Teaching descriptive statistics using R

[1]Dedi Rosadi, [2]Khabib Mustofa, [3]Iman Sanjaya, [4]Hendra Perdana, and [5]Krisna Mutiara Wati

[1]Statistics and Computing Research Group, Department of Mathematics, Gadjah Mada University, Indonesia;
[2]Department of Computer Science, Gadjah Mada University, Indonesia;
[3]Ministry of Research and Technology, Indonesia;
[4]Postgraduate Program in Mathematics, Department of Mathematics, Gadjah Mada University, Indonesia;
[5]Program Pasca Sarjana, Gadjah Mada University, Indonesia
Coorresponding Author: dedirosadi@gadjahmada.edu

Abstract. In this paper, we introduce the application of R for teaching statistics descriptive subject which is usually given in the undergraduate statistics course. As an illustration, we will consider the use of R for teaching the subject frequency distribution table, both using the R-CLI and R-GUI version. The R-GUI version used here is a part of Rplugin.SPSS which is currently under our extensive development. Rplugin.SPSS is a R Commander Plugin, which is a reorganized and an extented version of the menu of Statistics in R-Commander, a SPSS-like version menu. It can be also considered as the extended version of Rplugin.Econometrics (Rosadi, 2010). Further details and further examples can be found in Rosadi (2011, 2013).

Keywords: R Commander Plug-ins, *Open Source*, Statistical Analysis, SPSS.

Introduction

R (R Development Core Team, 2013), an open source programming environment for data analysis and graphics, has becoming the `lingua franca' of data analysis and statistical computing. It is available for a number of platforms (Windows, Mac OS and various Unix and Linux), frequently updated (latest version per AUGUST 2013 is 3.0.1), a "reliable" open-source software (its kernel developed by leading statistician and programmers, called as R-Core Team), an extensive online information and user-help (http://www.R-project.org and r-help@r-project.org), etc..

The functionality of R is based on the Add-on packages/library (similar to Toolbox in MATLAB). Default installation of R will automatically install and load several basic packages i.e., base, datasets, utils, grDevices, graphics, stats, methods. Beyond these basic packages, there are thousand of contributed packages, available in CRAN websites (http://cran.rproject.org/), ranging from the classic statistical methods until the most recently published in the statistical literature. Extensive description of capability of R, for instance has been grouped into some particular purposes in Task Views of R.

Unfortunately, for teaching purpose, R-CLI seems to be less user friendly and relatively difficult to use, especially if we compare it with the commerical softwares which have an extensive GUI capabilities, such as, among other, the popular SPSS. For solving this problems, Fox (2005, 2009) introduced RCmdr and its plugin extension, to facilitate the menu of R-GUI. In this paper, we discuss the capabilities of new GUI package for statistical analysis, which is compatible with the menu of SPSS, that we called as RcmdrPlugin.SPSS (Rosadi, 2013). In particular, we provide a simple example of the use of the menu for teaching statistics descriptive topics which is given in a standard basic statistics course for undergraduate student.

The rest of this paper is organized as follows. In the second section, we review the R-GUI, and especially the R-Commander which is the most popular R-GUI currently available for basic statistical analysis. For the third part, we provide a detail of design and implementation of the plugin. We provide an empirical example of application of plugin for analysis of real data.

R-GUI

The main interaction of R user is using CLI (Command Line Interface), therefore, for some user (beginners) it less interactive and not user-friendly, and it is relatively difficult to learn, especially compared to when studying the statistical software with extensive GUI. For this purpose, some statistican and programmers have been developed the R-GUI version, such as R-Commander (Fox, 2005), R Sciviews (http://www.sciviews.org/SciViews-R), JGR (R GUI Interface using Java, see http://www.rosuda.org), etc. (see http://www.sciviews.org for more information on R-GUI).

One of the most popular R-GUI package is R Commander (Rcmdr). It is developed using languange tcl/tk (Welch, Jones and Hoobs, 2003, Dalgaard P, 2001a,b, 2002) and provide the point and click GUI for doing some basic statistical analysis, such as:

1. Data Management menu, such as for data entry, exporting (from external formats) and manipulating data.
2. Basic statistics : such as descriptive statistics, mean, proportion and variance test, etc.
3. Advanced statistics tools: such as multivariate analysis, non parametric analysis, regression analysis, etc.
4. Various graphics for visualizing data.
5. Various functions for doing some analysis of statistical distribution.

In general, R commander is a useful GUI for doing the most common use basic statistical analysis. However, it less extensive when compare it with some popular statistical software, such as SPSS. Fortunately, it can be easily extended using suitable plug-in (Fox, 2009).

