A Short Introduction to R

A. Di Bucchianico
Eindhoven University of Technology
Laboratory for Industrial Mathematics Eindhoven
P.O. Box 513
5600 MB Eindhoven
The Netherlands
a.d.bucchianico@tue.nl
http://www.win.tue.nl/∼adibucch

Abstract

This note briefly describes the context of the statistical software **R**, as well as some basic commands for handling data, producing graphs, computing statistics, performing statistical tests and writing functions. There are some exercises with solutions to illustrate basic analyses.

Contents

1 The R initiative

There are many commercial statistical softwares available. Well-known examples include SAS, SPSS, S-Plus, Minitab, Statgraphics, GLIM, and Genstat. Usually there is a GUI (graphical user interface). Some softwares allow to perform analyses using the GUI as well as by typing commands on a command line. Larger analyses may be performed by executing scripts.

In the 1970's Chambers of AT&T started to develop a computer language (called S) that would be able to perform well-structured analyses. A commercial version of S appeared in the early 1990's under the name S-Plus. Ihaka and Gentleman developed a little bit later a free, open source language R which is very similar to S. Currently R is being maintained and continuously improved by a group of world class experts in computational statistics. Hence, R has gained enormous popularity among various groups of statisticians, including mathematical statisticians and biostatisticians. The R-project has its own web page at www.r-project.org. Downloads are available through the CRAN (Comprehensive R Archive Network) at www.cran.r-project.org.

2 R basics

There are several tutorials available inside R through Help or can be found on the web, *e.g.* through CRAN. The R reference card is very useful. Within R further help can be obtained by typing `help` when one knows the name of a function (*e.g.*, `help(pnorm)`) or `help.search` when one only keywords (*e.g.*, `help.search("normal distribution")`).

2.1 Data files

Assignment are read from right to left using the $\leftarrow$ operator:

$$\texttt{a} < -2 + \texttt{sqrt}(5)$$

There are several form of data objects. Vectors can be formed using the `c` operator (concatenation), *e.g.*,

$$\texttt{a} < -\texttt{c}(1, 2, 3, 10)$$

yields a vector consisting of 4 numbers. Vectors may be partitioned into matrices by using the `matrix` command, *e.g.*,

$$\texttt{matrix}(\texttt{c}(1, 2, 3, 4, 5, 6), 2, 3, \texttt{byrow} = \texttt{T})$$

creates a matrix with 2 rows and 3 columns.

The working directory may be set by `setwd` and displayed by `getwd()` (this will return an empty answer if no working directory has been set). Please note that directory names should be written with quotes and that the Unix notation must be used even if R is being used under Windows, *e.g.* `setwd("D:/MyData")` or one must double the slashes as follows: `setwd("D:\\MyData")`. A data set may be turned into the default data set by using the command `attach`; the companion command `detach`. Data files on a local file system may be read through the command `scan` when there is only one column or otherwise by

$$\texttt{read.table}(\text{“}\texttt{file.txt}\text{”}, \texttt{header} = \texttt{TRUE})$$

Both `read.table` and `scan` can read data files from the WWW (do not forget to put quotes around the complete URL).

Parts of data files may be extracted by using so-called subsetting. The command `d[r,]` yields the rth row of object d, while $d[, c]$ yields the cth column of object d. The entry in row r and column c of object d can be retrieved by using `d[r,c]`. Extracting elements that satisfy a certain condition may also be extracted by subsetting. *E.*g., `d[d<20]` yields all elements of d that do not exceed 20, while `d["age"]` extracts the column with name "age" (note the double quotes) from object d. The number of elements of an object d is given by `length(d)`. The names of the variables (columns) of a data frame or table can be retrieved by `colnames`. The collection of values attained by a variable can be obtained with `levels`.

2.2 Probability distributions in R.

Standard probability distributions have short names in R as given by Table 1. Several probability functions are available. Their names consists of two parts: the first part is the name of the function (see Table 2), while the second part is the name as in Table 1. *E.g.*, a sample of size 10 from an exponential distribution with mean 3 is generated in R by `rexp(10,1/3)` (R uses the failure intensity instead of the mean as parameter).
Table 3 lists several goodness-of-fit tests that are available in R, either directly or via the package `nortest` (see Subsection 2.4).

Distribution	Name in R
normal	norm
(non-central) Student T	t
Weibull	weibull
exponential	exp
(non-central) χ^2	chisq
Gamma	gamma
F	f

Table 1: Names of probability distributions in R.

