaces graphiques, en anglais « GUI » pour *Graphical User Interface*, qui permettent d'accéder à une partie des fonctions du langage .


- RGUI, l'interface graphique installée par défaut sous Windows.
- JGR, une interface graphique programmée en Java pour . Elle fonctionne pour tous les systèmes d'exploitation sur lesquels le langage Java est disponible et donc aussi bien pour Microsoft Windows que pour MacOS X et de nombreux systèmes de type Unix.
- Rattle, une interface graphique pour le data mining utilisant .
- R Commander, une interface graphique pour faire des statistiques usuelles avec .
- Statistical Lab.
- RExcel, pour exécuter les fonctionnalités de  et de R Commander à partir de Microsoft Excel.
- rggobi, une interface pour le logiciel GGobi spécialisé dans la visualisation de données multidimensionnelles.
- RKWard, une interface graphique basée sur les bibliothèques de KDE, une interface graphique disponible sur de nombreux systèmes de type Unix.












Il existe également des programmes qui facilitent l'écriture des instructions et des programmes en langage .

Ces programmes se regroupent en deux grandes catégories :

- Les éditeurs de texte comme le bloc-note de Microsoft Windows, Microsoft Wordpad ou Microsoft Word.
- Les environnements de programmation, appelés en anglais *Integrated Development Environments* (IDEs), comme Tinn-R, Emacs (*Emacs Speaks Statistics*), Jedit, Kate, WinEdt (R Package RWinEdt) ou Vim.

Pour les utilisateurs qui en auraient besoin, par exemple pour rédiger un mémoire ou un rapport de stage, il est également possible d'intégrer directement des résultats obtenus avec  dans des documents L^AT_EX à l'aide du langage Sweave ou dans des documents au format OpenDocument (ODF) à l'aide du langage odfWeave.

Il existe plusieurs sites internet, principalement en langue anglaise, consacrés au langage .

- La section sur  à l'Open Directory Project.
- RSeek et  site search qui sont des moteurs de recherche spécialisés dans la recherche de documents consacrés à .
- Plusieurs listes de diffusion permettent aux utilisateurs novices ou expérimentés de  de poser directement leurs questions aux autres utilisateurs afin qu'ils leur fassent partager leur expérience du langage.
- Le  Journal est un journal à comité de lecture comportant des articles consacrés aux problèmes de calcul statistique et au développement du langage . Il peut aussi bien intéresser de simples utilisateurs du langage que des programmeurs.
-  books comprend une liste importante de livres consacrés à .
- Le R Graphical Manual explique par l'exemple comment construire des graphiques à partir de n'importe quelle bibliothèque du langage. Il comporte également un index exhaustif des fonctions de toutes les bibliothèques d'extensions du langage .
- Le  wiki est un site coopératif de documentation sur le langage .

2 INSTALLATION DU LANGAGE

2.1 Les premières instructions

Ces premières instructions sont communes aux trois systèmes d'exploitation.


1. Rendez-vous sur le site <https://www.r-project.org/>.
2. Puis, à gauche sur la page d'accueil, vous trouverez un menu Download. Dans ce menu, cliquez sur CRAN.
3. Choisissez un site miroir proche de chez vous.

4. Un encadré blanc intitulé Download and Install R doit apparaître sur votre écran.


Installation de la version 3.6.1 (2019-07-05) de sous Windows

1. Cliquez sur Download R for Windows puis sur base.
2. Un encadré grisé doit apparaître dans lequel, à la première ligne, est inscrit Download R 3.6.1 for Windows (81 megabytes, 32/64 bit), la version 3.6.1 étant celle disponible au mois de septembre 2019. Cliquez dessus.
3. Procédez au téléchargement.
4. Exécutez le fichier que vous venez de télécharger en choisissant une installation par défaut.



Le téléchargement de  n'est pas très long. En effet, sa taille est de 80,0 Mo.


