# Les objets (partie 2)

*Thibault LAURENT*

*17 Octobre 2018*

## Contents



Ce document a été généré directement depuis **RStudio** en utilisant l'outil Markdown. La version .pdf se trouve ici.

**Rappel :**

Le répertoire de travail utilisé dans cette session (à vous de le modifier selon votre ordinateur):

```
setwd("Z:/m2_foad/cours_r/chapitre2")
```

Vous devrez également installé le package suivant :

```
install.packages("sas7bdat")
```

# Gestion des objets : Entrée - Sortie

Cette section est dédiée à la gestion des objets dans **R**. Son apprentissage doit vous permettre de vous familiariser avec les outils permettant de sauvegarder et de charger des objets. Par ailleurs, **R** étant avant tout un logiciel de traitement de données, un paragraphe sera consacré à la gestion des jeux de données (importation, exportation).

## Enregistrement d'un fichier de codes *.R*

Un fichier de codes doit porter l'extension *.R* et ne contenir que du code **R** et des commentaires. On présente ici un extrait du fichier *Z:/m2_foad/cours_r/chapitre2/chapitre2_partie1.R* :

```
# Ce fichier contient les lignes de commandes présentées dans le chapitre 2
# du cours d'introduction à R

# choix du répertoire de travail
setwd("Z:/m2_foad/cours_r/chapitre2")
# chargement des données
load(file("http://www.thibault.laurent.free.fr/cours/Ressource/diamants.RData"))

# Codes présentés dans la première section
# 1.1. Les vecteurs
# option 1
a.numeric <- c(1.2, 3.5, 5.4, 6.2, 8.1)
is.numeric(a.numeric)
```

```
# option 2
a.numeric <- numeric(5) # on crée un vecteur de numeric de taille 5
a.numeric[1] <- 1.2 # on affecte la valeur 1.2 au 1er élément du vecteur
a.numeric[2] <- 3.5 # etc.
a.numeric[3] <- 5.4
a.numeric[4] <- 6.2

# etc.
```

Lorsque vous sauvegardez ce fichier depuis **RStudio**, il vous suffit de faire *"File < Save as < chapitre2_partie1.R"*. Penser à le faire régulièrement au cours d'une session.

Dès que vous ouvrirez une nouvelle session de **RStudio**, aucun des objets que vous aurez créé dans une session précédente n'aura été conservé. Une façon d'exécuter tout le code contenu dans *chapitre2_partie1.R* est de faire en début de session :

```
source("Z:/m2_foad/cours_r/chapitre2/chapitre2_partie1.R")
```

Chaque utilisateur a sa façon de travailler. Plutôt que d'éxécuter toutes les instructions du fichier *chapitre2_partie1.R*, on peut vouloir ne récupérer que certains objets. C'est ce qu'on va voir dans la section suivante.

## Enregistrement et chargement des objets

L'enregistrement d'un objet dans **R** est très simple. Il se fait à l'aide de la fonction *save()* en précisant comme premier(s) argument(s), l'objet (ou les objets) à sauvegarder et avec l'argument **file=** le nom du fichier (avec une extension *.RData*). Lorsque vous redémarrerez une session, il suffira alors d'utiliser la fonction *load()* en précisant l'adresse complète du fichier à charger.

Reprenons l'objet **don** dans la 1ère partie du chapitre 2. Pour le créer, on avait du exécuter toutes les commandes suivantes :

```
age <- c(20, 21, 20, 25, 29, 22)
taille <- c(165, 155, 150, 170, 175, 180)
sexe <- c("F", "F", "F", "M", "M", "M")
don <- data.frame(age, taille, sexe)
row.names(don) <- c("sonia", "maud", "iris", "mathieu", "amin", "gregory")
don$diplome <- c("DU", "M2", "M2", "DU", "DU", "M2")
don <- cbind(don, pays = c("FR", "FR", "SNG", "CAM", "HAI", "BF"))
don <- rbind(don, c(21, 180, "F", "DU", "FR"))
don2 <- data.frame(age = c(20, 21), taille = c(180, 175), sexe = c("M", "F"),
                   diplome = c("DU", "DU"), pays = c("FR", "ESP"))
row.names(don2) <- c("pierre", "sonia")
don <- rbind(don, don2)
don3 <- data.frame(note_algebre = c(18, 10, 8, 15, 20, 5, 17, 12, 8),
                   nom = c("sonia1", "pierre", "7", "gregory", "amin",
                           "mathieu", "iris", "maud", "sonia"))
don <- merge(don, don3, by.x = "row.names", by.y = "nom")
```

