Introduction to R – a computing software for statistical analysis

Krzysztof Podgórski
Department of Mathematics and Statistics
University of Limerick

September 8, 2009

Quotation of the lecture

“I was still a couple of miles above the clouds when it broke, and with such violence I fell to the ground that I found myself stunned, and in a hole nine fathoms under the grass, when I recovered, hardly knowing how to get out again. Looking down, I observed that I had on a pair of boots with exceptionally sturdy straps. Grasping them firmly, I pulled with all my might. Soon I had hoist myself to the top and stepped out on terra firma without further ado.”

R. E. Raspe, Singular Travels, Campaigns and Adventures of Baron Münchausen, 1786.

Outline

1. The R Project for Statistical Computing
2. Statistical Tables using R
3. Data analysis with R
4. Bootstrap – in the end

Downloading and installing the R-package

R can be downloaded from the following website:

`http://www.r-project.org/index.html`

The highlights

The highlights

The highlights

- The package is available for all popular operating systems:

The highlights

- The package is available for all popular operating systems:
  - Windows

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux
- It is free!

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package
- Packages are available for download through a convienient facility

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package
- Packages are available for download through a convienient facility
- It is fairly well documented and the documentation is available either from the program help menu or from the website

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package
- Packages are available for download through a convienient facility
- It is fairly well documented and the documentation is available either from the program help menu or from the website
- It is the top choice of statistical software among academic statisticians but also very popular in industry specially among biostatisticians and medical researchers (mostly due to the huge package called **Bioconductor** that is built on the top of **R**)

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package
- Packages are available for download through a convienient facility
- It is fairly well documented and the documentation is available either from the program help menu or from the website
- It is the top choice of statistical software among academic statisticians but also very popular in industry specially among biostatisticians and medical researchers (mostly due to the huge package called **Bioconductor** that is built on the top of **R**)
- It is a powerful tool not only for doing statistics but also all kind of scientific programming

The highlights

- The package is available for all popular operating systems:
  - Windows
  - Mac OS 10
  - Linux
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package
- Packages are available for download through a convienient facility
- It is fairly well documented and the documentation is available either from the program help menu or from the website
- It is the top choice of statistical software among academic statisticians but also very popular in industry specially among biostatisticians and medical researchers (mostly due to the huge package called **Bioconductor** that is built on the top of **R**)
- It is a powerful tool not only for doing statistics but also all kind of scientific programming – **Chemoconductor???**

What is R? – only some basic information

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
    - linear and nonlinear modelling,
    - classical statistical tests,
    - time-series analysis,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
    - linear and nonlinear modelling,
    - classical statistical tests,
    - time-series analysis,
    - classification,
    - clustering,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
  - an effective data handling and storage facility,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
  - an effective data handling and storage facility,
  - a suite of operators for calculations on arrays, in particular matrices,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
  - an effective data handling and storage facility,
  - a suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for data analysis,

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
  - an effective data handling and storage facility,
  - a suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for data analysis,
  - graphical facilities for data analysis and display either on-screen or on hardcopy, and

What is R? – only some basic information

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented
  - linear and nonlinear modelling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering,
  - ...
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
  - an effective data handling and storage facility,
  - a suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for data analysis,
  - graphical facilities for data analysis and display either on-screen or on hardcopy, and
  - a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Outline

1. The R Project for Statistical Computing
2. Statistical Tables using R
3. Data analysis with R
4. Bootstrap – in the end

Textbook Appendix 2, Table A.1

The following is a fragment of the table of values of $F(x)$ for the standard normal cumulative distribution function from page 254 of the textbook

Table A.1 $F(z)$, the standard normal cumulative distribution function

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.4	0.0003	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005
−3.3	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007
−3.2	0.0007	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
−3.1	0.0010	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013
−3.0	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
−2.9	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
−2.8	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
−2.7	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
−2.6	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
−2.5	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
−2.4	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
−2.3	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
−2.2	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
−2.1	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
−2.0	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
−1.9	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
−1.8	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
−1.7	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537

