Introduction à l'application et à la programmation du logiciel *R* aux fins d'analyse statistique

Mamadou Sané

Présentation dans le cadre d'un stage doctoral au *Cdame*

7 Mars 2012

SOMMAIRE

Introduction

Présentation de R

- R est un système d'analyse statistique et graphique libre développé et distribué par le *R Development Core Team*.
- R est disponible sous Unix, Linux, Windows, et Macintosh
- R est un logiciel puissant permettant de réaliser la plupart des analyses statistiques

Introduction

R et les autres logiciels

- R peut faire les mêmes analyses statistiques que SPSS, SAS et Mystat
- R peut utiliser les données provenant d'autres logiciels statistiques
- R est orienté programmation et ne sort un output que si l'utilisateur le demande
- Il est possible d'utiliser des interfaces graphiques avec R (Rcommander, Rstudio).

Prise en main

Installation (1)

Windows

- À l'adresse http://cran.cict.fr/bin/windows/base/release.htm est proposé le téléchargement d'un fichier nommé R-2.X.X-win32.exe (les X étant remplacés par les numéros de la dernière version disponible);
- Exécutez-le et une fois l'installation terminée, vous devriez avoir une icône R sur votre bureau.

Prise en main

Installation (2)

Mac OS X

- Se rendre à la page suivante :

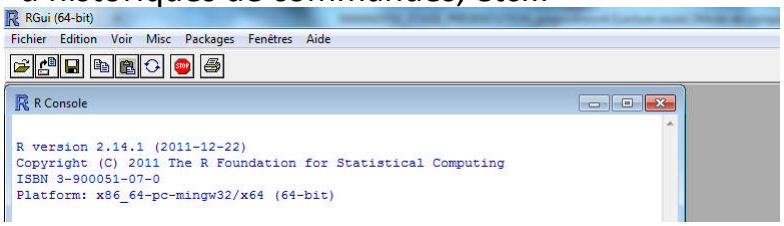
http://cran.r-project.org/bin/macosx/ ;

- télécharger le fichier nommé R-2.X.Y.dmg;
- double cliquer sur le fichier téléchargé; une fenêtre devrait s'ouvrir, contenant le programme d'installation;
- double cliquer sur le programme d'installation et suivre les instructions.

Prise en main

Ecran de travail et menu (1)

- Le menu Fichier comprend 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 de commandes, etc...



Prise en main

Ecran de travail et menu (2)

- Edition contient les habituelles commandes de copier-coller, ainsi que la boîte de dialogue autorisant la personnalisation de l'apparence de l'interface;
- Misc traite de la gestion des objets en mémoire et permet d'arrêter une procédure en cours de traitement.

Prise en main

Ecran de travail et menu (3)

- Packages automatise la gestion et le suivi des librairies de fonctions, permettant leur installation et leur mise à jour
- Fenêtres gère la définition spatiale des fenêtres (console=fenêtre des lignes de commande, fenêtre graphique, fenêtre de script…);
- Help donne accès aux manuels de références de R et à l'aide en ligne

Prise en main

Environnement de travail

L'environnement de travail comprend:

- Le répertoire courant de travail, là où sont enregistrés par défaut les objets créés;
- Les objets créés : graphiques, matrices, vecteurs, tableaux de données.

R donne la possibilité de sauvegarder l'environnement de travail et de le restaurer

Prise en main

Librairies

Un package ou une librairie est un ensemble de fonctions dédiées à une analyse statistique spécialisée. Quelques exemples:

- MASS contient les fonctions et les données de Venable, W. N. et Ripley, B. D. (2002). Modern applied statistics with S;
- Rcmdr est un interface graphique permettant de réaliser la plupart des analyses statistiques;
- rgrs est une librairie destinée de fonctions pour étudiants et chercheurs en sciences sociales

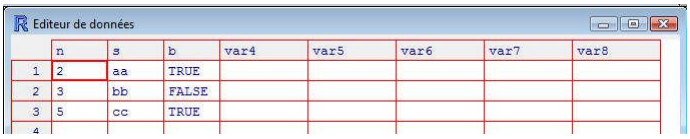
Acquisition et importation de données

Saisir dans R

On peut saisir directement une table de données (de taille assez faible) directement dans R en utilisant le format *data.frame*