Installation de la version 3.6.1 (2019-07-05) de sous macOS X

1. Cliquez sur Download R for (Mac) OS X.
2. Une liste de fichiers à télécharger, intitulée Files :, apparaît. Le premier élément de la liste est R-3.6.1.pkg, la version 3.6.1 étant celle disponible au mois de septembre 2019. Cliquez dessus. Attention, il vous faudra peut-être, en fonction de votre navigateur internet, appuyez sur la touche control en cliquant pour pouvoir télécharger le fichier.
3. Pour installer , double-cliquez sur l'icône du package d'installation R-3.6.1.pkg.


Installation de la version 3.6.1 (2019-07-05) de sous Linux

1. Cliquez sur Download R for Linux puis sur le nom de la distribution Linux installée sur votre ordinateur.
2. Suivez les instructions détaillées sur le site. Celles-ci varient trop d'une distribution à l'autre pour être reproduites ici mais si vous utilisez Linux vous ne devriez pas avoir de mal à installer ou à faire installer le langage par votre administrateur réseau.



Dans la grande majorité des situations, l'installation de  est très simple et ne nécessite que peu de connaissances techniques.

2.2 sous les trois principaux systèmes d'exploitation

L'interface graphique du langage  est très similaire d'un système d'exploitation à l'autre. Voici, en détails, la procédure à suivre dans le cas de l'environnement Windows.

Environnement Windows

- R fonctionne avec plusieurs fenêtres sous Windows. La fenêtre R Console est la fenêtre principale où sont réalisées par défaut les entrées de commandes et les sorties de résultats en mode texte. À celle-ci peuvent s'ajouter des fenêtres facultatives, telles que les fenêtres graphiques, les fenêtres d'informations (historique des commandes, aide, visualisation de fichier, etc), toutes appelées par des commandes spécifiques via la fenêtre R Console.
- Le menu File ou Fichier contient les outils nécessaires à la gestion de l'espace de travail, tels que la sélection du répertoire par défaut, le chargement de fichiers sources externes, la sauvegarde et le chargement d'historiques des commandes exécutées, etc.
- Le menu Edit ou Edition contient les habituelles commandes de copier-coller, ainsi que la boîte de dialogue autorisant la personnalisation de l'apparence de l'interface.
- Le menu View ou Voir permet d'afficher ou de masquer la barre d'outils et la barre de statut.
- Le menu Misc traite de la gestion des objets en mémoire et permet d'arrêter un calcul ou des calculs en cours de traitement.
- Le menu Packages automatise la gestion et le suivi des bibliothèques de fonctions, permettant leur installation et leur mise à jour de manière transparente depuis l'un des miroirs du CRAN (*Comprehensive R Archive Network*) <https://cran.r-project.org/>.
- Enfin, le menu Windows ou Fenêtres et le menu Help ou Aide assument des fonctions similaires à celles qu'ils occupent dans les autres applications Windows, à savoir la définition spatiale des fenêtres et l'accès à l'aide en ligne et aux manuels de références du langage R.

2.3 Installer des packages du langage R

Qu'est ce qu'un package ? Un package est une compilation d'outils. Certains sont déjà présents dans l'installation de base de R. En effet, lors de l'installation de R, un dossier `library` s'est créé par défaut. Il comprend les packages de base de R. Mais d'autres packages qui vous seront utiles pour réaliser vos analyses statistiques seront à télécharger puis à installer.

Pour les trois environnements

1. Reprenez la procédure de téléchargement de R vue à la section « Installer le langage R ».
2. Cette fois-ci au lieu de cliquer sur Windows, Mac OS X ou Linux, cliquez sur packages dans la liste intitulée Source Code for all Platforms.
3. Une page apparaît sur laquelle est indiqué le nombre de packages actuellement disponibles sur le CRAN (14 924 en septembre 2019). Pour obtenir la liste des packages rangés dans l'ordre alphabétique, cliquez sur Table of available packages, sorted by name.
4. La liste des packages apparaît alors.
5. Puis cliquez sur le package dont vous avez besoin.

