

# **Hands-on Course in Programming and Statistical Data Analysis with R**

## **R: User's Concise Manual; Programming and Data Analysis**

**By : M.A.B. Wolanowski**

## Préface – version française

Les meilleures choses dans la vie sont gratuites. Le système informatique R est un excellent outil gratuit pour analyse statistique de données. L'objectif de ce papier n'est pas de fournir une vue compréhensive de R. Néanmoins, il peut servir à résoudre une gamme extensive de problèmes.

Le traitement et l'éventail de thèmes abordés distinguent ce livre des autres publications concernant les domaines similaires. Ce livre n'a pas la forme descriptive adoptée conventionnellement dans les textes statistiques. Le **style concis** est une caractéristique importante de ce livre. L'auteur s'est efforcé à garder informations essentielles et sauter détails qui ne sont pas indispensables. Cela doit faciliter le processus d'apprentissage ainsi que recherche d'information.

On a exposé brièvement la base théorique des sujets discutés. Cependant, on a essayé de ne pas s'enfoncer dans les détails théoriques. Le livre est sensé d'être étudiée avec, **en même temps, taper et exécuter les commandes R discutées.**

Grace à la forme austère, le livre a environ 120 pages et décrit beaucoup de méthodes d'analyse statistique de données. En plus, il contient exemples et information sur écriture de logiciels, construction de graphiques et manipulation des dossiers informatiques. Le style concis et parcimonieux doit plaire aux lecteurs qui pensent que présentation descriptive fait obstacle à compréhension des aspects essentiels.

La première partie du livre facilite compréhension de la logique des logiciels R inclus dans la deuxième partie. En plus, connaissance de la première partie permet de se servir de R comme langage universel de programmation informatique. Cela peut attirer utilisateurs des ordinateurs portables.

Le livre peut servir de guide pour résoudre problèmes concernant analyse statistique de données. Il est destiné aux gens qui veulent soit apprendre R soit s'inspirer par les exemples présentés dans traitement de leurs problèmes. En plus des exemples nombreux, le texte contient définitions des fonctions ou procédures utiles, e.g.:

- (a) Fonction "expand\_wt" de paragraphe 3.6.2 élimine variable de poids de tableau de données.
- (b) Fonction "dist\_km" de paragraphe 3.6.3 rend la distance sphérique entre points géographiques décrits par leurs latitude et longitude.
- (c) Procédure de paragraphe 9.2 trouve la matrice de scores normaux à partir de données qui contiennent valeurs de variables discrètes.
- (d) Fonction "longf" de paragraphe 14.1 transforme "data frame" avec mesures répétées en forme longue.
- (e) Conversions de similitudes en distances et vice versa sont présentées aux paragraphes 11.1 et 12.1.
- (f) Distances entre vecteurs zéro-un sont discutées aux paragraphes 11.1 et 12.5. La fonction "mat\_sim" rend la matrice de similitudes entre objets représentés par vecteurs zéro-un. Leurs coordonnées indiquent présence ou absence de caractéristiques spécifiées.

Les étudiants qui font des études de mathématique ou informatique et qui s'intéressent aux modèles statistiques peuvent se servir de ce livre comme un texte autonome pour apprendre R. Ceux qui ne sont pas habitués à présentation mathématique de matériaux peuvent profiter du livre conjointement avec un papier plus conventionnel. Chaque utilisateur du système R peut utiliser le livre comme support pour: trouver méthodes statistiques appropriées et écrire programmes informatiques afin de résoudre problèmes spécifiques.

La forme compacte du livre justifie un cours universitaire en statistiques d'au moins un an comme condition préalable. On n'exige pas des connaissances spécifiques en informatique. Cependant, connaissance d'un langage informatique serait avantageuse et on en recommande fortement. Le langage C++ est préféré.