Si dans une nouvelle session, on souhaite effectuer des statistiques sur l'objet final **don**, les étapes intermédiaires ne sont plus nécessaires. Aussi, plutôt que d'exécuter à chaque fois l'ensemble de ces commandes, on aimerait pourvoir charger directement l'objet **don**. Pour cela, il faut d'abord le sauvegarder.

La sauvegarde de cet objet s'effectue par l'instruction :

```
save(don, file = "donnees_don.RData")
```

En l'état, le fichier sera enregistré dans le répertoire de travail courant. Si vous souhaitez l'enregistrer autre part, précisez un nouveau répertoire de travail avant la sauvegarde (avec la fonction *setwd()*) ou indiquez le chemin d'accès avant le nom du fichier :

```
save(don, file="C:/autre_repertoire/donnees_don.RData").
```

Lors d'une nouvelle session, l'instruction :

```
setwd("Z:/m2_foad/cours_r/chapitre2")
load("donnees_don.RData")
```

chargera l'objet **don**.

Une autre manoeuvre consiste à sauvegarder la totalité des objets créés lors d'une session par la commande :

```
save.image("chapitre2_objets.RData")
```

et de charger cette dernière lors d'une nouvelle session :

```
setwd("Z:/m2_foad/cours_r/chapitre2")
load("chapitre2_objets.RData")
```

Ce procédé peut être utile lorsque l'on souhaite sauvegarder un nombre important d'objets. On notera que lorsque vous quittez une session **R**, vous avez une fenêtre qui s'affiche en disant : *Save workspace image*. Si vous cliquez sur "oui", cela revient à exécuter la commande *save.image()*. Il peut cependant avoir l'inconvénient (dans le cas où l'environnement de travail est celui par défaut), de charger à l'ouverture de la nouvelle session plusieurs objets *parasites* inutiles pour l'utilisateur, d'où l'intérêt d'utiliser parfois la fonction *ls()* pour afficher les objets créés dans l'espace de travail et la fonction *rm()* (*remove*) qui permet de supprimer les objets qu'on n'utilise plus. Par exemple, on souhaite créer ici deux vecteurs **x** et **y** :

```
eps <- rnorm(7)
a <- 3
b <- 4
x <- c(18, 17, 19, 20, 15, 19, 20)
y <- a*x + b + eps
rm(a, b, eps)
ls()
```

```
## [1] "age"    "don"    "don2"   "don3"   "sexe"   "taille" "x"      "y"
```

Ici, on a supprimé les objets **a**, **b** et **eps** (objets intermédiaires utilisés pour simuler le vecteur **y**), une fois qu'on n'en avait plus besoin. Pour supprimer tous les objets d'un seul coup, on peut utiliser la commande suivante :

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

**Remarque :** lorsqu'on début du cours, on vous a demandé d'éxécuter la commande suivante, vous avez ainsi chargé un fichier *".RData"* dans votre environement courant. Nous avons utilisé la fonction *file()* pour préciser qu'il fallait aller cherché ce fichier sur internet. Si pour une raison ou une autre, vous risquez de travailler sans accès à internet, il est donc conseillé de sauvegarder le jeu de données **diamants** dans un répertoire courant de votre machine :

```
load(file("http://www.thibault.laurent.free.fr/cours/Ressource/diamants.RData"))
save(diamants, file = "donnees_diamants.RData")
```

Un fichier au format *.RData* n'est donc pas un fichier contentant des données à proprement dites (comme des fichiers *.txt*). Il s'agit d'un format propre à **R** et si vous essayez de l'ouvrir avec un bloc-note, vous ne pourrez

pas lire son contenu. La section suivante a pour objectif de montrer comment on importe des données issues de fichiers de différents formats.