6. Une brève description du package apparaît suivie d'une liste proposant plusieurs versions du package. La première (.tar.gz) est celle contenant le code source du package et ne sert a priori qu'aux utilisateurs de Linux. La deuxième (.tgz) est destinée aux utilisateurs de macOS X et la troisième (.zip) aux utilisateurs de Windows.
7. Sous macOS X et Windows, il faut alors démarrer l'interface graphique de R.
8. Sous macOS X, allez dans le menu Packages et utilisez le Package Manager. Indiquez alors à R le fichier .tgz que vous venez de télécharger.
9. Sous Windows, allez dans le menu Packages et choisissez Installer depuis un fichier .zip. Indiquez alors à R le fichier .zip que vous venez de télécharger.

Environnement Windows

Il existe une procédure alternative pour les utilisateurs d'un environnement Windows.

1. Reprenez la procédure de téléchargement de R vue à la section « Installer le logiciel R ».
2. Cette fois-ci au lieu de cliquer sur base, cliquez sur contrib.
3. Cliquez ensuite sur le dossier de la version R que vous avez installée.
4. Puis cliquez sur le package dont vous avez besoin.
5. Un fichier « .zip » est enregistré sur votre disque dur.
6. Démarrez l'interface graphique de R, allez dans le menu Packages et choisissez Installer depuis un fichier .zip. Indiquez alors à R le fichier .zip que vous venez de télécharger.



1. Lorsque vous aurez besoin de packages qui ne sont pas installés par défaut pour réaliser les analyses statistiques qui vous seront demandées dans les exercices, cela vous sera signalé. Il faudra alors installer ces packages supplémentaires sur votre ordinateur.
2. Beaucoup de bibliothèques contiennent un ou plusieurs document(s) détaillant leurs fonctionnalités et montrant leur application pas à pas à un exemple. Pour obtenir la liste des vignettes présentes sur votre ordinateur, il suffit d'exécuter la fonction vignette().




2.4 Quelques remarques sur la fenêtre R Console

- Ce qui est entré par l'utilisateur n'est pas de la même couleur que la réponse de R.
- R utilise le système anglo-saxon pour les nombres décimaux c'est-à-dire les décimales sont séparées par un point et non par une virgule comme en France.
- R distingue les majuscules des minuscules.
- Vous devez faire attention à l'utilisation du point virgule. En effet, sous R, ce dernier sert à séparer deux instructions.

- Vous pouvez rappeler les commandes déjà exécutées en utilisant la touche « Flèche vers le haut ».
- Vous pouvez parcourir la ligne de commande que vous êtes en train d'écrire en appuyant sur les touches « Flèche vers la gauche » et « Flèche vers la droite ».

3 TRAVAILLER AVEC

3.1 Démarrer

Pour démarrer , vous pouvez par exemple lancer le logiciel  en double-cliquant sur l'icône  qui se trouve par exemple sur votre bureau. La fenêtre R console s'ouvre. Elle vous affiche tout ce texte :


```
R version 3.6.1 (2019-07-05) -- "Actions of the Toes"
Copyright (C) 2019
The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.


Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou ' help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

[R.app GUI 1.70 (7684) x86_64-apple-darwin15.6.0]
```


Puis sous ce texte, le symbole `>`, appelé **prompt**, apparaît également. Il signifie que  est prêt à travailler.



Il ne faut jamais taper ce symbole au clavier car il est déjà présent en début de ligne sur la fenêtre R Console.

C'est à la suite de `>` que vous taperez les lignes de commande de . Une fois la commande tapée, vous devez toujours la valider en appuyant sur la touche « Entrée ».



Si votre commande est incomplète, le symbole `>` est remplacé par `+`. Ce symbole `+` signifie que  attend la suite de la commande. Si vous ne savez pas compléter la ligne de commande ou qu'elle présente une erreur vous pouvez appuyer sur la touche « Echap » pour annuler la commande et créer un nouveau symbole `>`.

3.2 Quitter R

Pour quitter R, vous utilisez la commande suivante :

```
> q()
Save workspace image? [y/n/c]
```



Sous l'environnement Windows ou macOS X, c'est une boîte de dialogue en français qui apparaîtra à l'écran et qui comportera la même question.

R vous propose de sauvegarder le travail effectué. Trois réponses vous sont proposées : y (pour *yes*), n (pour *no*) ou c (pour *cancel*, annuler).

