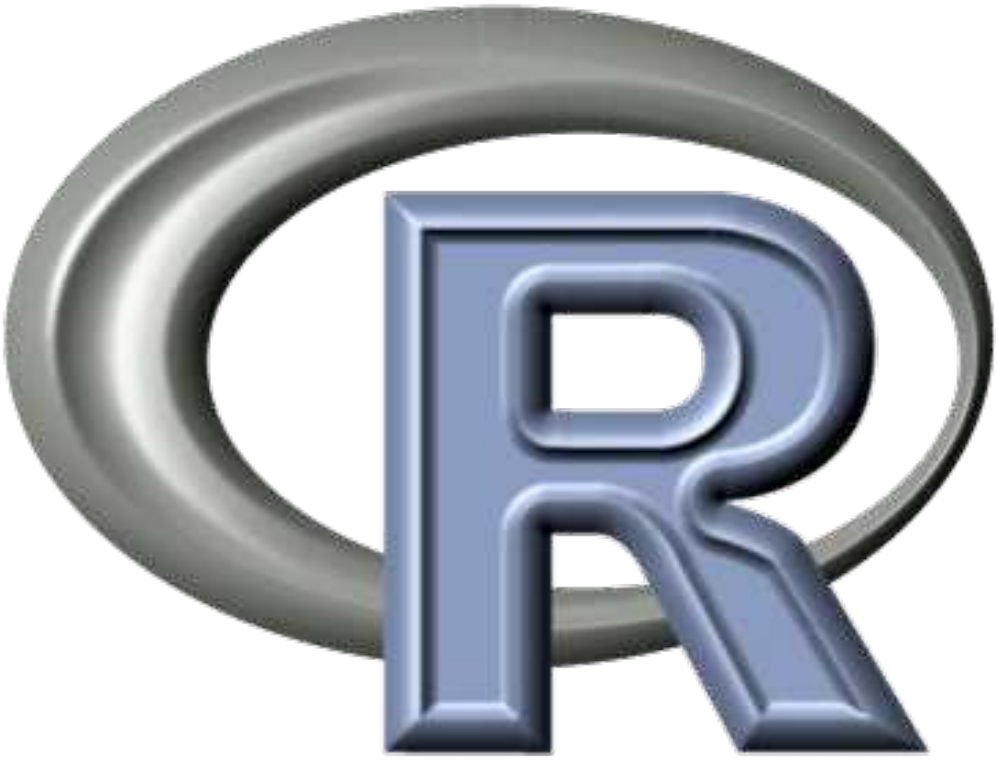
Analyses statistiques et représentations graphiques sous R

Sami Mestiri

Faculté des sciences économiques et de gestion Mahdia
mestirisami2007 ∂ gmail.com



Contents

Avant propos

La statistique est devenu de plus en plus une science de pleine actualité. Ses champs d'application se sont multipliés : Agronomie, Démographie, Économie, Marketing, Gestion de l'entreprise, Finance, Médecine, Sciences politiques, etc. Il est donc indispensable de connaitre les principes et les indicateurs fondamentaux de la statistique.

L'objet de ce livre est de fournir aux lecteurs des outils de traitements statistiques des données avec logiciel R. Chaque chapitre de comporte une partie théorique et des exercices d'applications. Les applications sont programmées sous le logiciel R.

Dans certains cas on est amené à étudier plusieurs caractères sur une même population. Dans ce cas là, l'intérêt peut porter non seulement sur chaque caractère, mais également sur les liens qui peuvent exister entre les variables. En fait, certaines études visent à étudier conjointement deux variables mesurés sur un même individu. Il s'agit par exemple de la répartition des employés d'une entreprise selon leurs salaires et ancienneté, ou bien la répartition des ménages selon le nombre d'enfants à charge et leurs revenus. Nous nous intéresserons donc dans ce livre à l'étude d'une distribution statistique à deux dimensions.

Dans ce livre nous focalisant notre intérêt sur deux points particuliers. Le premier est la présentation des outils d'analyse bivariée, le but étant d'initier les étudiants à l'élaboration d'une étude utile pour l'exploitation des relations entre deux variables. Le second point sera une introduction à l'étude des séries chronologiques pour la détermination du cycle économique.