## Importation et Exportation de fichiers de données

Il est très rare, lors d'une analyse statistique, que le jeu de données sur lequel est basée cette étude soit directement utilisable. Dans la grande majorité des cas, un "nettoyage" et/ou une fusion de plusieurs tables s'imposent. **R** peut très bien remplir cette tâche et pour des utilisateurs familiers avec d'autres outils, on retrouvera sous **R** des commandes identiques à **Perl** ou **SQL**. Toutefois, l'utilisation de ces outils est non triviale pour des non-experts en SGBD (système de gestion de bases de données).

Dans ce paragraphe, on va s'intéresser aux outils permettant d'importer des fichiers de données autres que ceux dont le format est reconnu par le logiciel. D'autre part, toutes les personnes ne travaillant pas avec le même logiciel, il peut être utile de connaître les fonctions permettant de "rapatrier" des données d'autres logiciels statistiques. C'est ce que nous verrons dans un deuxième temps.

### Importation et Exportation de fichier-texte

Les fonctions permettant de lire des données dans un fichier texte sont nombreuses et efficaces dans **R**. Ce type de fichiers porte en général l'extension ".txt" et ".csv". Les fichiers provenant de Excel (".xls" ou ".xlxs") ne rentrent pas dans ce type de fichiers.

#### La fonction *read.table()*

Si vous disposez d'un fichier-texte (format *.txt*), vous pouvez le lire avec la fonction *read.table()*. Cette fonction (qui renvoie un objet de classe **data.frame**) admet un seul argument obligatoire : l'emplacement du fichier-texte à importer. Les autres arguments optionnels servent à préciser certaines caractéristiques du fichier (symbole utilisé pour séparer deux cellules, symbole utilisé pour les décimales, identification des valeurs manquantes, etc.). Voici les valeurs par défaut utilisés :

- **header = FALSE** : la première ligne du fichier à importer ne contient pas le nom des variables.
- **sep = ""** : le séparateur entre deux cellules est la tabulation TAB.
- **dec = "."** : pour séparer la partie entière de la partie décimale, le symbole utilisé est le point.

Bien évidemment, tous les fichiers-texte ne respectent pas nécessairement le même encodage et c'est pourquoi on est souvent à amener à jouer sur ces paramètres.

Considérons les fichiers "dontxt_correct.txt" et "dontxt_problem.txt" : enregistrer-les dans votre répertoire de travail (dans mon cas, je les ai enregistrés dans le sous répertoire de travail "Ressource"). Ces deux fichiers ont les mêmes données, mais celles-ci ont des caractéristiques un peu différentes :

Commençons par importer le premier jeu de données *"dontxt_correct.txt"* sans modifier les arguments d'entrée :

```
dontxt_correct <- read.table("Ressource/dontxt_correct.txt")
```

Pour vérifier que l'importation s'est correctement effectuée, on peut utiliser la fonction *str()* :

```
str(dontxt_correct)
```

```
## 'data.frame':    84 obs. of  6 variables:
##  $ VER: int  565 401 417 449 816 301 571 566 428 421 ...
##  $ JA1: int  4 6 7 3 8 7 6 6 4 9 ...
##  $ RO1: num  113 57.3 52.1 112.2 90.7 ...
##  $ TYJ: Factor w/ 2 levels "Sem","WE": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ TrH: Factor w/ 5 levels "0h-24h","0h-6h",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ IMP: int  0 2 0 2 2 1 1 2 2 0 ...
```

On vérifie qu'on ait le nombre de variables et d'observations attendu et aussi que les variables ont été codées dans le bon type, ce qui semble être le cas ici.