Si vous tapez y, cela permet que les commandes exécutées pendant la session et les objets enregistrés en mémoire soient conservés et soient donc « rappelables » et « réutilisables ».

Si vous tapez n, vous quittez R qui oubliera tout le travail que vous avez réalisé. Attention, vous risquez de tout perdre !

Si vous tapez c, la procédure de fin de session sous R est annulée.

3.3 Sauvegarder sous R

Si vous quittez R en choisissant la sauvegarde de l'espace de travail, deux fichiers sont créés :

1. le fichier `.Rdata` contient des informations sur les variables utilisées,
2. le fichier `.Rhistory` contient l'ensemble des commandes utilisées.

3.4 Consulter l'aide de R

Il y a quatre sources principales d'aide :

1. les fichiers d'aide,
2. les manuels,
3. les archives R-help,
4. et enfin R-help lui-même.

Pour une fonction, dont le nom est `fonction`, vous pouvez consulter une fiche de documentation en tapant `?fonction` ou `help("fonction")`. Grâce à cette aide, il suffit que vous reteniez le nom de la fonction, mais pas forcément toute la syntaxe.

Exemple

Vous cherchez à obtenir des informations sur la fonction `read.table`, vous tapez alors la commande suivante :

```
> ?read.table
```

ou encore

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

Pour une bibliothèque d'extension, dont le nom est `package`, vous pouvez consulter une fiche de documentation en tapant :

```
> help(package="package")
```

Grâce à cette aide, il suffit que vous reteniez le nom de la bibliothèque où se trouve la fonction que vous souhaitez utiliser, mais pas forcément le nom exact de cette fonction.

Les pages d'aide sont généralement très détaillées. Elles contiennent souvent, entre autres :

- une section *See Also* qui donne les pages d'aide sur des sujets apparentés.
- une section *Description* qui précise ce que fait la fonction.
- une section *Examples* avec des lignes de commande illustrant ce que fait la fonction documentée. Ces exemples peuvent être exécutés directement en utilisant la fonction `example`.

Exemple


Tapez la commande suivante :

```
> example(plot)
```

Pour afficher successivement les différents graphiques ainsi créés, vous devez cliquer plusieurs fois de suite sur la fenêtre où sont situés les graphiques.

3.4.1 Affichage de l'aide dans la console

Lorsque l'aide s'affiche dans la console, vous pouvez faire défiler le texte ligne par ligne avec la touche « Entrée » ou « Flèche vers la bas » ou page par page en appuyant sur la barre « Espace ». Une fois arrivé à « END », tapez `q`.

Il s'agit du mode d'affichage par défaut de l'aide dans un terminal  et donc dans un environnement Linux.

3.4.2 Affichage de l'aide en-dehors de la console

Lorsque l'aide s'affiche dans la console, elle n'est pas facile à consulter. Il existe des versions au format `.html` de tous les fichiers d'aide dans les trois environnements Linux, MacOS X et Windows. Pour s'en servir à la place des versions texte qui s'affichent dans la console, il faut utiliser l'option `help_type="html"` de la fonction `help`.

Exemple

Vous cherchez à obtenir des informations au format `.html` sur la fonction `read.table`, vous tapez alors la commande suivante :


```
> help("read.table", help_type="html")
```

L'option `help_type="text"` de la fonction `help` permet d'afficher l'aide au format texte `.txt`.

Exemple

Vous cherchez à obtenir des informations au format `.txt` sur la fonction `read.table`, vous tapez alors la commande suivante :




```
> help("read.table", help_type="text")
```

Le format `.html` est celui utilisé par défaut lorsque vous utilisez l'interface graphique de  dans un environnement Mac OS X ou Windows.


La fonction `help.start` permet d'accéder à la page d'accueil de l'aide au format `.html`.

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



3.4.3 Changer le mode d'affichage par défaut de l'aide

Il est possible de modifier une des options de  qui gère le mode d'affichage par défaut des fichiers d'aide de . Pour afficher, pour toute la durée d'une session , les fichiers au format :

- `.html`, tapez l'instruction `options(help_type = "html")` dans la console.
- `.txt`, tapez l'instruction `options(help_type = "text")` dans la console.