Chapitre1 : Initiation à l'utilisation de R

R est un logiciel libre distribué par "GNU Public Licence" (qui est une licence qui fixe les conditions légales de distribution des logiciels libres du projet GNU) et dérive du langage S (le logiciel S-PLUS). Il présente des caractéristiques remarquables comme la possibilité d'effectuer du calcul matriciel et d'autres opérations complexes, du stockage et de manipulation des données. Il contient des nombreuses fonctions pour les analyses statistiques et des outils graphiques flexibles. Il s'adresse à un large public formé de spécialistes et non-spécialistes en informatique.

Il a été initialement créé, en 1996, par Robert Gentleman et Ross Ihaka du département de statistique de l'Université d' Auckland en Nouvelle Zélande. Depuis 1997, il s'est formé une équipe "R Core Team" qui développe R. Il est conçu pour pouvoir être utilisé avec les systèmes d'exploitation Unix, Linux, Windows et MacOS.

R est devenu un logiciel libre et gratuit en 1995. R est à la fois un langage de programmation et un progiciel de fonctions statistiques. La version de base de R contient déjà un grand nombre de fonctions statistiques et graphiques permettant, par exemple, de calculer une moyenne ou une variance ou de tracer un histogramme.
Le logiciel R disponible gratuitement sur le site CRAN, à l'adresse suivante : cran.r-project.org. Ce logiciel est comparable à Matlab à certains égards, mais se révèle beaucoup plus puissant dans le domaine des traitements statistiques. De nombreux chercheurs ont développé au cours des années des fonctions plus avancées qui sont disponibles à tous les utilisateurs de R. Ces fonctions sont regroupées en bibliothèques qui sont disponibles pour téléchargement sur le site du projet R : http ://www.r-project.org/.

Le " Comprehensive R Archive Network" (CRAN Réseau d'archives de R globales) est un ensemble de sites qui fournit ce qui est nécessaire à la distribution de R, ses extensions sa documentation, ses fichier sources et ses fichiers binaires. Le site maître du CRAN est situé en Autriche à vienne, nous pouvons y accéder par L'URL : http ://cran.r-project.org/

1 Installation de logiciel R

Le logiciel est télécharger sur le site web officiel de R (http ://www.r-project.org/), il faut ensuite se diriger dans download CRAN. Choisissez un site miroir proche de chez vous (par exemple : en France), les téléchargements seront probablement plus rapides, vous trouvez ensuite un encadrement légendé download and install R.
Pour Windows, cliquez sur Windows puis base. Cliquez ensuite sur le fichier "exist" (par exemple : R-2.14.1-win.exe). Le programme d'installation est alors téléchargé sur votre ordinateur. Il suffit de cliquer dessus et de suivre les instructions. Un dossier portant le nom de la version de R télécharger (R-2.14.1-win.exe dans ce cas) est crée. Il est situé, à partir du disque dur C, dans la série de dossiers suivante : program files /R.
Dans ce dossier se trouve le dossier library qui comprend les packages base de R. Un autre élément utile doit être localisé : le fichier .R data. Celui-ci n'est pas apparent au début. Il contiendra tous les objets que vous créerez et sauvegardez dans R. Sur mon ordinateur, il apparaît, à partir du disque dur C.

2 Création et manipulation des données

On peut interagir directement avec R en ligne de commande, ou bien éditer un fichier texte (avec l'extension .R) sous son éditeur préfère.
Les données manipulées sous R sont stockées dans un espace de travail. Le répertoire de travail actuel peut être visualise en tapant getwd(), tandis que l'accès a des répertoires se fait grace à la commande setwd().
On peut à n'importe quel moment de la session visualiser le contenu de cet espace de travail a l'aide de la commande ls(), et supprimer une variable à l'aide de la commande rm(), en lui passant comme argument le nom de la variable. Pour supprimer l'ensemble des variables contenues dans l'espace de travail, on utilise la commande: rm(list=ls())
Pour quitter l'environnement, il suffit de taper q(). Comme on vient de le dire, le travail sous R s'effectue via une session, que l'on peut à tout moment sauver à l'aide de la commande save.image().