Design and Development of RcmdrPlugin.SPSS

Rcmdr Plug-in is a plug-in that allow the package developers to develop GUI to their R package, with the R Commander providing most of the necessary intrastructure (in a similar way as the add-ins menu in Microsoft excel). Rcmdr Plugin introduced by John Fox and illustrated using his package called as RcmdrPlugin.TeachingDemos (Fox, 2009). Currently, in CRAN server, it has been available about 29 Rcmdr plug-ins, see also Rosadi, (2010). This Plug-in contains GUI menu for doing several statistical method, and in particular can be used as the teaching tools for various statistical courses. However, in our opinion, most of the plugin suffers from several problems, such as:

- The menu are not well organized, it scattered on several sub menus
- The input and output dialog for doing a particular analysis is less interactive and too simple, especially if we compare it with GUI menu available in the commercial statistical softwares, e.g. SPSS

Motivated by these reasons, we develop a new plugins for R Commander that we called Rplugin.SPSS. Rplugin.SPSS is a R Commander Plugin, which is a reorganized and an extented version of the menu of Statistics in R-Commander, a SPSS-like version menu. It can be also considered as the extended version of our plugin Rplugin.Econometrics (Rosadi, 2010).

There are several goals design and developments for this plugin, as follows:

- Open Source and Multiplatforms
  RcmdrPlugin.SPSS is design to be multiplatforms, available for Windows and Linux. The plug ins is open source and will be released under GPL, and it is available in CRAN server. Some additional information will be available in the personal homepage of the corresponding author (http://dedirosadi.staff.ugm.ac.id).
- The ease of use

The menu sctructure of RcmdrPlugin.SPSS has been designed to group appropriately as the purpose of the analysis, self-defined and easy to be accessed, a SPSS-like version menu . Furthermore, when design the GUI layout, we only used the standard version of package tcltk and tcl/tk, and avoid to use the function that requires to install the extended version of tcltk and tcl/tk.

- Input – Output dialog
  For design of the input and output GUI dialog and layout, we use the comprehensive style and try to be compatible with GUI input and output that available in the commercial software, especially SPSS. Furthermore, since R is used as the computation engine in the background, make the output of a particular menu can be easily programmable to give the output as the standar procedure in statistical analysis. For instance, the menu for doing time series analysis, which is called ARIMA analysis, can be used to identify, estimate the model, do diagnostic check and forecast based on the model.
- Menu coverage
  As the goal of the plugin is to be SPSS-like menu, the coverage of the menu of R-GUI in RcmdrPlugin.SPSS will be similar to SPSS. Since this plugin is under development, in the current version of the menu, only some menus has been implemented completely. However, some analyses are more complete than the menu available in SPSS. For instance, the GUI menu for doing time series analysis has been extensively developed in Rosadi (2010) and now becoming part of the plugin, where it capability far beyond the analysis offered in menu time series analysis in SPSS.

In the next section, we provide an empirical example of using the plugin, especially for making frequency distribution table and its graphics, a standard topic discussed in the statistics descriptives topic given in the standar basic statistics course.

RcmdrPlugin.SPSS: Empirical example

Raw data which has been collected during a survey needs to be organized systematically for doing meaningful analysis to the data. There are several methods for organizing the data

a) Making its quantitative frequency distribution, i.e., data is grouped according to its quantitative frequency, for example in the term of the frequency distribution table
b) Making its qualitative frequency distribution, i.e., data is grouped according to its quaitative group
c) Data is arranged based on time order, i.e., in the term of time series.
d) Data is arranged based on geograhical location, i.e., spatially grouped data

One of the standard method for making the frequency distribution table is using Sturges method. For example, let assume we have adult height data, consisting of 50 respondent data (in centimetres)

176	167	180	165	168	171	177	176	170	175
169	171	171	176	166	179	181	174	167	172
170	169	175	178	171	168	178	183	174	166
181	172	177	182	167	179	183	185	185	173
179	180	184	170	174	175	176	175	182	172

Using Sturges method, we can obtain the following frequency distribution table

Table 4.1 The frequency distribution table of adult height

Interval	Frequency
164,5 - 167,5	6
167,5 - 170,5	7
170,5 - 173,5	8
173,5 - 176,5	11
176,5 - 179,5	7
179,5 - 182,5	6
182,5 - 185,5	5
Total	50

Based on this table, we can calculate the table cumulative frequency distribution, either in terms of the frequency or in the percentage. We further can describe the data using graphics such as the histogram, poligon and ogive.