Le livre doit être accessible aux lecteurs qui connaissent les fondements de statistiques et d'informatique. La base nécessaire en statistiques (et plus) serait fournie par chacune des publications:

(a) Ronald E. Walpole and Raymond H. Myers, "Probability and Statistics for Engineers and Scientists", Macmillan Company, 1985.

(b) William Mendenhall and James E. Reinmuth, "Statistics for management and Economics", Wadsworth Publishing Company, 1978.

L'auteur essaie de montrer comment utiliser R pour résoudre problèmes de types différents. Cet objectif est compatible avec inclusion des exemples basés sur données artificielles. Naturellement, les exemples présentés n'appartiennent tous à cette catégorie.

***On s'attend que les lecteurs tapent et exécutent les commandes R discutées.*** Par conséquent, graphiques et certains autres résultats sortis d'ordinateur ne paraissent pas dans le texte. Naturellement, il y a des commentaires sur ces résultats. On utilise base de données de petites tailles. Donc, on peut les saisir à l'ordinateur manuellement, sans accéder à source extérieure. Les lecteurs doivent confronter les résultats observés sur le moniteur de l'ordinateur avec les remarques dans le livre.

***Reproduction et exécution de programmes R contenus dans le texte faciliteront compréhension de thèmes considérés et joueront le rôle d'exercices.*** On juge cela nécessaire pour novice qui souhaite d'apprendre R de ce livre. Les exemples fournis dans le texte illustrent les applications majeures de méthodes présentées. Naturellement, lecteurs peuvent utiliser ces techniques aussi dans les exercices de nombreux manuels en statistiques.

Les méthodes quantitatives sont entrées dans tous les secteurs qui embauchent les diplômés des universités et grands écoles. Les étudiants et employés à chaque niveau souhaitent d'acquérir ou d'améliorer leurs compétences en application de méthodes statistiques. Cette assertion se rapporte à tous les secteurs professionnels, y compris: industrie, administration publique, santé, éducation, etc.

La plupart de gens considérés s'intéresse aux techniques statistiques bien établis. Une proportion substantielle de ces individus n'a pas expérience forte dans le domaine de statistique ou programmes informatiques. Il est probable que ces personnes chercheront un livre ayant les caractéristiques:

- (a) Il est compact et on peut assimiler rapidement ces contenus; on peut retrouver facilement détails spécifiques.
- (b) Il explique comment appliquer méthodes de Gestion de Données ainsi que celles d'Analyse Statistique.
- (c) Le texte ne ressemble pas au livre de cuisine et il y a description de logique à la base de méthodes discutées.
- (d) Il traite une vaste gamme de méthodes statistiques.
- (e) On peut l'étudier sans avoir base solide en statistique ou informatique.
- (f) Il explique comment mettre en œuvre les méthodes présentées avec système informatique disponible facilement et gratuit.
- (g) Il contient exemples qui montrent à lecteur comment résoudre ses problèmes concernant analyse de données.

Ce livre possède tous ces particularités. En plus, il explique comment écrire programmes en R et il permet d'utiliser R comme langage universel de programmation informatique.

Pendant plusieurs ans l'auteur de ce livre employait le système SAS dans projets concernant analyse de donnée et construction de modèles statistiques. Un moment est arrivé quand SAS n'était plus disponible pour lui. A ce point, il a décidé de basculer vers R. Ces notes reflètent le processus de son apprentissage. Le livre est inspiré par les publications:

- (i) Documentation of CRAN.R project.
- (ii) P A Cornillon et al., "Statistiques avec R", Presses Universitaires de Rennes, 2008.
- (iii) A.J.Dobson, "An Introduction to Statistical Modelling", Chapman and Hall, 1983.
- (iv) B.S.Everitt and T.Hothorn, "A Handbook of Statistical Analysis using R", Taylor & Francis Group, 2010.
- (v) K.G.Joreskog and D.Sorbom, "Lisrel 8 User's Reference Guide", SSI, 1996 2001.
- (vi) SAS Institute Inc., "Categorical Data Analysis, Course Notes", SAS Institute Inc., 1991.