A présent, importons le deuxième fichier de données, toujours sans modifier les arguments d'entrée :

```
dontxt_problem <- read.table("Ressource/dontxt_problem.txt")
```

On utilise la fonction *str()* pour regarder le type des données et aussi la fonction *head()* pour afficher les premières lignes :

```
str(dontxt_problem)
```

```
## 'data.frame':    85 obs. of  7 variables:
##  $ V1: Factor w/ 85 levels "1","10","11",..: 85 1 12 23 34 45 56 67 78 84 ...
##  $ V2: Factor w/ 82 levels "100,00","1035,00",..: 82 52 29 33 39 73 17 55 53 38 ...
##  $ V3: Factor w/ 25 levels "0,00","1,00",..: 25 13 17 20 10 22 20 17 17 13 ...
##  $ V4: Factor w/ 85 levels "100,00","101,00",..: 85 9 59 56 8 80 46 77 76 79 ...
##  $ V5: Factor w/ 3 levels "Sem","TYJ","WE": 2 1 1 1 1 1 1 1 1 1 ...
##  $ V6: Factor w/ 6 levels "0h-24h","0h-6h",..: 6 5 5 5 5 5 5 5 5 5 ...
##  $ V7: Factor w/ 4 levels "0","1","2","IMP": 4 1 3 1 3 3 2 2 3 3 ...
```

```
head(dontxt_problem, 2)
```

```
##         V1     V2   V3     V4  V5        V6  V7
## 1 RowNames    VER  JA1    RO1 TYJ       TrH IMP
## 2        1 565,00 4,00 113,00 Sem 7h30-8h30   0
```

On se rend compte immédiatement que l'objet **dontxt_problem** n'est pas conforme. On en déduit donc que les paramètres par défaut de la fonction *read.table()* ne sont pas adéquats pour ce jeu de données. Parmi les choses qui ne vont pas, la première ligne contenant le nom des variables est considérée comme une observation. Pour éviter cela, il faut indiquer que cette ligne contient le nom des variables. Ceci peut être fait par l'ajout de l'argument **header=TRUE**.

Le symbole qui permet de séparer la partie entière de la partie décimale est la virgule, qui n'est pas celui par défaut. Il est donc nécessaire de l'indiquer en utilisant l'option **dec="."**.

Finalement, pour importer correctement ce jeu de données, il suffisait de faire :

```
dontxt_clean <- read.table("Ressource/dontxt_problem.txt", header = TRUE, dec = ",")
str(dontxt_clean)
```

```
## 'data.frame':    84 obs. of  7 variables:
##  $ RowNames: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ VER     : num  565 401 417 449 816 301 571 566 428 421 ...
##  $ JA1     : num  4 6 7 3 8 7 6 6 4 9 ...
##  $ RO1     : num  113 57.3 52.1 112.2 90.7 ...
##  $ TYJ     : Factor w/ 2 levels "Sem","WE": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TrH     : Factor w/ 5 levels "0h-24h","0h-6h",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ IMP     : int  0 2 0 2 2 1 1 2 2 0 ...
```

**Remarque** : pour la lecture du premier jeu de données, il n'était pas nécessaire d'ajouter l'option **header=TRUE**. En effet, la première ligne du fichier *"dontxt_correct.txt"* contient seulement une valeur de moins que le nombre de colonnes dans le fichier. Aussi, **R** a compris directement que la première ligne contenait le nom des variables et que la première colonne n'était pas une variable, mais correspondait au nom des individus. Ainsi, la première colonne du fichier n'apparaît pas sous forme d'une variable, mais comme étant l'identifiant des individus. Pour afficher les identifiants des individus, on fait :

```r
row.names(dontxt_correct)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
## [15] "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28"
## [29] "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39" "40" "41" "42"
## [43] "43" "44" "45" "46" "47" "48" "49" "50" "51" "52" "53" "54" "55" "56"
## [57] "57" "58" "59" "60" "61" "62" "63" "64" "65" "66" "67" "68" "69" "70"
## [71] "71" "72" "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" "83" "84"
```

