SUPPORTING FREEWARE FOR STATISTICAL LECTURES – RKWard

Kateřina Janurová – Martina Litschmannová – Robert Skopal – Pavlína Kuráňová – Michal Běloch

Abstract

Until recently, the only statistical software used in the introductory statistical lessons in VSB-TU Ostrava were STATGRAPHICS Plus 5.1 and SPSS Statistics V23.0. Fact that neither of them is free to use creates two main problems. First problem is related to the licence limitation that doesn't allow software to be distributed outside the class so students can't use it on their own computers at home. The second problem is that after finishing the course in most cases students won't actually use such software in practice nor their further studies. In order to enable students to transfer their knowledge from lectures to practice, the appropriate statistical freeware is needed for the introductory course. Among other solutions under consideration, the free open-source GUI (Graphical User Interface) RKWard for R programming language appears to be the most suitable one, because its build in plugin architecture allows teacher to adapt the GUI for the requirements of the specific subject. Plus the rising popularity of R creates an opportunity to introduce students to modern programming language used in a wide range of areas. Our paper describes the RKWard extension and modification that were made for the purposes of the introductory statistical lectures at VŠB-TU Ostrava.

Key words: statistical lectures, R software, GUI RKWard, freeware

JEL Code: A

Introduction

The statistical software used in the introductory statistical lessons in VSB-TU Ostrava caused the students inconsiderable problems in transferring their knowledge from lectures to practice. This was mainly caused by the fact that the software was not available freely, but only with a limited license, so the students were not allowed to use it outside the school labs. Therefore, we searched for the statistical software, which is free to use, easy to learn and as close to specific needs for teaching the introductory statistical lessons as possible.

Our choice fell on R language complemented with the RKWard Graphical User Interface, due to several reasons. The first is the fact that R language has become the most popular tool used among statisticians and data miners due to its open-source license, platform independency and wide variety of available libraries. The second is the extensibility and customizability of RKWard GUI, which can be easily modified by the new plugins.

The aims of our paper are to present the RKWard extension and modification that were made for the purposes of the introductory statistical lectures at VŠB-TU Ostrava and to encourage the teachers to use or even modify RKWard for their own statistical lessons. It should be mentioned that the second part of our project for innovation of the introductory statistical lessons was the recording of short video tutorials for learning to use the new environment. All the tutorials are available in the czech language at the web pages of our lessons (Litschmannova, 2016).

1   RKWard Introduction

RKWard is a comprehensive environment for statistical computing. It combines the user-friendly GUI, which is comparable to commercial statistical software, with the power of development environment for R programming language. RKWard includes:

1. Spreadsheet-like data editor with ability to copy and paste from other statistical software e.g. Excel, possibility of sorting, subsetting and recoding of categorical data and automatic control of categorical entry data.

2. Data import in several import format e.g. SPSS, Stata, R data and CSV in different data format.

3. Workspace browser with possibility of showing data, sourced functions and loaded packages and their functions.

4. GUI dialogs for a variety of statistical functions, analysis and plots.

5. Plot preview and browsable history.

6. R console for direct writing of R commands for cases where the actual list of available statistical analysis is not sufficient.

7. Script editor for writing R functions.

8. Syntax highlighting, code folding and code completion in both R console and script editor.

9. R package management

10. Plugin architecture for RKWard extension and customizing.

The detailed description of particular parts of RKWard and their functionality is not the aim of this article, therefore we recommend the readers interested in this topic the article by Stefan Rödiger (Rödiger, 2012).

2 RKWard Modification

Apart from the RKWard benefits mentioned in previous sections, the key advantage in using RKWard for teaching statistics is the possibility to modify the environment for the needs of specific subject. Because of its plugin architecture the modifications and extensions are easy to make. RKWard contains complete GUI dialogs or reusable GUI components from which the new developer/teacher can learn relatively quickly.

The exhaustive description of plugin-writing process, from which the authors of this article have learned, is given by the web pages of RkWard Developer Team (Friedrichsmeier, 2015).

3 RKWard modification for the introductory statistical lectures at VŠB-TU Ostrava

RKWard for the introductory statistics course at the VŠB-TU Ostrava was modified by the four new plugins that were created according to the procedure described by Friedrichsmeier and Michalke (Friedrichsmeier, 2015)). The modification includes plugin for exploratory analysis of numerical and categorical variables, plugin for working with the basic probability distributions and plugin for the statistical inference for estimating confidence interval and hypothesis testing in the form of direct input of summary characteristics as well as raw input data.

3.1 Summary plugin

This plugin provides elementary exploratory analysis of selected variables. The user can choose from a variety of possible summary characteristics for analysis of numerical variable, as shown on Fig. 1. The results are given in the form of table, as shown on Fig. 2.

