Università
Ca' Foscari
Venezia

Tutorato di Probabilità e Statistica

Marco Fiorucci

Università Ca' Foscari di Venezia

mfiorucc@dsi.unive.it

Università
Ca' Foscari
Venezia

Introduction

R: The most powerful and most widely used statistical software

Università Ca' Foscari Venezia

The New York Times

Jan 2009

Data Analysts Captivated by R's Power

"R is really important to the point that it's hard to overvalue it," said Daryl Pregibon**, a research scientist** at **Google**, which uses the software widely. "It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

Forbes

Nov 10 2010

Names You Need to Know in 2011: R Data Analysis Software

"R is rapidly augmenting or replacing other statistical analysis packages at universities"

Features

- Open source, development- flexible, extensible
- Large number of statistical and numerical methods
- High quality visualization and graphical tools
- Extended by a very large collection of rapidly developing packages

Università
Ca' Foscari
Venezia

Why is it called R?

The name is partly based on the (first) names
of the first two R authors and
partly a play on the name of the Bell Labs language 'S

Initially written by Robert Gentleman, & Ross Ihaka,
Dept of Statistics, University of Auckland, New Zealand
(1996)

Short R History

Î1991: Ross Ihaka, Robert Gentleman begin work on a project that will become R

1993: The first announcement of R

1995: R available by ftp

1996: A mailing list is started and maintained by Martin Maechler at ETH

1997: The R core group is formed

2000: R 1.0.0 is released

Università Ca' Foscari Venezia

Short R History Continued

2001: Bioconductor for the analysis andcomprehensio of genomic data using R

2008: The Omegahat project to enable connectivity between R and other languages

2010: Former co-founder and employees of SPSS found Revolution Analytics, a company which offers a commerical package around R.

2011: Rstudio Project provide a free open source integrated development environment (IDE) for R

Università Ca' Foscari Venezia

R as a calculator

Simple computations are evaluated immediately in the CONSOLE (Try it)

3+4

6/18

4*3

4**3 or 4^3

sqrt(25)

log(100) vs. log10(100); exp(3) vs. 10**3

help(log)

But this is BORING!

Università
Ca' Foscari
Venezia

R language elements

Università Ca' Foscari Venezia

Objects

Everything in R is an named OBJECT.
Examples of objects are vectors, matrices, lists, dataframes, functions.

Object names start with a letter, then can contain letters, numbers, periods (.) or underscores (_) arbitrary length.

x;X ; x47 ; thisisaverylongnamethatishardtoread
Object names are **case sensitive**

Objects are not explicitly typed and the type is determined by examining the object, i.e. is an object a number, a character, a function etc. The type of an object can be dynamically changed at any time,

Università Ca' Foscari Venezia

What objects are in my workspace?

objects() # displays all objects in my workspace

ls() # alternate method for UNIX gurus

print(objectname) # displays contents of object

objectname # acts like print() in most cases (but not always)

str(objectname) # display "structure" of object

rm(objectname) # removes an object from workspace

Try it.

Università Ca' Foscari Venezia

Object assignment.

```
x <- 10    # numeric (integer or real or complex)
w = 15     # alternate way to assign
qq = x * w     # product of two objects values
z <- 14.7
y <- TRUE      # notice upper case; also use T and F
name <- "Carl"
second.name <- "James"
ww <- sqrt(x)
ww <- sqrt(name) # oops not a valid operation
ww <- 15 + TRUE # what do you get?
```

Principal types of values

- numeric (integers or real numbers), 16 digits of precision
- character (a.k.a. string) - length 0+, delineated by matching "
- logical (TRUE or FALSE which are interpreted as 1 or 0 in arithmetic operations)
- function (you can create your own functions - very useful)
- factor (value lables)

Università Ca' Foscari Venezia

Object data structures

Data is stored in workspace in 5 data structures (MOST COMMON)

SCALAR, i.e. single value,

VECTOR, i.e. set of values of SAME TYPE,

MATRIX, i.e. rectangular set of values of SAME TYPE
array, i.e. multidimensional set of values of SAME TYPE

DATAFRAME, i.e. collection of VECTORS of same length of DIFFERENT TYPES

LIST, i.e. general collection of ANY objects (data and others) of ANY type

Vectors

VECTORS is a set of values ALL OF THE SAME TYPE

```
age <- c(56, 56, 28, 23, 22)
height <- c(185, 162, 185, 167, 190)
f.names <- c(" Carl ", "Lois", " Matthew ", " Marianne ")
over.30 <- c(TRUE, TRUE, FALSE, FALSE, FALSE)
odd <- c(2.3, "Carl") # surprising, but look at result!
length(age)
length(f.names)
str(age) # what is the structure of age?
str(f.names)
```

Università Ca' Foscari Venezia

Dataframes

DATAFRAMES are collections of VECTORS of SAME length, but DIFFERENT types. Not the same as a matrix (this must be same type)

```
age <- c(56, 56, 28, 23, 22)

height <- c(185, 162, 185, 167, 190)

f.names <- c(" Carl ", "Lois", " Matthew ", " Marianne" , "David")

over.30 <- c(T, T, F, F, F)

schwarz <- data.frame( f.names, age, height, over.30, stringsAsFactors=FALSE)
```

schwarz

str(schwarz)

length(schwarz) # number of vectors, not length of vectors

dim(schwarz)

nrow(schwarz)

ncol(schwarz)

names(schwarz)

```
data()

data(cars)

cars

names(cars)

cars$speed

speed

attach(cars)

speed
```

head(cars)

tail(cars)

summary(cars)

hist(speed)

boxplot(speed)

cars[speed>18,]

Università Ca' Foscari Venezia

Frequence Table

```
table(speed)

classi <-2*(0:15)

cut(speed,breaks=classi)

table(cut(speed,breaks=classi))

table(cut(speed,breaks=10))

Freqsum<-
cumsum(table(cut(speed,breaks=classi))/length(sp
eed))
```

```
Freqsum<-
cumsum(table(cut(speed,breaks=classi))/length(sp
eed))

hist(speed,breaks=10)
```

Provate a far variare il numero di classi e osservate come varia l'istogramma corrispondente.

Università Ca' Foscari Venezia

References

1. http://www.revolutionanalytics.com/
2. http://www.burns-stat.com/documents/tutorials/why-use-the-r-language/
3. Carl James Schwarz. Introduction to R Software (slides).