Exemple:

n = c(2, 3, 5); s = c("aa", "bb", "cc"); b = c(TRUE, FALSE, TRUE) ; df = data.frame(n, s, b) ;edit(df)



	n	s	b	var4	var5	var6	var7	var8
1	2	aa	TRUE					
2	3	bb	FALSE					
3	5	cc	TRUE					

Acquisition et importation de données

Importation

- *read.table*: utilisé pour l'importation de fichiers en forme de tableau. Le séparateur par défaut est l'espace

Exemple: *read.table("donnees1_Howell_p347_EX11_1_comp.txt",header=TRUE, sep="\t",dec=",") # pour importer un fichier texte avec comme séparateur tabulation*

- *read.delim*, *read.delim2*, *read.csv*, *read.csv2* jouent le même rôle mais avec des séparateurs par défaut différents

Acquisition et importation de données

Importation

- *read.spss* permet d'importer des fichiers SPSS

- *scan* permet de lire des données de tout type

Exemple: *scan(file = "donnees1_Howell_p347_EX11_1_comp.txt", what =list(0,0,0,""),sep = "\t",skip=1)*

```
> scan(file = "donnees1_Howell_p347_EX11_1_comp.txt", what =list(0,0,0,""),sep = "\t",skip=1)
Read 6 records
[[1]]
[1] 15 10 25 15 20 18

[[2]]
[1] 30 15 20 25 23 20

[[3]]
[1] 40 35 50 43 45 40

[[4]]
[1] "Masculin" "Masculin" "Feminin"  "Feminin"  "Feminin"  "Masculin"
```

Manipulation des données

Qualitatives
Quantitatives
Mixtes

Discrètes

- Tableau de fréquences: commande table

Exemple: library(MASS) ; Data = read.table("EX10-2.dat", header=TRUE); Time.freq= table(Data$Time); Tabcroise = table(Data$Time,Data$Rating);

```
> library(MASS) ; Data = read.table("EX10-2.dat", header=TRUE); Time.freq= table(Data$Time);Tabcro$
> Time.freq

 0  1
13  7
> Tabcroise

   20 30 40 50 60
 0  1  3  4  4  1
 1  0  1  2  3  1
```

Manipulation des données

Qualitatives
Quantitatives
Mixtes

Quantitatives

- Fonctions:

  moyenne = mean; médiane = median;

  quantile = quantile; écart-type = sd;

  étendue = range; distance inter quartile = IQR

Mixtes

- tapply(var. quant., var. discrète., fonction)

Graphiques

Qualitatives
Quantitatives

Variables quantitatives

- Histogramme: hist
- Diagramme en tiges et feuilles: stem
- Nuage de points: plot
- Diagramme en boîte et moustaches: boxplot

Variables discrètes

- Diagramme à barre: barplot
- Diagramme circulaire: pie

Survol des analyses statisques

Test ind. Chi 2
Reg. linéaire simple
Anova

Variables discrètes

- Fonctions:
- Tableau croisé : freq = table(var1, var 2)
- Profils lignes (total des pourcentages en ligne=100)
  library(rgrs); lprop(freq);

Profils colonnes (idem en colonne) cprop(freq)

Test chi 2: chisq.test(freq)

Graphique associé: mosaicplot(var1 ~ var2)

Survol des analyses statisques

Regression linéaire simple

- Modèle: *md = lm(Y ~X, data=)*
- Résultats de l'estimation: *summary(md)*
- Coefficients: *md$coefficients*
- Prédiction avec int. de conf.: *predict(md, nouv. donn.,interval=``confidence'')*
- Nuage de points avec droite de regression:

plot(Y~X,data);abline(md$coefficients);

Diagnostic: *plot(md)*

Survol des analyses statisques

Analyse de variance simple

Comparer les moyennes de différents groupes d'une variable quantitative

- Modèle: *md = aov(Y ~Groupe, data)*
- Résultats de l'estimation: *summary(md)*
- Taille des effets: *model.tables(md)*
- Tests de comparaison multiples deux à deux: *pairwise.t.test(Y, Groupe)*

Exportation des résultats

Tableaux

- Copier/Coller vers Excel, Word
  install.packages("R2HTML",dep=TRUE);
  library(R2HTML); library(rgrs); copie(tableau);
- Exportation vers Word via un fichier:
  copie(tableau, file=TRUE, filename=``tableau1.html'')

Graphiques

- Via interface graphique: fichier/sauver sous
- Via commande: *dev.copy()*

Programmer avec R

Boucles et vecteurs

- La commande *c()* permet de créer un vecteur

Exemple: *x = c(0,0,1,0,1,1,1,1,0,0)*

- La commande *seq(debut, fin, pas)* permet de créer des suites arithmétiques de raison *pas*
- *Exemple : y = seq(0,100,5)*
- La commande pour les boucles est

for (indice de début:indice de fin){} et *while(condition){}.*

Exemple: pour créer un vecteur à partir de x où nous remplaçons 1 par femme et 0 par homme