Bien que ce livre soit autonome, on recommande fortement aux lecteurs de se familiariser avec le papier (iv) ci-dessus. Cela leur donnera une compréhension plus fine de méthodes discutées et de leurs applications. Le livre de B.S.Everitt and T.Hothorn contient nombreux exemples réaliste utilisant larges bases de données. La référence (ii) a une description instructive de techniques présentés dans paragraphes 1.4 (Tables) et 1.5 (Cutting a continuous variable).

**AVERTISSEMENT: La lecture de ce livre sans taper et exécuter en même temps les commandes R discutée serait désagréable et décourageant.**

\*\*\*-----\*\*\*

## Preface

The best things in life are free. The R system is an excellent free tool for statistical analysis of data. The objective of these notes is not to give a comprehensive view of R. However, familiarity with the presented text will enable to solve a vast range of problems.

The treatment and the range of discussed issues distinguish this paper from other publications concerning similar areas. The book does not have conventional descriptive form adopted in statistical papers. The **concise style** is an important feature of the book. Depicting any topic, the author tried to retain essential information and to skip details which are not indispensable. This should facilitate both the learning process and information retrieval.

The theoretical background underlying the discussed topics is not neglected. However, an effort was made to keep theoretical considerations to minimum. The book is meant to be studied with **simultaneously typing and executing the discussed commands of R**.

Due to the austere form, the whole book contains just over 120 pages and covers a wide range of statistical data analysis techniques. Furthermore, it provides information on writing computer programs, producing graphs and performing file operations. The concise and parsimonious style should appeal to the readers who are of the opinion that descriptive presentation impedes apprehension of the essential aspects.

The first part of the book facilitates understanding the logic of the R programs included in the second part. Moreover, familiarity with the first part will allow utilising R as a general purpose programming language. This is likely to be found attractive by the personal computer users.

The text may serve as a guide in solving various problems from the area of statistical data analysis. The book is addressed to people who wish either to learn the R system or to be inspired by the presented examples in solving their own statistical problems. In addition to numerous examples, the text contains definitions of some useful utility functions or procedures, e.g.:

- (a) Function "expand\_wt" in paragraph 3.6.2 eliminates weight variable from data table.
- (b) Function "dist\_km" in paragraph 3.6.3 returns spherical distance between two geographical points given by their latitude and longitude.
- (c) Procedure of paragraph 9.2 determines matrix of normal scores from data involving values of discrete variables.
- (d) Function "longf" in paragraph 14.1 transforms data frame containing repeated measures into long form data frame.
- (e) Conversions of similarities to distances and vice versa are presented in paragraphs 11.1 and 12.1.
- (f) Distances between zero\_one vectors are discussed in paragraphs 11.1 and 12.5. The function "mat\_sim" returns a matrix of similarities between items represented by zero-one vectors. Their coordinates indicate presence or absence of some characteristics.

The students majoring in mathematics or computer science and interested in statistical models can learn R using the book as an autonomous text. Those not accustomed to compact mathematical presentation of material may profit from the book as a companion to a more conventional paper. Any user of the R system can employ the book as aid in both: choosing appropriate statistical methods and writing R programs for solving specific problems.

The compact form of the book justifies at least one year university course in statistics as prerequisite. No specific knowledge of computer programming is assumed. However some familiarity with any programming language would be advantageous and is strongly recommended. The language C++ is preferred.

The book should be accessible to readers having some basic familiarity with statistics and computer programming. The necessary statistical background (and more) would be provided by anyone of the publications:

(a) Ronald E. Walpole and Raymond H. Myers, "Probability and Statistics for Engineers and Scientists", Macmillan Company, 1985.

(b) William Mendenhall and James E. Reinmuth, "Statistics for management and Economics", Wadsworth Publishing Company, 1978.