Function	Name in R
d	density
p	probability = cumulative distribution function
q	quantile
r	random numbers

Table 2: Names of probability functions in R.

2.3 Graphics in R

The standard procedure in R to make 1D and 2D plots is plot. Histogram are available through hist. These commands can be supplied with options to allow for titles, subtitles, and labels on the x-axes:

```
plot(data,main=''Title'',sub=''Subtitle'',xlab=''X-axis'',ylab=''Y-axis'')
```

Quantile-quantile plots are available through qqplot, while qqnorm yields a plot of the quantiles of a data set against the quantiles of a fitted normal distribution (normal probability plot). Dot plots are available through stripchart. A Box-and-Whisker plot is also available for exploratory data analysis through boxplot (if the data set is a data frame like produced by read.table, then multiple Box-and-Whisker plots are produced). The empirical cumulative distribution function is available through ecdf. Histograms are available through hist or histogram; the library **MASS** (cf. Subsection 2.4) contains an improved version truehist. The library **Hmisc** (cf. Subsection 2.4) has the command histbackback to produce histograms for comparing two data sets. Kernel density estimators are available through density. Graphics can be saved to files by choosing File and Save as in the menu of the R console. In order plot variables conditional on other variables, use the split command or more generally, use the formula notation (*e.g.*, $x \sim y1 + y2$ to plot the values of x conditional on the values of both $y1$ and $y2$. Plots may be combined by using the option add = TRUE. Legends may be added by the legend command. The position can be entered with keywords like "topright" as the first argument. Functions may be plotted using the curve command. Graphical options can be set globally using par; a list of graphical options can be obtained with par(). Local options like lwd (line width), pch (point character = point type), cex (character expansion = magnification factor for characters) can be added to most plotting functions.,,

Test	Name in R	Package
Shapiro-Wilks	shapiro.test	stats
Kolmogorov (1-sample)	ks.test	stats
Smirnov (2-sample)	ks.test	stats
Anderson-Darling	ad.test	nortest
Cramér-von Mises test	cvm.test	nortest
Lilliefors test	lillie.test	nortest

Table 3: Names of goodness-of-fit tests in R.

2.4 Libraries in R

Extensions to the basic functions are available through libraries of functions. Libraries need to be installed once and need to be loaded when needed using `library`. In the Windows interface of R, these libraries can be loaded or installed by choosing the option `Packages` in the menu. Libraries may also contain ways to improve exchange of files with other software like Matlab or WinEdt. Examples of useful libraries include:

`survival`: library for survival analysis (Cox proportional hazards etc.)

`qcc`: SPC library

2.5 Basic statistics in R

Summary statistics of a data set can be obtained from `summary`, or by using individual commands like `mean`, `sd`, `mad`, and `IQR`. Standard hypothesis tests are also available, *e.g.*, `t.test` yields the standard tests for means of one or two normal samples.

2.6 Functions in R

Analyses that have to be performed often can be put in the form of functions, *e.g.*,

$$\begin{aligned} \texttt{mysimplef} \quad <- \quad & \texttt{function}(\texttt{data}, \texttt{mean} = 0, \texttt{alpha} = 0.05) \\ & \{\texttt{hist}(\texttt{data}), \texttt{t.test}(\texttt{data}, \texttt{conf.level} = \texttt{alpha}, \texttt{mu} = \texttt{mean})\} \end{aligned}$$

This means that typing `mysimplef(data,4)` uses the default value $\alpha = 0.05$ and tests the null hypothesis $\mu = 4$.

2.7 Editors for R

Instead of pasting commands from an ordinary text editor into the R console, one may also use WinEdt as R editor by using the RWinEdt package. Another choice is Tinn-R, which is a free R editor that helps the user by showing R syntaxis while typing.

3 Exercises

Exercise 1 *Calculate in* R *the probability that a random variable with a* χ^2 *distribution with* 3 *degrees of freedom is larger than* 5.2*.*

Exercise 2 *Compute the* 0.99 *quantile of the standard normal distribution.*

Exercise 3 *Generate* 20 *random numbers from an exponential distribution with mean* 3*. How can you check that you choose the right parametrization?*

Exercise 4 *Generate* 40 *random numbers from a normal distribution with mean* 10 *and variance* 2*. Make a histogram and play with the number of bins to convince yourself of the influence on the shape. Check normality through a normal probability plot, a plot of the density and an appropriate goodness-of-fit test. Also test the mean and variance of the sample.*