À la fin de l'aide, il y a presque toujours une ou plusieurs lignes de commande d'exemple. Il est judicieux de les exécuter afin d'avoir une idée de ce que les fonctions ont besoin comme informations et des résultats qu'elles vous renvoient. Pour cela, il suffit de copier ces lignes, de les coller sur la fenêtre `R console` et d'observer ce qui se passe. Vous pouvez, bien sûr, les modifier suivant vos besoins. De plus, ces lignes de commande sont souvent un exemple d'analyse de données et pointent parfois sur d'autres fonctions utiles à essayer. La notion de « fonctions » sous  sera développée par la suite.

3.5 Affecter

En fait  est une grosse calculatrice, mais vous n'avez pas installé le logiciel  pour cette fonctionnalité-là. En effet, vous aimeriez parfois réutiliser le résultat d'une opération arithmétique sans avoir à le ressaisir ou à le copier/coller. Pour cela, vous



affecterez des valeurs à des objets et utiliserez l'opérateur `<-`, appelé **opérateur d'affectation** ou d'assignation. Cet opérateur prend une valeur quelconque à droite et la place dans l'objet indiqué à gauche.



Exemple

```
> n<-28
> N<-20
```

Que signifient ces deux lignes de commande ?

La première ligne de commande peut donc se lire ainsi « mettre la valeur 28 dans l'objet nommé `n` ». La deuxième ligne de commande peut donc se lire ainsi « mettre la valeur 20 dans l'objet nommé `N` ».



 ne vous affiche rien dans la console. Pour que  vous renvoie le résultat de cette affectation, il faut lui demander de l'imprimer. Vous verrez comment y parvenir au paragraphe suivant.



Le signe `=` convient également pour faire des affectations.

Exemple

```
> m=1973
> m
```

```
[1] 1973
```

3.6 Afficher

Quand vous affectez un nom à un objet, l'affichage de celui-ci n'est pas automatique. Il faut que vous le demandiez en tapant uniquement le nom donné à l'objet.

Exemple

```
> n
```

```
[1] 28
```

Vous pouvez aussi utiliser ces objets dans des calculs.

Exemple



```
> N+n
```

```
[1] 48
```

Vous apprendrez dans la suite comment affecter et visualiser à l'aide de la même ligne de commande.

Vous pouvez utiliser autant d'objets que vous souhaitez. Les objets peuvent contenir non seulement des nombres comme vous venez de le voir mais aussi des chaînes de caractères, qui sont alors indiquées par des guillemets droits, et d'autres choses encore.

3.7 Supprimer

Par défaut  conserve en mémoire tous les objets créés lors de la session. Il n'y a que dans le cas où vous quittez , sans sauvegarder la session, que les objets que vous avez créés sont supprimés.



Il est donc conseillé de supprimer régulièrement les objets que vous avez créés lors de votre session. Attention, vous devez être sûr de ne plus en avoir besoin car ils seront définitivement perdus.

À retenir

Pour savoir quels sont les objets qui ont été créés pendant votre session, utilisez, à votre convenance, la fonction `ls` ou la fonction `objects`.

Si vous souhaitez supprimer

- un objet, par exemple l'objet `m`, utilisez `remove`, abrégée en `rm`.

Exemple

```
> rm(m)
```

- plusieurs objets, par exemple les objets `n` et `N`, utilisez à nouveau `rm`.

Exemple

```
> rm(n, N)
```

- tous les objets en mémoire, utilisez `rm`, combinée avec d'autres fonctions.

Exemple

```
> rm(list = ls())
```

3.8 comme calculatrice

Exemple

Tapez la ligne de commande suivante et validez par la touche « Entrée » du clavier :

```
> 2 + 8
```

```
[1] 10
```

Le chiffre 1 entre crochets indique l'indice du premier élément de la ligne. Le second chiffre, 10, est le résultat de la ligne de commande.

Exemple

Tapez la ligne de commande suivante :

```
> 120:155
```

```
[1] 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135  
[17] 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151  
[33] 152 153 154 155
```

Dans le résultat ci-dessus, l'indice de l'élément 120 est 1 et celui de 143 est 24.



