

Introduction au langage *R* avec *RStudio*

Dominic Comtois

Octobre 2018

© *Tous droits réservés. Aucune partie de cet ouvrage ne peut être reproduite ou retransmise sans l'autorisation expresse de l'auteur.*

Contact: dominic.comtois@gmail.com

Table des matières

1	Introduction	1
1.1	Pourquoi apprendre <i>R</i> ?	1
1.2	Quelques domaines d'utilisation	1
1.3	<i>R</i> est un langage de programmation	2
1.4	Approche pédagogique	2
2	De la ligne de commande à <i>RStudio</i>	5
2.1	Ligne de commande	5
2.2	Interface graphique de base	6
2.3	<i>RStudio</i>	7
2.4	Installation des <i>packages</i> requis	8
2.5	Interfaces de type pointer-cliquer	9
2.5.1	<i>R Commander</i>	9
2.5.2	<i>JGR & Deducer</i>	10
2.5.3	<i>Rattle</i>	11
2.5.4	<i>Radiant</i>	12
2.6	Exercices	13
3	Les variables	15
3.1	Affectation et modification	15
3.2	Noms de variables	16
3.2.1	Mots réservés	16
3.2.2	Bonnes pratiques	16
3.3	Copie et suppression	17
3.4	Liste des variables disponibles	17
3.5	Exercices	17
4	Les opérateurs	19
4.1	Opérateurs arithmétiques	19
4.2	Opérateurs de comparaison	20
4.3	Opérateurs logiques	20
4.4	Autres opérateurs	21
4.5	Exercices	21
5	Les vecteurs	23
5.1	Création de vecteurs de taille > 1	23
5.2	Mode des données	23
5.3	Les conversions	24
5.3.1	Conversions automatiques	24
5.3.2	Conversions explicites	25

5.4	L'indexation	25
5.4.1	Extraction	25
5.4.2	Affectation	26
5.5	Taille et ajout d'éléments	26
5.6	Le recyclage	27
5.6.1	Cycles incomplets	27
5.6.2	Incohérences lors d'affectations	27
5.7	Les noms (<i>names</i>)	28
5.8	Exercices	29
6	L'environnement <i>RStudio</i>	31
6.1	Configuration initiale	31
6.2	Les projets <i>RStudio</i>	31
6.3	L'interface <i>RStudio</i>	33
6.3.1	Volet supérieur droit	33
6.3.2	Volet inférieur droit	35
6.4	L'éditeur de scripts	36
6.5	Raccourcis clavier à connaître	36
6.6	<i>Addins</i>	38
6.7	Exercices	38
7	Les fonctions	39
7.1	Les arguments	39
7.1.1	Ordre des arguments	40
7.1.2	L'argument "... "	40
7.2	Fonctions statistiques de base	41
7.3	Fonctions statistiques et données manquantes	41
7.4	Opérateurs et fonctions	41
7.5	Arguments obligatoires <i>vs</i> facultatifs	42
7.6	Effets secondaires	42
7.6.1	Exercice	42
7.7	Fonctions génératrices de séquences	42
7.7.1	<code>seq()</code>	42
7.7.2	<code>rep()</code>	43
7.8	<code>is.na()</code>	44
7.9	Fonctions d'identification	44
7.10	Générateurs de nombres aléatoires	44
7.11	Fonctions génériques	45
7.12	Écrire ses propres fonctions	45
7.12.1	Utilisation de <code>return()</code>	46
7.13	Exercices	47
8	Matrices et tableaux	49
8.1	Les matrices	49
8.1.1	<code>length()</code> , <code>dim()</code> , <code>nrow()</code> et <code>ncol()</code>	50
8.1.2	Les noms de rangées et de colonnes	50
8.1.3	L'indexation des matrices	51
8.1.4	Opérations matricielles	52
8.2	Statistiques simples par rangées ou colonnes	53
8.2.1	Solution générale: <code>apply()</code>	53
8.3	Les tableaux (<i>arrays</i>)	53