Fig. 1: The input window for the exploratory analysis of numerical variable
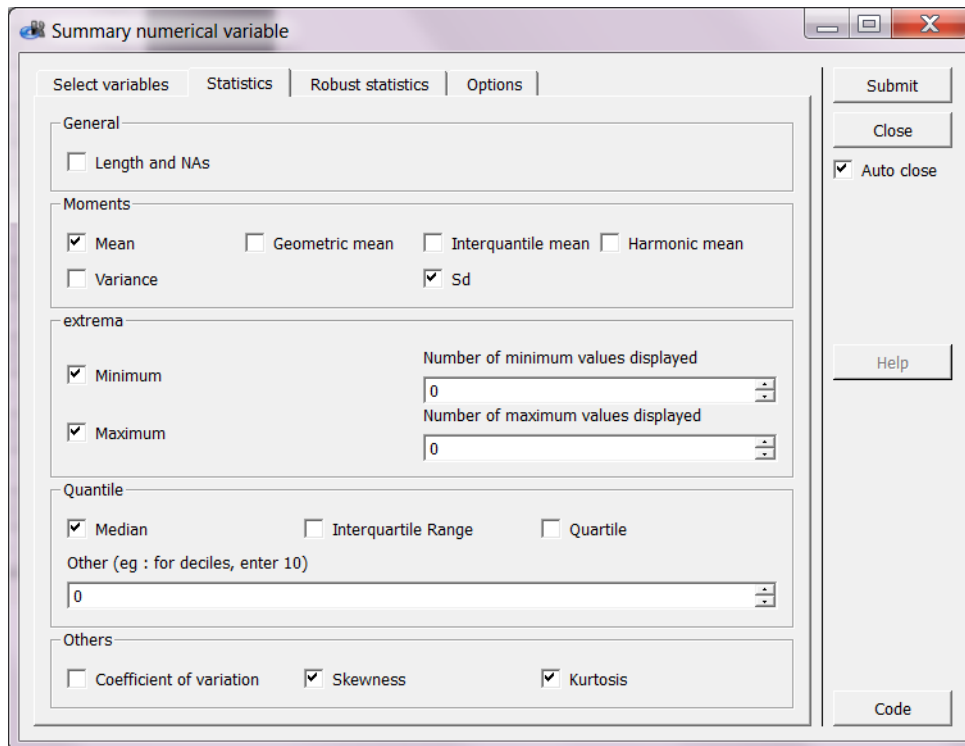
Fig. 2: The table of results of exploratory analysis of numerical variable

Univariate statistics

Parameters

- Omit missing values: yes

Wed May 11 13:20:48 2016

Variable Name	Mean	sd	Minimum	Maximum	Median	Skewness	Kurtosis
diabetes.1	8.043556	0.2549964	7.33	8.76	8.065	-0.03691537	0.11375541
diabetes.2	7.354333	0.3316626	6.44	8.05	7.350	-0.49589637	0.04558032

Run again

The results of analysis of categorical variable are given in the form of a frequency table.

3.2 Probability models plugin

Probability models plugin was created to teach the basic principles of probability distributions. It contains the set of discrete distributions: Binomial, Geometric, Hypergeometric, Negative-binomial, Poisson and the set of continuous distributions: Exponential, Normal, Weibull, Chi-Squared, Student and Fisher-Snedecor.

The calculation of probability values is available for each discrete and continuous distribution; in addition the quantile and density values can be calculated for all continuous

distribution. But the main advantage of this plugin in teaching statistics is the graphical output, where the students can see the meaning of calculated probability values on plotted graphs of distribution function and probability mass/density function. Fig. 3 shows graphical output for Poisson distribution with parameter $\lambda t = 5$ and number of trials $k = 4$. The corresponded graphs for Weibull distribution with parameters shape $\Theta = 5$ and scale $\beta = 3$ and time $t = 7$ are showed on Fig. 4.

Fig. 3: Output of calculated probability values for Poisson distribution with corresponding probability mass function and distribution function
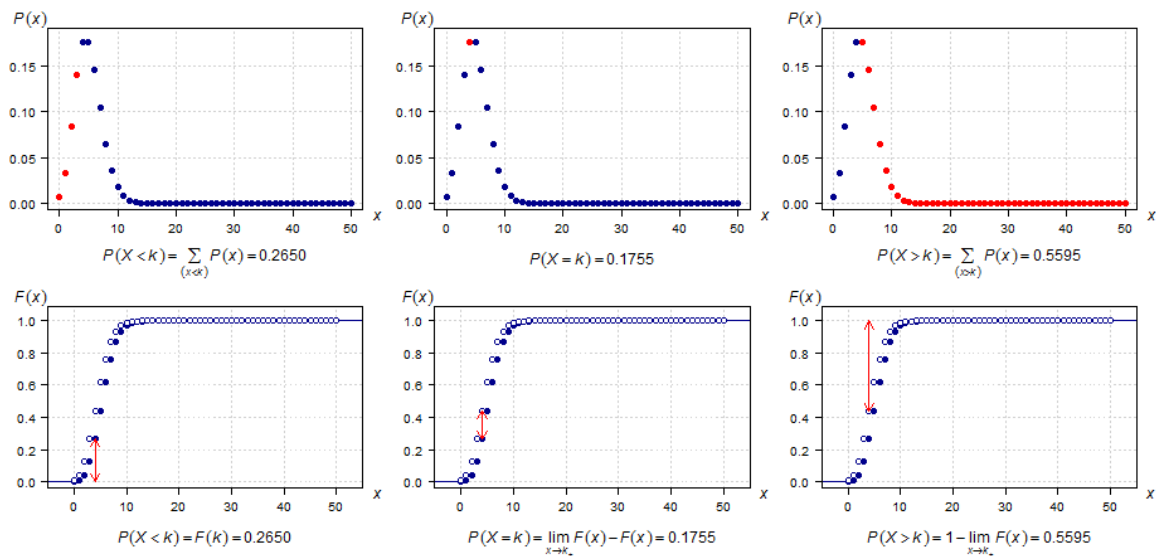
Poisson distribution