Le chiffre qui se trouve entre les crochets indique l'indice du premier élément de la ligne sur laquelle il se trouve.

Une fonction que vous allez utiliser très souvent est la fonction `sqrt`, qui n'est rien d'autre que la racine carrée d'un nombre.

Exemple


Tapez la ligne de commande suivante :


```
> sqrt(4)
```

Le résultat s'affiche sous la forme :

```
[1] 2
```

4 ÉCRIRE ET COMPILER DES SCRIPTS SOUS

Il est souvent plus pratique de composer le code  dans une fenêtre spécifique du logiciel : la fenêtre de script.

Les entrées **Nouveau script** ou **Ouvrir un script** permettent de créer un nouveau script de commandes  ou d'accéder à un ancien script sauvegardé lors d'une session précédente d'utilisation du logiciel.

Pour exécuter des instructions à partir de la fenêtre de script il suffit de procéder par copier-coller ou de se servir du raccourci clavier "ctrl+R".

Pour sauvegarder un script, il suffit, lorsque la fenêtre de script est active, de sélectionner l'entrée « Sauver » du menu « Fichier ».

Des scripts s'écrivent avec un éditeur de texte.

Exemple

Winedt, MS Wordpad, Open Office, vi, emacs ou Xemacs, etc, sont des éditeurs de texte.



Souvent il est proposé aux débutants d'utiliser le traitement de texte de Word car la plupart des mises en forme du texte n'affecte pas l'exécution des lignes de commande lors du « copier-coller ». Attention! Il faut désactiver la mise en forme automatique du texte.



Certains éditeurs, comme emacs, Xemacs ou Tinn-R peuvent vous être conseillés car ils vous rendront la vie plus facile en fournissant la tabulation intelligente, la coloration syntaxique et le complément des commandes.

Indépendamment de l'éditeur, vous sauvegardez les scripts dans un de vos répertoires. Ensuite deux solutions s'offrent à vous :

1. soit vous utilisez le copier-coller sur la fenêtre R Console,
2. soit vous les lisez avec l'une des trois commandes suivantes :

```
> source(file="C://chemin//vers//nomdefichier//
+ fichier.R",echo=T)
```


ou

```
> source(file="../../../repertoire/fichier.R",echo=T)
```


ou encore

```
> source("fichier.R",echo=T)
# Si "fichier.R" est dans le répertoire de travail
```



1. Notez l'utilisation des slashes pour séparer les répertoires, même sous l'environnement Windows.
2. Notez la présence du symbole #. Sous , ce symbole est le symbole des commentaires. Tout ce qui suit un # est alors ignoré. Sachez qu'un élément clef d'une bonne écriture de script est la présence abondante de commentaires.



Écrire des scripts lisibles et commentés est une habitude à prendre dès maintenant car cela vous rendra l'utilisation du langage  bien plus facile.

Les quatre avantages d'écrire des scripts

1. Écrire des scripts est un gain de temps car vous évitez de réécrire des lignes de commande déjà écrites.
2. Si vous avez beaucoup de lignes de commande à écrire, c'est beaucoup plus simple de les manipuler, de les modifier dans un éditeur de texte.
3. Écrire des scripts est un outil de collaboration puissant. C'est souvent pratique de pouvoir envoyer à un de vos camarades ou de vos collègues, par fichier attaché dans un mail par exemple, votre code et le fichier de données

brut associé et de savoir qu'il lui suffit d'exécuter la fonction source sur votre code pour effectuer votre analyse sur sa machine.

4. Enfin il n'existe pas de message d'alerte dans R sauf quand vous quittez R. Vous pouvez alors perdre des données sans vous en rendre compte. Le seul moyen de trouver l'erreur est de recommencer l'écriture de la ligne de commande.



Il est à noter qu'il existe également des logiciels de programmation libres et gratuits comme Tinn-R qui sont destinés à vous faciliter la rédaction de scripts.

5 R SANS LES MAINS

Il existe plusieurs interfaces graphiques qui ont vocation à faciliter l'utilisation du langage R. En voici deux parmi celles-ci qui méritent une attention particulière.