The author attempts to show how various types of problems can be solved using R. This objective is compatible with inclusion of examples based on artificial data. Naturally, not all presented examples belong to this category.

***The readers are expected to key in and to execute the discussed commands.*** Consequently, graphics and some other computer output are not included in the text. Naturally, there are brief comments concerning the results printed by computer. An effort was made to use small data bases. Then, data can be easily reproduced by typing, with no need to access an external source. The readers should confront results observed on the computer screen with the comments provided in the book.

***Copying and executing the R programs included in the text will facilitate comprehension of the discussed topics and will play the role of exercises.*** This is deemed essential for a novice who wishes to learn R from this book. The examples provided in the text illustrate the major applications of the presented methods. Naturally, readers may use these techniques also in exercises from numerous statistics textbooks.

The quantitative methods have entered every sector which provides employment to graduates of the tertiary education institutions. The students and employees at any level wish to acquire or to improve their skills in application of statistical methods. This statement pertains to all professional sectors, including: management, economics, finance, industry, public administration, health, education, etc.

Most of people under consideration are interested in well-established statistical techniques. A significant proportion of these individuals have a limited experience in the area of statistics or computer programming. They are likely to look for a book having the following features:

(a) It is compact and its contents can be quickly digested; any specific details can be easily retrieved.

- (b) It explains how to apply methods of both, Data Manipulation and Statistical Analysis.
- (c) The text does not resemble a cookery book and an outline of logic underlying discussed techniques is provided.
- (d) It covers an extensive range of statistical methods.
- (e) It does not require a solid background in statistics or computer science.
- (f) It explains how to implement presented methods using software which is readily available at no cost. This software can be downloaded from internet and installed easily in a personal computer.
- (g) It contains examples to guide a reader in solving his own problems concerning data analysis.

This book has all these characteristics. In addition, it explains how to write R programs and it enables using R as a general purpose programming language.

During a number of years, the author of these notes employed the SAS system to perform tasks from the area of modelling and statistical analysis of data. At some point, SAS was no longer available to him. Then, he decided to switch to R. These notes reflect his learning process. The book was inspired by the following publications:

- (i) Documentation of CRAN.R project.
- (ii) P A Cornillon et al., "Statistiques avec R", Presses Universitaires de Rennes, 2008.
- (iii) A.J.Dobson, "An Introduction to Statistical Modelling", Chapman and Hall, 1983.
- (iv) B.S.Everitt and T.Hothorn, "A Handbook of Statistical Analysis using R", Taylor & Francis Group, 2010.
- (v) K.G.Joreskog and D.Sorbom, "Lisrel 8 User's Reference Guide", SSI, 1996 2001.
- (vi) SAS Institute Inc., "Categorical Data Analysis, Course Notes", SAS Institute Inc., 1991.

Although this book is self-contained, the readers are strongly advised to familiarize themselves with the above paper (iv). It will give them a deeper insight into the discussed methods and their applications. The book by B.S.Everitt and T.Hothorn contains a number of realistic examples involving huge data bases. The reference (ii) has an instructive description of techniques used in paragraphs 1.4 (Tables) and 1.5 (Cutting a continuous variable).

**WARNING: *Reading this book without simultaneously typing and executing the discussed R commands would be unpleasant and disheartening.***

---

# CONTENTS

## Part I: Data Manipulation

### 1 Introduction

- 1.1 General information
  - 1.1.1 Internet
  - 1.1.2 R Console
  - 1.1.3 Getting help
  - 1.1.4 Divers
  - 1.1.5 Special values
  - 1.1.6 Organizing an analysis
- 1.2 Objects in R
  - 1.2.1 Vectors
  - 1.2.2 Matrices
- 1.3 Factors and lists
  - 1.3.1 Factors
  - 1.3.2 Lists
- 1.4 Tables
- 1.5 Cutting a continuous variable
- 1.6 Selection of elements from vectors or matrices
  - 1.6.1 Function **which**
  - 1.6.2 Vectors
  - 1.6.3 Matrices
- 1.7 Programming in R
  - 1.7.1 Statements
  - 1.7.2 Operators
  - 1.7.3 Loops