X_recode = x

for (i=1:length(x)){if x[i]=0 then X_recode[i] = ``homme'' else X_recode[i] = ``femme'' }

Les instructions conditionnelles *if… else* peuvent être évitées en faisant simplement *X_recode[x==0] = ``homme''; X_recode[x==1] = ``femme'';*

Programmer avec R

Fonctions

- Les fonctions de type *apply* permettent d'éviter d'écrire des boucles

Exemple:

x <- list(a = 1:10, beta = exp(-3:3), logic = c(TRUE,FALSE,FALSE,TRUE)) # liste d'elements

lapply(x,mean) # applique la fonction moyenne à chaque élément de la liste

- On peut écrire sa propre fonction

Exemple:

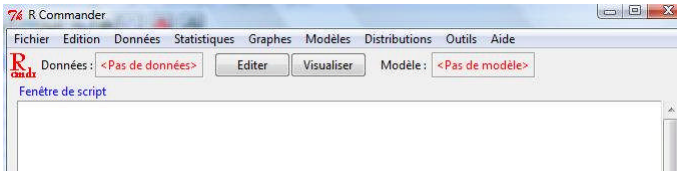
mafonction <- function(S) { R=(S-mean(S))/sd(S)
return(R) }
 # cette fonction sert à standardiser la variable
Data1 = read.table(``Ex9-22.dat'', header=TRUE)
mafonction(Data1$Expend)

```
> mafonction= function(S) {
+  R=(S-mean(S))/sd(S)
+  return(R) } # cette fonction sert à standar
> Data1 = read.table("Ex9-22.dat", header=TRUE
> mafonction(Data1$Expend)
 [1] -1.1008605814 -0.8271606914 -1.0612364686
 [7] -0.3098458861 -0.0647634109 -0.5050313304
[13] -1.3393390377 -0.3832238727 -0.1564858942
[19] -0.8293620310  0.1883906427 -0.7779974404
[25]  0.1480327501  0.7519335796  0.1869230830
[31]  0.8253115662 -0.1374076177 -0.5226420472
[37]  0.9830742373  1.0138929917 -0.0339446566
[43]  0.3894463259  0.8832801756  1.1474409272
[49] -0.4243155452  0.0005429971
```

Programmer avec R

Le script de Rcmdr

- library(Rcmdr) lance l'interface graphique de R commander;
- La plupart des analyses peuvent être réalisées via l'interface graphique de Rcmdr et le script sauvegardé. Ce script peut servir de code de programmation.
- Toutes les commandes saisies peuvent être récupérées grâce au menu Fichier/sauver l'historique des commandes



RÉFÉRENCES UTILES

Trouver l'aide à propos de R

En ligne

- *Help(''commande'')* ou *?commande* renvoit à l'aide en ligne
- Les exemples peuvent être directement exécutés avec *example(''commande'')*
- Lancer la documentation web avec *help.start()*
- Pour faire une recherche sur R, lancer le moteur de recherche: http://www.rseek.org/

Ressources officielles

- Accessible sur http://www.r-project.org/:http://cran.r-project.org/manuals.html; http://cran.r-project.org/doc/manuals/R-intro.html pour les débutants; http://cran.r-project.org/doc/manuals/R-data.html pour la manipulation des données; http://cran.r-project.org/doc/FAQ/R-FAQ.html pour les FAQ

RÉFÉRENCES UTILES

Ouvrages consultés

•Braun, W. J. et Murdoch, D. J. (2009). *A first course in statistical programming with R*. Cambridge, Royaume-Uni : Cambridge press.

•Howell, D. C.(2009). Méthodes statistiques en sciences humaines (6e édition). Bruxelles, Belgique : Éditions De Boeck université.

•Fox, J. (2011). *An R companion to applied regression*, 2e édition. Los Angeles, Californie : Sage.

•Lafaye de Micheaux, P. et collab.(2011). Le logiciel R : maîtriser le language, effectuer des analyses statistiques. New York, New Jersey: Springer.

•Spector, P. (2008). *Data manipulation with R*, 4e édition. New York, New Jersey : Springer.

•Venable, W. N. et Ripley, B. D. (2002). *Modern applied statistics with S*, 4e édition. New York, New Jersey : Springer.