- RStudio, <https://www.rstudio.com>, est avant tout un outil puissant pour écrire facilement des scripts, des fonctions voire des bibliothèques R. C'est un programme à installer séparément de R et disponible pour les trois environnements Windows, macOS X et Linux. Il présente dans un même environnement le script, la console R, la liste ainsi que le contenu des objets présents dans la mémoire de R ainsi que les pages d'aide consultées ou les graphiques produits. Il permet également de transformer automatiquement un script R en un fichier au format html où sont intercalées les commandes R avec les résultats de celles-ci.
- Rcmdr (RCommander) est une bibliothèque pour le langage R, disponible pour les trois environnements Windows, macOS X et Linux. C'est avant tout une interface graphique pour un grand nombre de fonctions usuellement utilisées en statistique. Son utilisation rend la pratique de R proche de celles d'autres logiciels de statistique « à menus et boîtes de dialogue » comme par exemple SPSS, Minitab ou Statistica. Elle permet en outre d'importer facilement des fichiers au format .csv, SPSS, SAS ou Minitab. Cette bibliothèque permet, elle aussi comme RStudio, de transformer automatiquement un script R en un fichier au format html où sont intercalées les commandes R avec les résultats de celles-ci.

Pour vous montrer la facilité avec laquelle il est possible de produire ces rapports html ainsi que le rendu final obtenu, vous trouverez en ligne, pour chacun des chapitres du livre, des rapports, faits avec RStudio et Rcmdr.

Pour aller plus loin

- Un certain nombre de livres écrits principalement en anglais, mais aussi en français, paraissent chaque année. Une liste de manuels d'introduction ou très spécialisés est à votre disposition sur le site <https://cran.r-project.org/manuals.html> dans l'onglet Documentation.
- Pour un public francophone, un point de départ peut être le polycopié d'Emmanuel Paradis, téléchargeable en ligne, intitulé « R pour les débutants », 77 pages (au mois de septembre 2019), qui a la particularité d'exister également en version anglaise « R for Beginners ». Les deux documents

sont disponibles à cette adresse <https://cran.r-project.org/> dans la rubrique « Documentation », sous-rubrique « Contributed documentation ».

- Plusieurs milliers de pages d'enseignement de statistiques sous R, rédigées en langue française, sont disponibles à cette adresse : <http://pbil.univ-lyon1.fr/R/>.
- Il existe aussi des groupes ou des foires aux questions autour de R auxquels vous pouvez vous abonner.
- Il a existé les R News mais maintenant ces nouvelles sont remplacées par le R Journal. Les articles de ces revues ont pour objectif de mettre en avant certaines bibliothèques de fonctions particulièrement intéressantes.

Installation de RStudio



RStudio est un environnement de développement intégré (Integrated Development Environment, abrégé IDE en anglais), gratuit, libre et multiplateforme pour R. Il est disponible sous la licence libre AGPLv3 ou bien sous une licence commerciale, soumise à un abonnement annuel.

RStudio intègre la possibilité d'écrire des *notebooks* combinant de manière interactive du code R, du texte mis en forme en Markdown et des appels à du code Python ou Bash.

Relativement léger et ergonomique il est vivement conseiller de l'utiliser lorsqu'il faut réaliser des développements plus conséquents.

1. Installez RStudio sur votre ordinateur. Pour cela, téléchargez-le à l'adresse suivante : <https://www.rstudio.com/products/rstudio/download>. Vous trouvez dans le bas de la page la section **Installers for Supported Platforms**. Cliquez alors sur **RStudio 1.2.1578 - Windows 10/8/7 (64-bit)** (à l'heure où le livre est écrit la dernière version de RStudio est datée 17 septembre 2019). Une fois le fichier téléchargé, lancez le fichier exécutable.

2. Lancez RStudio.

En bas à gauche vous trouvez la console. Cette console correspond à celle que obtenez en lançant R directement.

Cependant ce n'est pas très pratique de travailler directement dans la console. La saisie y est réalisée ligne à ligne. En cas d'erreur de saisie, il faut généralement tout recommencer. Un avantage de RStudio est de pouvoir visualiser toutes les lignes de commande dans un seul fichier via un script.