### 2 Reading, Exporting and Saving data files

- 2.1 Reading data
- 2.2 Description of the created object
- 2.3 Function **read.table**
- 2.4 Reading data in FORTRAN format
- 2.5 Exporting and saving data files

### 3 Functions

- 3.1 Arguments by default
- 3.2 User defined functions
- 3.3 Functions of the apply family
  - 3.3.1 Function **apply**
  - 3.3.2 Function **sweep**
  - 3.3.3 Functions **lapply**, **sapply** and **tapply**
  - 3.3.4 Function **aggregate**
  - 3.3.5 Function **by**
- 3.4 Miscellaneous functions
- 3.5 Statistical functions
  - 3.5.1 Descriptive statistics



- 3.5.2 Probability distributions
    - 3.5.3 Samples
  - 3.6 Examples
    - 3.6.1 Ranks
    - 3.6.2 Eliminating weight variable from data frame
    - 3.6.3 Geographic distances
- 4 Overview of R objects
  - 4.1 Data overview
    - 4.1.1 Economic sectors
    - 4.1.2 Groups of observations
  - 4.2 Graphics
    - 4.2.1 Scatterplots
    - 4.2.2 Histograms
    - 4.2.3 Boxplots
    - 4.2.4 Several graph windows
  - 4.3 Estimating density functions
    - 4.3.1 R function **density**
    - 4.3.2 Nonparametric kernel density estimators

## Part II: Data Analysis

- 5 Elementary Inferences
  - 5.1 Student's t Test
    - 5.1.1 Tests about population mean
    - 5.1.2 Independent samples t test
    - 5.1.3 Paired t test
  - 5.2 Non parametric Tests
    - 5.2.1 Wilcoxon Mann Whitney Rank Sum Test
    - 5.2.2 Wilcoxon Signed Rank Test
  - 5.3 Independence in Contingency Tables
  - 5.4 McNemar's Test
  - 5.5 Correlation
    - 5.5.1 Function **cor.test**
    - 5.5.2 Pearson's Correlation
  - 5.6 Comparison of Variances
- 6 Linear Models
  - 6.1 Model formulae
  - 6.2 Generalized linear models
  - 6.3 Maximum likelihood estimation
  - 6.4 Function **glm**
  - 6.5 Linear Regression
    - 6.5.1 Regression with **glm**
    - 6.5.2 The **lm** function
- 7 Binomial and Poisson models
  - 7.1 Binomial models
  - 7.2 Logistic regression
  - 7.3 Poisson models