The same "table" in R

Here is a simple code in R that produce the same values as in the table

```
#Preceding line with the symbol '#' makes it a comment in R
#The following line produce a single value of the standard normal cumulative
#function. It is the value corresponding to the first value in the table

pnorm(-3.4)

#[1] 0.0003369293

#Then the first row of the table

z=seq(-3.4,-3.31,by=0.01)
pnorm(z)

# [1] 0.0003369293 0.0003494631 0.0003624291 0.0003758409 0.0003897124
# [6] 0.0004040578 0.0004188919 0.0004342299 0.0004500872 0.0004664799

#And all values from the table

z=seq(-3.4,3.4,by=0.01)
pnorm(z)

   [1] 0.0003369293 0.0003494631 0.0003624291 0.0003758409 0.0003897124 0.0004040578 0.000418
  [11] 0.0004834241 0.0005009369 0.0005190354 0.0005377374 0.0005570611 0.0005770250 0.000597
  [21] 0.0006871379 0.0007113640 0.0007363753 0.0007621947 0.0007888457 0.0008163523 0.000844
  [31] 0.0009676032 0.0010007825 0.0010350030 0.0010702939 0.0011066850 0.0011442068 0.001182
  [41] 0.0013498980 0.0013948872 0.0014412419 0.0014889987 0.0015381952 0.0015888696 0.001641
  [51] 0.0018658133 0.0019262091 0.0019883759 0.0020523590 0.0021182050 0.0021859615 0.002255
  [61] 0.0025551303 0.0026354021 0.0027179449 0.0028028146 0.0028900681 0.0029797632 0.003071
  [71] 0.0034669738 0.0035726010 0.0036811080 0.0037925623 0.0039070326 0.0040245885 0.004145
```

There is more than meets the eye in the table

There is more than meets the eye in the table

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

There is more than meets the eye in the table

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- The density represents distribution of probability for a random variable associated with it.

There is more than meets the eye in the table

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- The density represents distribution of probability for a random variable associated with it.
- The area under the density represents the probability so the that the total area under it is equal to one.

There is more than meets the eye in the table

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- The density represents distribution of probability for a random variable associated with it.
- The area under the density represents the probability so the that the total area under it is equal to one.
- The area accumulated up to certain value z represents probability that a corresponding random variable takes value smaller than z and this probability defines the cumulative distribution function $F(z)$ which is tabularized.

All this can be seen in **R**

The following code explores various aspects of the standard normal distribution

```
#Plotting the density function of the standard normal variable

z=seq(-3,3,by=0.01)
plot(z,dnorm(z),type='l',col="red",lwd=4)

#Plotting the cumulative distribution function (that one from the table)

plot(z,pnorm(z),type='l',col="red",lwd=4)

#And plotting them one at the top of the other


par(mfrow=c(2, 1))

plot(z,dnorm(z),type='l',col="red",lwd=4)

plot(z,pnorm(z),type='l',col="red",lwd=4)

#Side by side

par(mfrow=c(1, 2))

plot(z,dnorm(z),type='l',col="red",lwd=4)

plot(z,pnorm(z),type='l',col="red",lwd=4)
```

Outline

1. The R Project for Statistical Computing
2. Statistical Tables using R
3. Data analysis with R
4. Bootstrap – in the end

Data from Table 1.1 of the textbook

Table 1.1 Random and systematic errors

Student	Results (ml)					Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise, unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased

This is also given in the text file `Table1_1.txt` contents of which is given below

```
A    10.08    10.11    10.09    10.10
B    9.88    10.14    10.02    9.80
C    10.19    9.79    9.69    10.05
D    10.04    9.98    10.02    9.97
```

Reading data from a file to **R**

```
#Reading the data from
Titra=read.table("Table1_1.txt", row.names = 1)

Titra

#      V2    V3    V4    V5
#A 10.08 10.11 10.09 10.10
#B  9.88 10.14 10.02  9.80
#C 10.19  9.79  9.69 10.05
#D 10.04  9.98 10.02  9.97

#Listing the first row
Titra[1,]

#and the last column
Titra[,4]
```

Means and standard deviations

Example 2.1.1

Find the mean and standard deviation of A's results.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	10.08	−0.02	0.0004
	10.11	0.01	0.0001
	10.09	−0.01	0.0001
	10.10	0.00	0.0000
	10.12	0.02	0.0004
Totals	50.50	0	0.0010

$$\bar{x} = \frac{\sum x_i}{n} = \frac{50.50}{5} = 10.1 \text{ ml}$$

$$s = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)} = \sqrt{0.001/4} = 0.0158 \text{ ml}$$

Note that $\sum(x_i - \bar{x})$ is always equal to 0.