Exercise 5 *A telecommunication company has entered the market for mobile phones in a new country. The company's marketing manager conducts a survey of* 200 *new subscribers for mobile phones. The data of her survey are in the data set* `telephone.txt`*, which contains the first month's bills. Make an appropriate plot of this data set. What information can be extracted from these data? What marketing advice would you give to the marketing manager?*

Exercise 6 *Dried eggs are being sold in cans. Each can contains two different types of eggs. As part of a quality control programme, the fat content of eggs is being investigated. The investigation is divided over* 6 *different laboratories. Every laboratory receives the same number of eggs of both types. Testing the fat content of eggs is destructive, so that each egg can only be investigated once. Since measuring the fat content is time consuming, the measurements are divided over* 2 *laboratory assistants within each laboratory.*

A quality manager applies a certain statistical procedure and claims that there is a significant difference between the fat contents measured by the laboratories. This report causes confusion, since the 6 *laboratories all have a good reputation and there is no reason to expect large variation in fat content of the eggs. Find an explanation by making appropriate plots of the data set* `eggs.txt`.

Exercise 7 *Supermarket chain ATOHIGH has two shops A and B in a certain town. Both shops are similar with respect to size, lay-out, number of customers and spending per customer. The populations of the parts of town of the two shops are quite similar. Management decides to experiment with the lay-out of shop A, including different lighting. After some time a survey is performed on a Saturday afternoon among* 100 *customers in both shops. The survey is restricted to customers which are part of a family of at least* 3 *persons. The data set* `supermarket.txt` *contains the data of this survey. Perform a statistical analysis of this data set, both by producing appropriate plots and by computing appropriate summary statistics. What can you conclude about the spending in both shops?*

Exercise 8 *Write an* `R` *function that computes the empirical survivor function given a data set. The empirical survivor function at a point x counts the proportion of observations exceeding x.*

Exercise 9 *Write an* `R` *function that produces a confidence interval for the variance from a sample from the normal distribution.*

Exercise 10 *Write a function that plots a plot of the values of a data set against the quantiles of a given Weibull distribution. Test your function with a*

a) a random sample of size 30 *from a Weibull distribution with shape parameter* 3 *and scale parameter* 2*.*

b) a random sample of size 30 *from a Gamma distribution with shape parameter* 7.57 *and rate parameter* 0.235*. Check using formulas that the mean and variance of this Gamma distribution is approximately equal to the mean and variance of the Weibull distribution in 1).*

Exercise 11 *In the 19th century French physicists Fizeau and Foucault independently invented ways to measure the speed of light. Foucault's method turned out to be the most accurate one. The Foucault method is based on fast rotation mirrors. The American physicists Michelson and Newcomb improved Foucault's method. The data set* `light.txt` *contains measurements by Newcomb from 1882. The data are coded times needed by light to travel a distance of* 3721 *metres across a river and back. The coding of these measurements was as follows: from the original times in microseconds measured by Newcomb first* 24.8 *was subtracted, after which the results were multiplied with* 1000*.*

a) Compute a 95% *confidence interval for the average speed of light.*

b) State the assumptions on which the above interval is based. Check your assumptions with a suitable plot. What is your conclusion?

c) Make a box plot of the data. What do you observe?

d) Plot the data against the observation number. Provide a possible explanation for what you observed in part c).

e) Recompute a 95% confidence interval for the average speed of light using your findings of part c).

f) Use the WWW to find the currently most accurate value for the speed of light. Test whether this new value is consistent with the measurements of Newcomb.

Exercise 12 *To improve rain fall in dry areas, an experiment was carried out with 52 clouds. Scientists investigated whether the addition of silver nitrate has a positive effect on rainfall. They chose 26 out of a sample of 52 clouds and seeded it with silver nitrate. The remaining 26 clouds were not treated with silver nitrate. The data set* `clouds.txt` *records the rainfall in feet per acre.*

a) Apply a t-test to investigate whether the average rainfall increased by adding silver nitrate. Argue whether the data are paired or not.

b) The t-test assumes normally distributed data. Check both graphically and by using a formal test whether the data are normally distributed. What conclusion should you draw on the test performed in part a)?

c) Because the scientists thought that addition of silver nitrate should have a multiplicative effect, they suggested transforming the data. What transformation should be a logical candidate? What effect does this transformation have on the normality assumption? Apply the t-test of part a) to the transformed data. What is your final conclusion on the addition of silver nitrate?