### Autres fonctions utiles

Il existe d'autres fonctions permettant de lire des données dans un fichier-texte. Elles sont toutes plus ou moins proches de *read.table()* car elles ne différent que par le changement d'un ou quelques paramètres d'entrée. Parmi elles, on peut citer *scan()*, *read.csv()*, *read.delim()*, *read.fwf()*. Il est bon de les connaître (si leur utilisation est vraiment nécessaire), mais dans la grande majorité des cas, la fonction *read.table()* est suffisante pour arriver à ses fins.

On peut également citer la fonction *read.csv2()*, qui marche très bien lorsqu'on a sauvegardé un fichier sur Excel au format *".csv"* (en utilisant le "séparateur point-virgule").

L'exportation de données vers un fichier-texte se fait au moyen de la fonction *write.table()*. Les deux arguments obligatoires sont le nom de l'objet à exporter et le nom du fichier à créer. Les autres arguments sont les mêmes que ceux de la fonction *read.table()*, i.e. qu'ils servent à définir différents codes. Pour continuer avec l'exemple ci-dessus, on enregistre la table **.txt** en donnant l'instruction :

```r
write.table(dontxt_correct, "fichier-sortie.txt").
```

**Compléments :** on a vu qu'il est nécessaire d'avoir une petite idée de la structure du fichier initial de données avant d'utiliser le process d'importation. Pour cela, on peut afficher ligne par ligne les éléments de n'importe quel fichier texte en utilisant la commande **readLines**. Par exemple, pour savoir ce qu'il y a dans les deux premières lignes du fichier texte *"dontxt_problem.txt"*, on procède ainsi :

```r
readLines("Ressource/dontxt_problem.txt", n = 2)
```

```
## [1] "RowNames\tVER\tJA1\tRO1\tTYJ\tTrH\tIMP"
## [2] "1\t565,00\t4,00\t113,00\tSem\t7h30-8h30\t0"
```

Ceci nous permet de voir que les colonnes sont espacées par une tabulation et que la virgule délimite les nombres décimaux.

### Fichiers Excel (".xls" ou ".xlsx")

Pour le moment, nous recommandons de convertir les fichiers *.xls* et *.xlsx* au format *.csv* (avec le délimiteur point virgule) et de les importer en utilisant les fonctions vues ci-dessus. Le logiciel Excel étant un logiciel propriétaire, il existe une alternative provenant du monde du libre LibreOffice qui permet de faire ce type de conversion.

### Factor ou character

La plupart des fonctions ci-dessus possèdent un argument d'entrée qui permet à l'utilisateur de décider si les chaînes de caractère seront codées en **factor** ou **character**. Par défaut, on a vu que les chaînes de caractère étaient codées en **factor**. Or, ceci peut présenter certains inconvénients que nous présenterons dans le cours "R avancé". Pour choisir un codage des chaînes de caractère en **character**, on ajoute l'option **stringsAsFactors = FALSE**.

### Importation de données d'un autre logiciel statistique

Un module a été spécialement conçu pour pouvoir importer des fichiers de données produit par d'autres logiciels statistiques. Il s'agit de la librairie **foreign**, librairie intégrée à la version de base de **R**. Noter toutefois qu'il faudra la charger avant chaque utilisation par :

```
library("foreign")
```

Plusieurs systèmes statistiques sont pris en compte (**Minitab**, **SPSS**, **Stata**, ...), mais nous n'en verrons ici que quelques uns. Avant de commencer, il faut savoir qu'étant donné leur coût, il est très rare que plusieurs logiciels statistiques cohabitent. Cependant, il peut arriver que ce module soit utile si un travail est effectué à plusieurs dans différents environnements.