8.4	Exercices	54
9	Les <i>data frames</i>	55
9.1	<i>data frames</i> inclus dans l'installation <i>R</i>	55
9.2	Différences et similarités avec les matrices	56
9.2.1	Comportement des fonctions	56
9.3	L'opérateur <code>\$</code>	56
9.3.1	Extraction de colonnes avec <code>\$</code>	56
9.3.2	Ajout de colonnes avec <code>\$</code>	57
9.4	Chargement et visualisation	57
9.5	Indexation pour <i>data frames</i> et <i>tibbles</i>	57
9.6	Saisie de données dans un <i>data frame</i>	59
9.6.1	Via l'interface graphique	59
9.6.2	Via la console ou l'éditeur de scripts	61
9.7	Conversions matrices \leftrightarrow <i>data frames</i>	61
9.8	Structures de données vues jusqu'ici	62
9.9	Les listes	62
9.10	Exercices	63
10	Les facteurs	65
10.1	Composantes d'un facteur	65
10.2	Création à partir d'un vecteur	65
10.3	Représentation numérique interne	66
10.4	L'option <i>stringsAsFactors</i>	66
10.5	Exercices	67
11	Les chaînes de caractères	69
11.1	Guillemets simples et doubles	69
11.2	Concaténer avec <code>paste()</code>	69
11.3	Séquences d'échappement	70
11.3.1	Chemins de fichiers sous <i>Windows</i>	71
11.4	Formatage avec <code>sprintf()</code>	72
11.5	Rechercher/remplacer: <code>sub()</code> et <code>gsub()</code>	73
11.6	Les expressions rationnelles	74
11.6.1	<code>grep()</code> et <code>grepl()</code>	74
11.7	Exercices	75
12	Importation et exportation de données	77
12.1	Fichiers texte (délimités)	77
12.1.1	Importation depuis l'interface <i>RStudio</i>	77
12.1.2	Lecture avec flexibilité accrue: <code>read.table()</code>	79
12.1.3	Écriture avec <code>write.table()</code>	80
12.2	Fichiers de données <i>R</i> (.Rdata)	81
12.2.1	Chargement	81
12.2.2	Sauvegarde	81
12.3	Fichiers <i>Excel</i>	81
12.3.1	Lecture avec <code>readxl</code>	82
12.3.2	Écriture avec <code>XLConnect</code>	83
12.4	Tables <i>SAS</i>	83
12.4.1	Fichiers <i>.sas7bdat</i>	83
12.4.2	Fichiers <i>SAS Transport</i>	84

12.5	Fichiers de données <i>SPSS</i>	84
12.6	Exercices	84
13	Rapports sommaires	85
13.1	Tableaux de fréquences	85
13.1.1	<code>table()</code> et <code>summary()</code>	85
13.1.2	<code>summarytools::freq()</code>	85
13.2	Statistiques descriptives univariées	86
13.2.1	<code>summary()</code> et <code>fivenum()</code>	86
13.2.2	<code>summarytools::descr()</code>	86
13.3	Tableaux croisés	88
13.3.1	<code>table()</code>	88
13.3.2	<code>summarytools::cTable()</code>	88
13.3.3	<code>gmodels::CrossTable()</code>	88
13.3.4	Contingences pour 3 variables ou plus	89
13.4	Sommaires complets avec <code>dfSummary()</code>	89
13.5	Traitement en sous-groupes	90
13.5.1	<code>by()</code>	91
13.5.2	<code>aggregate()</code>	91
13.6	Autres fonctions utiles	92
13.6.1	<code>str()</code>	92
13.6.2	<code>head()</code> et <code>tail()</code>	92
13.6.3	<code>summary()</code>	93
13.7	Exercices	94
14	Transformation des données	95
14.1	Tri avec <code>order()</code>	95
14.2	Recodage	95
14.2.1	Variables numériques (découpage)	95
14.2.2	Recodages complexes par indexation	97
14.2.3	Recodages avec <code>ifelse()</code>	97
14.3	Combinaison de colonnes avec <code>cbind()</code>	98
14.4	Exercices	99
14.5	Jointures avec <code>merge()</code>	100
14.6	Transposition	101
14.7	Restructuration	102
14.7.1	Avec <code>reshape2</code>	102
14.7.2	Avec <code>tidyr</code>	103
14.8	Exercices	103
15	Manipulations avec <i>dplyr</i>	105
15.1	L'opérateur <code>%>%</code>	105
15.2	Les fonctions <i>dplyr</i>	106
15.2.1	Sélection de colonnes et d'observations	106
15.2.2	Calculs avec <code>mutate()</code>	107
15.3	Calculs par sous-groupes	107
15.4	Jointures	109
15.4.1	<code>left_join()</code>	109
15.4.2	<code>inner_join()</code>	110
15.5	Exercices	111