Means and standard deviations much faster and better

```
#Comuting means

rowMeans(Titra)

#       A       B       C       D
#10.0950  9.9600  9.9300 10.0025

#and standard deviation

apply(Titra,1,sd)
#          A          B          C          D
#0.01290994 0.15055453 0.23036203 0.03304038
```

Outline

1. The R Project for Statistical Computing
2. Statistical Tables using R
3. Data analysis with R
4. Bootstrap – in the end

Nitrate ion concentration from Table 2.1

Table 2.1 Results of 50 determinations of nitrate ion concentration, in μg ml^{-1}

0.51	0.51	0.51	0.50	0.51	0.49	0.52	0.53	0.50	0.47
0.51	0.52	0.53	0.48	0.49	0.50	0.52	0.49	0.49	0.50
0.49	0.48	0.46	0.49	0.49	0.48	0.49	0.49	0.51	0.47
0.51	0.51	0.51	0.48	0.50	0.47	0.50	0.51	0.49	0.48
0.51	0.50	0.50	0.53	0.52	0.52	0.50	0.50	0.51	0.51

Also in the file `Table2_1.txt`

```
0.51    0.51    0.51    0.50    0.51    0.49    0.52    0.53    0.50    0.47
0.51    0.52    0.53    0.48    0.49    0.50    0.52    0.49    0.49    0.50
0.49    0.48    0.46    0.49    0.49    0.48    0.49    0.49    0.51    0.47
0.51    0.51    0.51    0.48    0.50    0.47    0.50    0.51    0.49    0.48
0.51    0.50    0.50    0.53    0.52    0.52    0.50    0.50    0.51    0.51
```

The mean concentration

```
#Getting data in a vector
x=scan('Table2_1.txt')

mean(x)
#[1] 0.4998

sd(x)
#[1] 0.01647385
```

Pulling ourselves by bootstraps

- If we would repeat our experiment of collecting 50 samples of nitrate concentrations many times we would see the range of error.
- But it would be a waste of resources and not a viable method.
- Instead we resample 'new' data from our data and use so obtained new samples for assessment of the error.
- The following **R** code does the job.

```
#Getting data in a vector
m=mean(x)
bootstrap=vector('numeric',500)
for(i in 1:500)
{
bootstrap[i]=mean(sample(x,replace=T))-mean(x)
}

#The distribution of estimation error
hist(boostrap)
```

We can safely say that the nitrate concentration is 49.99 ± 0.005.

Statistical joke of the week

A prisoner had just been sentenced for a heinous crime and was returned to his cell. An inquisitive guard could not wait to ask him about the outcome.

Statistical joke of the week

A prisoner had just been sentenced for a heinous crime and was returned to his cell. An inquisitive guard could not wait to ask him about the outcome.

Guard: “What did you get for a sentence?”

Statistical joke of the week

A prisoner had just been sentenced for a heinous crime and was returned to his cell. An inquisitive guard could not wait to ask him about the outcome.

Guard: "What did you get for a sentence?"
Prisoner: "I could choose life or 100 years."

Statistical joke of the week

A prisoner had just been sentenced for a heinous crime and was returned to his cell. An inquisitive guard could not wait to ask him about the outcome.

Guard: “What did you get for a sentence?”
Prisoner: “I could choose life or 100 years.”
Guard: “And what did you choose?”

Statistical joke of the week

A prisoner had just been sentenced for a heinous crime and was returned to his cell. An inquisitive guard could not wait to ask him about the outcome.

Guard: "What did you get for a sentence?"
Prisoner: "I could choose life or 100 years."
Guard: "And what did you choose?"
Prisoner: "Well, life, obviously. Statistically speaking that is shorter."