For this particular data, we can the following R script to do the analysis:

```
>tinggibadan <- read.table("tinggibadan.txt", header=FALSE, sep="\t",
  na.strings="NA", dec=".", strip.white=TRUE) #importing data
>  library(fdth) #loading the required library
>  tabelfrektb1=fdt(tinggibadan$tb, start=164.5, end=185.5, h=3) #set table
>  tabelfrektb1
    Class limits  f   rf rf(%) cf cf(%)
   [164,168)  6 0.12    12  6    12
   [168,170)  7 0.14    14 13    26
   [170,174)  8 0.16    16 21    42
   [174,176) 11 0.22    22 32    64
   [176,180)  7 0.14    14 39    78
   [180,182)  6 0.12    12 45    90
   [182,186)  5 0.10    10 50   100
>summary(tabelfrektb1, col=c(1:2, 4, 6),format=TRUE, pattern='%.2f')
   Class limits  f rf(%) cf(%)
 [164.00, 168.00)  6    12    12
 [168.00, 170.00)  7    14    26
 [170.00, 174.00)  8    16    42
 [174.00, 176.00) 11    22    64
 [176.00, 180.00)  7    14    78
 [180.00, 182.00)  6    12    90
 [182.00, 186.00)  5    10   100
>  tabelfrektb2=fdt(tinggibadan$tb) #using default method of sturges
>  tabelfrektb2
```

The graphics can be obtained using the following script:

```
>  #histogram
>  plot(tabelfrektb1)
>  plot(tabelfrektb1, type='fh') #Absolut freq. histogram
>  plot(tabelfrektb1, type='rfh') #Relative freq. hist
>  plot(tabelfrektb1, type='rfph') #Relative freq. hist %
>  plot(tabelfrektb1, type='cdh') # Cumulative histogram
>  plot(tabelfrektb1, type='cfh')# Cumulative freq. hist
>   plot(tabelfrektb1,type='cfph')#Cumulative freq. % hist
>   # Poligons
>   plot(tabelfrektb1, type='fp')# Absolut freq. poligon
>   plot(tabelfrektb1, type='rfp')# Relative freq. poligon
>   plot(tabelfrektb1, type='rfpp')#Relative freq. %
>   plot(tabelfrektb1, type='cdp') # Cum. density poligon
>   plot(tabelfrektb1, type='cfp') # Cum. freq. poligon
>   plot(tabelfrektb1, type='cfpp') # Cum. freq. %
>   plot(tabelfrektb1, type='d') #density plot
```

The above script can be easily done using frequency distribution table menu in RcmdrPlugin.SPSS. Further detail is given in Rosadi (2011, 2013).

Conclusion

As we already mention above, RcmdrPlugin.SPSS is the plug-in R Commander designed for doing statistical analysis using a SPSS-like version menu. Some parts of the menu are even more comprehensive than the menu offered in SPSS. For the future, we expect that RcmdrPlugin.SPSS will become one of the best alternative software for doing statistical analysis.

Acknowledgement. This research receive financial support from Penelitian Unggulan UGM, funded from DP2M DIKTI.

References

Dalgaard, P., 2001a, The R-Tcl/Tk interface. In Kurt Hornik and Fritz Leisch, editors, Proceedings of the 2nd International Workshop on Distributed Statistical Computing, March 15-17, 2001, Technische Universität Wien, Vienna, Austria, 2001. URL http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/. ISSN 1609-395X.

Dalgaard, P., 2001b, A Primer on the R-Tcl/Tk Package, R News vol. 1/3, September 2001

Dalgaard, P., 2002, Changes to the R-Tcl/Tk package, R News vol. 2/3, December 2002

Fox, J. , 2005, The R Commander : A Basic –Statistics Graphical User Interface to R,. Journal of Statistics Software, Vol.14, Issue 9.

Fox, J., 2009, RcmdrPlugin.TeachingDemos,. [Online] Available at www.cran.r-project.org

R Development Core Team, 2013, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Rosadi, D., 2010, Rplugin.Econometrics: R-GUI for Teaching Time Series Analysis”. in Proceedings of COMPSTAT 2010, 19th International Conference on Computational Statistics, Paris-France, 22-27 Agustus 2010. ISBN 978-3-7908-2603-6

Rosadi, D., 2011, Econometrics and Time Series Analysis using R: Application for Economics, Business and Finance, Andi Offset, Yogyakarta (in Bahasa Indonesia)

Rosadi, D., 2013, Statistical Analysis using R, Draft Version, incoming book (in Bahasa Indonesia)

Welch, B., Jones, K. and Hobbs, J., 2003, Practical Programming in Tcl and Tk, 4th eds, Prentice Hall