Définition 1.1

Un script est un fichier de type texte dans lequel il est possible de saisir directement une séquence d'instructions pour l'exécuter par la suite. La saisie est réalisée soit dans un éditeur de texte quelconque (Bloc-Notes, Notepad++, etc.), soit directement sous RStudio.

Pour lancer un script depuis RStudio, allez dans le menu et allez vers : **File → New File → R Script**. Une fenêtre d'édition s'ouvre sur laquelle vous pouvez écrire à la

suite vos commandes. Rappelez-vous bien que pour séparer deux commandes, il faut faire un saut à la ligne. Pour exécuter une commande particulière positionnez-vous sur la ligne où elle est écrite et lancez l'exécution via :

— par le bouton Run


— par le raccourci clavier Ctrl + Entrée (cmd + Entrée sous Mac).


Pour exécuter plusieurs commandes à la fois, sélectionnez les lignes correspondantes et lancez l'exécution de la même façon.

Exercices

1.1 Stockage d'une variable



Il est possible de déclarer et stocker une variable de type quantitative ou qualitative. Pour que  «se souvienn» d'une variable, il suffit de l'initialiser. Cela passe par l'affectation d'une valeur à cette variable. L'opération d'affectation est réalisée par l'opérateur `<-`.

La fonction `str` décrit l'objet. Cette fonction peut être utilisée sur n'importe quel objet de .

1. Affectez à `maVariable` la valeur 10 et affichez le résultat dans votre Console.
2. Quelle est la structure de `maVariable` ? Pensez à utiliser la fonction `str`.
3. Multipliez par 2 `maVariable`.
4. Multipliez `maVariable` par elle-même.
5. Changez la valeur de `maVariable` en 11 et affichez le résultat dans votre Console.
6. Affectez à `maVariable2` la valeur `NULL` et affichez le résultat dans votre Console.
7. Quelle est la structure de `maVariable2` ?
8. Affectez à `maVariable3` la chaîne de caractères `Hello World` et affichez le résultat dans votre Console.
9. Quelle est la structure de `maVariable3` ?

1.2 Manipuler des vecteurs



Le symbole `:` sous  s'interprète comme « jusqu'à ».

1. Affichez les valeurs de `1 « jusqu'à » 10` dans votre console.
2. Affectez à la variable `a` les valeurs de `1 « jusqu'à » 10` et affichez `a`.
3. Quelle est la nature de `a` ?
4. Est-ce possible d'additionner `a` avec lui-même ? Si oui, alors exécutez le calcul.

5. Est-ce possible de multiplier a avec une valeur réelle ? Si oui, alors exécutez le calcul avec le réel 10.
6. Vérifiez, pour chaque élément de a , s'il est supérieur à 5.
7. Sélectionnez les éléments de a supérieurs à 5.
8. Construisez le vecteur b composé des chiffres allant de 5 à 8.
9. Affichez le premier, le quatrième et le cinquième élément du vecteur b . Que remarquez-vous pour le cinquième élément ?
10. Affectez 9 au septième élément de b et affichez b .
11. Affichez les nombres de 5 à 30 de 5 en 5 grâce à la fonction `seq`.

1.3 Définir une fonction



Avec `R`, vous pouvez créer vos propres fonctions. **Exemple** : Soit la fonction f définie par $f(x) = 3 * x^2$. Tapez les commandes suivantes afin de stocker dans `R` la fonction f :

```
> f <- function(x){return (3*x^2)}
```

Créez la fonction f définie par : $f(x) = \frac{5 * x^5}{4}$. Calculez $f(3)$. Représentez graphiquement la fonction f sur l'intervalle $[-10; 10]$.

1.4 Structures de contrôle



Comme pour les langages informatiques usuels, `R` dispose de structures de contrôle sur des valeurs :

```
> a <- 3
> if (a > 5) {
+ print("a est plus grand que 5")
+ } else {
+ print ("a est plus petit que 5")
+ }
```

```
[1] "a est plus petit que 5"
```

Affichez 5 fois, à l'aide d'une boucle `for`, le message `Hello world`.