De plus, certaines revues statistiques comme CSBIGS ou Journal of Statistical Software éditent des articles de statistique appliquée. Pour pouvoir vérifier les résultats obtenus dans ces études, la publication des jeux de données est obligatoire et leur format dépend du logiciel qui a été utilisé par les chercheurs . Dans ce cas, la bibilothèque **foreign** peut être fort utile.

### SAS

Un fichier de données SAS porte l'extension *.sas7bdat*. Le package **sas7bdat** permet l'importation d'un jeu de données au format *.sas7bdat* à l'aide de la fonction *read.sas7bdat()*.

Si l'on prend la table de données baseball.sas7bdat et qu'on la copie dans le répertoire "Ressource", cela donne :

```
require("sas7bdat")
```

```
## Loading required package: sas7bdat
```

```
don.sas <- read.sas7bdat("Ressource/baseball.sas7bdat")
head(don.sas)
```

```
##             name         team no_atbat no_hits no_home no_runs no_rbi
## 1  Aldrete, Mike SanFrancisco      216      54       2      27     25
## 2 Allanson, Andy    Cleveland      293      66       1      30     29
## 3    Almon, Bill   Pittsburgh      196      43       7      29     27
## 4 Anderson, Dave   LosAngeles      216      53       1      31     15
## 5    Armas, Tony       Boston      425     112      11      40     58
## 6    Ashby, Alan      Houston      315      81       7      24     38
##   no_bb yr_major cr_atbat cr_hits cr_home cr_runs cr_rbi cr_bb   league
## 1    33        1      216      54       2      27     25    33 National
## 2    14        1      293      66       1      30     29    14 American
## 3    30       13     3231     825      36     376    290   238 National
## 4    22        4      926     210       9     118     69   114 National
## 5    24       11     4513    1134     224     542    727   230 American
## 6    39       14     3449     835      69     321    414   375 National
##   division position no_outs no_assts no_error salary
## 1     West       1O     317       36        1     75
## 2     East        C     446       33       20    NaN
## 3     East       UT      80       45        8    240
## 4     West       3S      73      152       11    225
## 5     East       CF     247        4        8    NaN
## 6     West        C     632       43       10    475
```

### STATA

Si l'on considère les données contenues dans le fichier automiss.dta (extrait du site de Stata), l'importation se réalise avec la fonction *read.dta()* comme suit :

```
automiss <- read.dta("Ressource/automiss.dta")
head(automiss)
```

```
##            make price mpg rep78 headroom trunk weight length turn
## 1   AMC Concord  4099  22    NA      2.5    11   2930    186   40
## 2     AMC Pacer  4749  17    NA      3.0    NA   3350    173   40
## 3    AMC Spirit  3799  22    NA      3.0    12   2640    168   35
## 4 Buick Century  4816  20     3      4.5    16   3250    196   40
## 5 Buick Electra  7827  15     4      4.0    20   4080    222   NA
## 6 Buick LeSabre  5788  18     3      4.0    21   3670    218   43
##   displacement gear_ratio  foreign
## 1          121       3.58 Domestic
## 2          258       2.53 Domestic
## 3          121         NA Domestic
## 4           NA       2.93 Domestic
## 5          350       2.41 Domestic
## 6           NA       2.73 Domestic
```

### SPSS

SPSS, comme Stata, ne pose aucun problème d'importation. Pour la table donspss.sav, le code est :

```
don.spss <- read.spss("Ressource/donspss.sav", to.data.frame = T)
head(don.spss)
```

```
##   ROWNAMES VER JA1       RO1 TYJ       TRH IMP
## 1        1  565   4 113.00000 Sem 7h30-8h30   0
## 2        2  401   6  57.28571 Sem 7h30-8h30   2
## 3        3  417   7  52.12500 Sem 7h30-8h30   0
## 4        4  449   3 112.25000 Sem 7h30-8h30   2
## 5        5  816   8  90.66667 Sem 7h30-8h30   2
## 6        6  301   7  37.62500 Sem 7h30-8h30   1
```