## 8 ANOVA and MANOVA

### 8.1 Model outline

### 8.2 Analysis with R

## 9 Principal Component Analysis (PCA)

### 9.1 Outline

### 9.2 Normal scores

## 10 Exploratory Factor Analysis

### 10.1 Background

#### 10.1.1 Orthogonal factors

#### 10.1.2 Factor rotation

#### 10.1.3 Factor scores

#### 10.1.4 ML estimation

### 10.2 Function **factanal**

## 11 Cluster Analysis

### 11.1 Background

### 11.2 Hierarchical clusters

### 11.3 k means clustering

### 11.4 Clustering with R

#### 11.4.1 Hierarchical clusters in R

#### 11.4.2 k\_means clusters in R

## 12 Multidimensional Scaling (MDS)

### 12.1 Introduction

### 12.2 Classical (Metric) MDS

### 12.3 Non metric (Ordinal) MDS

### 12.4 MDS in R

#### 12.4.1 Classical MDS in R

#### 12.4.2 Ordinal MDS in R

### 12.5 Distances between zero\_one vectors

## 13 Log Linear Models

### 13.1 Background

### 13.2 Function **loglin**

## 14 Random effects models

### 14.1 Repeated measures

### 14.2 Linear mixed effects model (lme)

#### 14.2.1 Model description

#### 14.2.2 Longitudinal data

### 14.3 Binomial response variable

## Part 1: Data Manipulation

### 1 Introduction

This chapter contains a list of several internet sites. They can be used to download R or its modules and to obtain information on various aspects of the system. Some examples of R commands to get help are given. Vectors, matrices and other R objects are defined. The discussion of operations involving these objects includes solving a system of linear equations, converting a continuous variable into an ordered class variable, building a contingency table, etc. The assignment, relational, arithmetic and logical operators are described. Conditional commands, loops and other elements of R programming are presented. An example demonstrates how to convert frequency matrix to data frame and to contingency table.

#### 1.1 General information

##### 1.1.1 Internet

To load and install R for Windows, go to the site: <http://cran.r-project.org> and click the relevant option. The R FAQ (Frequently Asked Questions) can be found at

<http://CRAN.R-project.org/doc/FAQ/R-FAQ.html>.

The basic R module can be downloaded and installed from the above site.

The HSAUR (Handbook of Statistical Analysis Using R; B.S.Everitt and T.Hothorn, 2010) gives the following Internet addresses:

<http://CRAN.R-project.org/bin/windows/base/release.htm>.

To download the installer file "rw2090.exe", activate the site and follow the instructions.

<http://CRAN.R-project.org/web/packages/>

This address contains a list of available packages

<http://CRAN.R-project.org/web/views/>

To attain the package "**sem**" (Structural Equation Models), go to this site, then, click "SocialSciences".

<http://CRAN.R-project.org/manuals.html>

A collection of manuals can be found at this address.

##### 1.1.2 R Console

The R Console is the screen obtained after having launched the R system. The symbol ">" is called "prompt". It indicates that the R system is ready to process user's command.

The R program is installed with some high priority packages, including: base, cluster, datasets, graphics, lattice, MASS, mgcv, nlme, spatial, stats, stats4, etc. The following commands show how to install and use packages.

```
install.packages("HSAUR2")
```

"HSAUR2" is the package name; internet connection is needed; then, you may choose: USA  
CA1

```
install.packages("sandwich")
```

This command will download and install the package "sandwich".

```
library("sandwich")
```

The library statement calls a specified installed package and allows it to be used in R  
program.

```
vignette("sandwich", package="sandwich")
```

Vignette is a document describing package.

```
library()
```

Execute this statement to print a list of all packages stored in the directory "... /R/R-  
211~1.1/library"

### 1.1.3 Getting help

The manuals, notes entitled "An Introduction to R", etc. are accessible by typing on the  
console: **help.start()** and pressing the key [Enter].

#### Example 1:

- (i) Execute the command **help.start()**. Then, click successively: "R Data  
Import/Export"; "Spreadsheet like data"; "Fixed width format files". In the  
displayed text find information on functions **read.fortran** and **read.fwf**

Execute the commands:

```
help(package="HSAUR2")  
# get overview of the package "HSAUR2"; it must be already installed.  
help("mean")      # get help for the word in quotes  
?mean             # the same action as above  
help("read.fortran") # get help for the word in quotes  
?read.table       # get help for the specified function  
?lme4::lmer        # get help for the function lmer from package lme4.  
?remove.packages  # get help on the indicated function.
```

### 1.1.4 Divers

In this book, the R function names are written in bold.

" <- " is the assignment operator

" # " indicates that all characters between this sign and the end of line constitute a

comment and should be ignored by the compiler.

```
options(prompt="R> ")      # Change the prompt to "R> ".
x<- sqrt(25) + 2; print(x)  # Evaluate the expression, store result in x; print x.
x; sqrt(25) + 2             # Print x; then, print the expression value.
ls()                        # List names of all objects; x will be one of them.
rm( x )                    # Remove object x.
rm( list = ls() )          # Remove (almost) everything in the working environment.
q()                         # Quit the R program; the system will ask whether to save the R workspace.
```

The normal distribution with mean  $\mu$  and variance  $\sigma^2$  is denoted by  $N(\mu, \sigma^2)$ . The multivariate normal distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$  is denoted by  $N(\mu, \Sigma)$ .