16 Tests statistiques	113
16.1 Khi-carré	113
16.2 Équivalence entre 2 proportions	114
16.3 Test-T	114
16.3.1 Échantillon unique	114
16.3.2 Échantillons indépendants	114
16.3.3 Échantillons pairés	115
16.4 Anova	115
16.5 Exercices	117
17 Les graphiques	119
17.1 La fonction générique <code>plot()</code>	119
17.2 Vecteurs numériques	120
17.2.1 Histogrammes	120
17.2.2 Boîtes à moustaches	120
17.2.3 Combiner <code>hist()</code> et <code>boxplot()</code>	121
17.3 Facteurs ou vecteurs de valeurs numériques discrètes	121
17.3.1 Graphiques en bâtons	121
17.3.2 Diagrammes circulaires	122
17.4 Paramètres graphiques	123
17.5 Exercices	124
17.6 Graphiques à deux variables	124
17.7 Ajout d'éléments graphiques	126
17.8 Exercices	127
17.9 Graphiques <i>ggplot</i>	128
17.9.1 Histogrammes	128
17.9.2 Boîtes à moustaches	128
17.10 Interfaces graphiques à la rescousse	129
17.10.1 Exercice	129
18 Les structures de contrôle	131
18.1 Conditions	131
18.1.1 <code>if - else if - else</code>	131
18.1.2 la fonction <code>ifelse()</code>	133
18.2 Boucles	134
18.2.1 Boucles <code>for()</code>	135
18.2.2 Boucles <code>while()</code>	135
18.2.3 Boucles <code>repeat()</code>	135
18.2.4 Application concrète	136
18.3 Exercices	137
19 Où trouver de l'aide	139
19.1 Menu Aide de <i>RStudio</i>	139
19.2 Sites Web	139
19.3 Fonctions et <i>packages</i>	139
19.3.1 <code>example()</code>	139
19.3.2 <code>demo()</code>	140
19.3.3 <code>apropos()</code>	140
A Installation de <i>R</i> et de <i>RStudio</i>	141
A.1 Installation de <i>R</i>	141

A.2	Installation de <i>RStudio</i>	145
B	Classe, type et mode	147
C	Expressions rationnelles	149
C.1	Métacaractères	149
C.2	Quantifiants	149
C.2.1	Forme <i>greedy</i>	149
C.2.2	Forme <i>non-greedy</i>	150
C.2.3	Forme possessive	150
C.3	Séquences d'échappement	150
C.4	Ressources	151
D	Formats de données chronologiques	153
E	Distributions et nombres aléatoires	155
F	Fichiers de configuration	157
F.1	Configuration initiale	157
F.1.1	Alternative au fichier <i>.Rprofile</i>	157
G	Corrigé des exercices	159
	Index	168

Liste des tableaux

4.1	Les opérateurs arithmétiques	19
4.2	Les opérateurs de comparaison	20
4.3	Les opérateurs logiques	20
4.4	Autres types d’opérateurs	21
5.1	Mode des vecteurs	24
7.1	Fonctions statistiques de base	41
7.2	Paramètres de la fonction <code>seq()</code>	43
7.3	Paramètres de la fonction <code>rep()</code>	43
8.1	Paramètres de la fonction <code>matrix()</code>	49
8.2	Opérations matricielles communes	52
9.1	Certains <i>data frames</i> inclus dans <i>R</i>	56
9.2	Résultat de fonctions avec matrices et <i>data frames</i>	56
9.3	Principales formes d’indexation	59
9.4	Structures de données abordées jusqu’ici	62
11.1	Paramètres de la fonction <code>paste()</code>	69
11.2	Séquences d’échappement	71
11.3	Éléments de gabarits <code>sprintf()</code>	72
12.1	Paramètres de <code>read.table()</code> pris en charge par l’interface <i>RStudio</i>	79
12.2	Paramètres additionnels de la fonction <code>read.table()</code>	80
14.1	Paramètres de la fonction <code>merge()</code>	101
15.1	Terminologie <i>R-base</i> , <i>dplyr</i> et SQL	106
B.1	Propriétés des objets les plus courants	147
D.1	Formats communs de données chronologiques (voir <code>?strptime</code>)	154