En tapant ls(), on constatera que nos variables sont toujours présentes dans l'espace de travail (celui-ci a en fait été sauve dans un fichier .RData, dans le répertoire de travail). La commande history() permet de lister les dernières commandes. Enfin, on peut sauver toutes les commandes ayant servi à analyser des données dans un fichier script avec l'extension .R, et taper source ('script.R',echo=T), depuis l'invite de commande, pour effectuer l'analyse.
R est un langage en ligne de commande (interprète), comme Matlab, et peut manipuler de nombreux types de données à l'aide de commandes prédéfinies. On va s'intéresser aux principaux types d'objets, à savoir :

-les vecteurs
-les matrices
-les facteurs
-les listes d'objets
-les dataframe

et voir quels sont les moyens de créer, manipuler et traiter de tels objets.
Notons que, comme dans tout langage, certains mots-clés sont réserves et ne peuvent être utilises comme noms de variable ou de fonction :

FALSE Inf NA NaN NULL TRUE
break else for function if in next repeat while

2.1 Les vecteurs

Le vecteur est le type de base de R. Un nombre est simplement un vecteur à un seul élément. Notons dans un premier temps que R peut être utilise comme un simple calculateur :

```
2+2
[1]4
2* pi
[1]6.283185
```

Certaines constantes (e.g. pi) et fonctions (exp, cos, etc.) sont connues de R et peuvent être appelées directement. On peut également assigner une valeur à une variable à l'aide de l'opérateur← (ou ¿) :

```
x ← 10
```

La variable peut ensuite être utilisée comme dans n'importe quel langage interprété :

```
x+2
[1]12
y ← x+4
y
[1]14
```

Bien que nous n'ayons utilisé que des variables de type numérique, il existe d'autres modes de représentation des données (ou classes) dans un vecteur :

character, integer, logical, complex, list.
Enfin, les types NULL et NA jouent un rôle particulier puisqu'ils désignent respectivement l'absence de valeur (i.e. un ensemble vide) et le codage par défaut d'une valeur manquante.
Différentes fonctions permettent de générer facilement des vecteurs :c, seq, rep. Par exemple, on peut créer le vecteur X = [1 2 3 4 5] de différentes façons :

```
X ← c(1,2,3,4,5)
X
[1]1 2 3 4 5
rm(X)
```

La fonction seq() est utile pour générer des suites d'entiers, et possède plusieurs arguments pour spécifier la séquence recherchée (voir ?seq).

```
Y ← seq(1:5)
Y
[1]1 2 3 4 5
```

On peut dupliquer un vecteur à l'aide de la commande rep() :

```
Z ← rep(X,2)
Z
[1]1 2 3 4 5 1 2 3 4 5
```

La fonction rev() permet quant a elle de renverser l'ordre de la séquence (1 2 3 devient 3 2 1).

```
W ← rev(X)
W
[1]5 4 3 2 1
```

Les opérations arithmétiques s'appliquent aussi sur des vecteurs. Par exemple, on peut additionner deux vecteurs a et b, et effectuer les produits scalaires ou membre a membre classiques :

```
a ← 1:5
b ← 5:9
a × 2
[1]2 4 6 8 10
a+b
[1]6 8 10 12 14
a × b
[1]5 12 21 32 45
```

De nombreuses fonctions permettent également de classer les éléments (ou les indices) d'un vecteur, de les sommer, etc.

```
u ← c(3,2,7,1,4)
u
[1]3 2 7 1 4
sort(u) # ordonner les éléments du vecteur u
[1]1 2 3 4 7
```

order(*u*) # donner le rang des éléments du vecteur u

[1]1 3 2 5 4

sum(u) # donner la somme des éléments du vecteur u

[1]17

length(u) donner la longueur du vecteur u

[1]5

Enfin, une particularité de R est que les éléments d'un vecteur peuvent avoir des noms. La fonction names() permet en effet d'associer une étiquette à chacun des éléments d'un vecteur :

x ← 1:5

names(*x*) ← *c*(a , b , c , d , e)

x

a b c d e

1 2 3 4 5

2.2 Les matrices

Pour les matrices, on peut soit générer un vecteur de taille n, et le réarranger pour créer une matrice de taille (l,m) , soit directement créer une matrice à l'aide de la fonction matrix(), en spécifiant en arguments le nombre de lignes et de colonnes, par exemple :

M ← ¿*c*(1,2,3,4,5,6,7,8,9,10),*nrow*=2,*ncol*=5¿

	[,1]	[,2]	[,3]	[,4]	[,5]
x	1.00	3.00	5.00	7.00	9.00
y	2.00	4.00	6.00	8.00	10.00

Les fonctions cbind et rbind permettent de manipuler des vecteurs de manière à former une matrice par concaténation sur les colonnes ou sur les lignes.

x ← *c*(1,3,5,7,9)

y ← *c*¿

A ← *cbind*(*x*,*y*)

	x	y
[1]	1.00	2.00
[2]	3.00	4.00
[3]	5.00	6.00
[4]	7.00	8.00
[5]	9.00	10.00