### 1.1.5 Special values

If the R system encounters problems in evaluating an expression, it may print an error message or return special value: Inf (Infinity) or NaN (Not a Number). The statements: **log(0)** and **log(-1)** print -Inf and NaN, respectively. The value Inf can be used as function argument. The following calls of the standard normal density function return the value zero: **dnorm(Inf)**; **dnorm(-Inf)**.

In R, the symbol NA (Not Available) denotes a missing value. An arithmetic expression containing NA has the value NA. For instance, NA is printed by the statements: **x<- NA**; **print(x + 5)**. Some functions have option "na.rm=". If na.rm=TRUE, observations or elements containing NA are omitted from the calculations performed by the function. The statement: **x<- c( 2, NA, 3, 7)** creates a vector x with missing second coordinate. Then, **mean(x)** returns NA and **mean(x, na.rm=T)** returns the mean of the non missing elements

A reserved word, NULL represents the null object in R. NULL is often returned by expressions and functions whose value is undefined.

### 1.1.6 Organising an analysis

It is recommended to create a separate directory for every R project. The images of sessions concerning the project will be saved in that directory. When the directory exists, open R console, in menu Fichier (File) click "Changer (Change) le repertoire courant ..." and select the required directory. Then, execute at least one command followed by the command **q()** and in the dialog box click "Save the session image". Two icons will appear in the directory window: icon "R workspace" and icon "history".

It may be comfortable to create a text file containing all steps to perform certain task. Any text editor can be used. A text file called "task1.rtf" can be invoked to the R Console by the command:

```
source("task1.rtf", echo=TRUE)
```

The R prompt '>' or '+' at the beginning of line should not appear in the command file. In this way, the analysis can be reproduced, maybe with updated data. The above command applies when the source file is in the R working directory; otherwise, the full path of "task1.rtf" is required.

## 1.2 Objects in R

The R language uses the notion of a class. It is defined in the C language. The class is a generalisation of the concept of user defined type. It consists of data and functions used to access these data.

<b>class(x)</b>	# returns the class of an object x.
<b>mode(x)</b>	# returns the mode of an object.

Examples of a class are: integer, matrix, list, etc. A mode can be: NULL, logical (binary), character (string of characters), numeric, complex, etc.

Objects are class entities, e.g. integer, matrix, etc. A list of object attributes is given by the command: **attributes(object)**.

### 1.2.1 Vectors

A vector is composed of coordinates (components). They have all the same mode. The **length(x)** returns the attribute "length" of vector x, i.e. the number of its coordinates. Vector is an elementary data structure in R; for instance, the number 5 is a vector of length 1. The following statements return the indicated numeric vectors:

5:7	# vector: 5, 6, 7;
<b>c</b> (5, -3, 2)	# vector: 5, -3, 2;
<b>seq</b> (2, 4, by=0.5)	# vector: 2.0, 2.5, 3.0, 3.5, 4.0;
<b>seq</b> (1, 2, length=5)	# vector: 1.00, 1.25, 1.50, 1.75, 2.00;
<b>rep</b> (1, 3)	# vector: 1, 1, 1;
<b>rep</b> ( <b>c</b> (1, 2), each=3)	# vector: 1, 1, 1, 2, 2, 2;
<b>rep</b> ( <b>c</b> (1, 2), 3)	# vector: 1, 2, 1, 2, 1, 2;

The statements below create vectors of characters:

<b>c</b> ("A", "BC", "B")	# vector: "A", "BC", "B";
<b>rep</b> ("A", 3)	# vector: "A", "A", "A";
<b>paste</b> ("A", 1:3, sep="_")	# vector: "A_1", "A_2", "A_3";
<b>paste</b> ("A", 1:3, sep="")	# vector: "A1", "A2", "A3";
<b>paste</b> ("x", v, sep="")	# attaches to x coordinates of a numerical vector v;
<b>paste</b> ("x", 1:3, collapse="+")	# vector with 1 component: "x1+x2+x3";
<b>substr</b> ("ABCDEF", 2:5)	# vector: "BCDE", substring.

The function **sort** arranges entries of a vector in increasing order; the function **rev** reverses the order:

<b>v</b> <- <b>sort</b> ( <b>v</b> )	# increasing order
<b>v</b> <- <b>sort</b> ( <b>v</b> , decreasing=TRUE)	# decreasing order
<b>v</b> <- <b>rev</b> ( <b>sort</b> ( <b>v</b> ))	# decreasing order

Suppose that **v** is a numerical vector and **logiv** is a logical vector, i.e. the components of **logiv** are either TRUE or FALSE. If **v** and **logiv** are of the same length, **v[logiv]** is a vector

containing the coordinates of  $v$  such that the corresponding coordinates of  $\log v$  have the value TRUE.

For instance, the expression " $v > 5$ " designates a logical vector of the same length as  $v$ ; its coordinate has the value TRUE if the corresponding coordinate of  $v$  is greater than five. The command: `v[ v > 5 ]` prints a vector containing all the coordinates of  $v$ , which are greater than five. The assignment: `v[ v == 0 ] <- 1e-10` replaces zero coordinates of  $v$  by  $1e-10$ .

### 1.2.2 Matrices

Matrix is a table with elements having all the same mode. If  $m$  is a matrix, the **length**( $m$ ) returns the total number of its elements. The **dim**( $m$ ) gives the dimensions of matrix  $m$ ; for a 2 dimensional matrix, it is the number of rows and the number of columns. The names of dimensions constitute an optional attribute. It is returned by the function **dimnames**. The **nrow**( $m$ ) and **ncol**( $m$ ) give the number of rows and columns, respectively. A matrix can be created by the function **matrix**. The function **array** is used to create tables of dimension greater than 2.

```
m<- matrix(1:6, nrow=2, ncol=3)
      # the values 1, 2, ..., 6 are entered column by column.
rnames<- c("row1", "row2")
      # the function c concatenates specified elements into a vector.
cnames<- c("col1", "col2", "col3")
```

```
dimnames(x)<- list(rnames, cnames) # The function list is discussed below.
```

If  $m$  is a matrix, the  $m[, k]$  designates its  $k$ 'th column. It is a vector. In R, a distinction is made between a vector and a matrix. A vector  $v$  of length  $n$  can be transformed to  $(n \times 1)$  matrix  $m$  by the assignment:

```
m<- as.matrix(v)
```

Then,  $m[,1]$  is a vector.

The functions **rownames** and **colnames** give the names of rows and columns of a matrix. The **rownames** and **colnames** can be used on the left and on the right side of the assignment operator.

For a numeric matrix  $m$ , **sin**( $m$ ) and **exp**( $m$ ) apply the functions **sin** and **exp** element by element;  $m^2$  returns a matrix whose entries are squares of the corresponding entries of  $m$ . For matrices  $A$ ,  $B$  and a vector  $v$ :

<b>t</b> ( $A$ )	returns	transpose of $A$
<b>det</b> ( $A$ )	returns	determinant of $A$
<b>eigen</b> ( $A$ )	returns	eigenvalues and eigenvectors of $A$
<b>solve</b> ( $A$ )	returns	inverse of $A$
<b>solve</b> ( $A, v$ )	returns	solution, $x$ , of $Ax = v$
<b>diag</b> ( $v$ )	returns	diagonal matrix with coordinates of $v$ on the diagonal
$A + B$	is	sum of $A$ and $B$
$A \%*\% t(B)$	is	matrix multiplication; rows of $A$ by columns of $B$