Parameters

- λt = Expected value: 5.00

Mon Jun 13 15:27:21 2016

Plots

- k = Number of trials: 4

Run again

Fig. 4: Output of calculated probability values for Weibull distribution with corresponding density function and distribution function
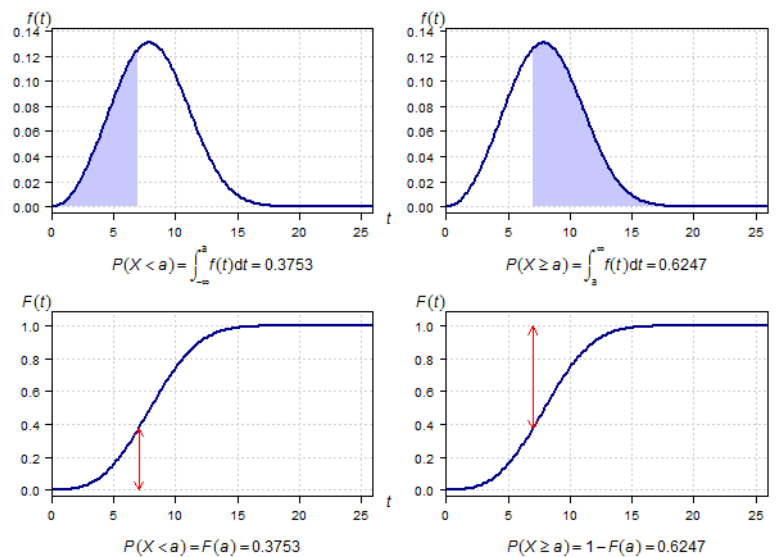
Weibull distribution

Parameters

- Θ = Shape: 3.000000000
- β = Scale: 9.000000000

Mon Jun 13 15:38:57 2016

Plots

- a = Time: 7

$$P(X < a) = \int_{-\infty}^{a} f(t)dt = 0.3753$$

$$P(X \geq a) = \int_{a}^{\infty} f(t)dt = 0.6247$$

$$P(X < a) = F(a) = 0.3753$$

$$P(X \geq a) = 1 - F(a) = 0.6247$$

Run again

3.3 Inference plugin

Inference plugin was created in order to provide the calculation of confidence interval and hypothesis testing on the basis of direct input of data summary characteristics for one or two input samples.

For one sample, the calculations are available for normal mean, normal standard deviation and binomial proportion, which can be calculated either by Wald test or by Clopper-Pearson test (Crow, 1956). In the case of two samples, the difference of normal means and binomial proportions (based on Wald test and Wilson test (Brown, 2002)) and ratio of normal standard deviations can be calculated. Fig. 5 shows the input window for the statistical inference for difference of binomial proportions.

716

Fig. 5: Input window for calculation of confidence interval and hypothesis testing on the basis of direct input of data summary characteristics for two input samples

The benefit of this plugin for teaching statistics is the output notation of confidence intervals and the null and alternative hypotheses, see Fig. 6, so the students can learn their correct form of in natural way.

Fig. 6: Output of calculated confidence intervals for difference of binomial proportions

Confidence intervals for the difference of proportions of binomial distribution using Wald test

$P(-0.25958 < \pi_1 - \pi_2 < 0.05958) = 0.95$

$P(\pi_1 - \pi_2 > -0.23392) = 0.95$

$P(\pi_1 - \pi_2 < 0.03392) = 0.95$

3.4   Hypothesis testing plugin

Hypothesis testing plugin is placed in the RKWard menu hierarchy under the Analysis menu. It is similar to the inference plugin in the sense of calculating confidence intervals and hypothesis testing, but the input is in the form of raw data.

One sample analysis is available for the mean and variance of normal distribution and for the median (provided by Wilcoxon test and Sign test (Lemmer, 1987)). Two-sample analysis is different for independent samples, where the difference/ratio of means (by the t-test and Welch's test (Welch, 1947)) and variances (by the F-test) of normal distribution or medians can be tested and for the paired data, where the analysis of difference of means of normal distribution and medians is provided (by the paired versions of t-test, Wilcoxon and Sign test). Data input for the two sample analysis is available in two formats: either in the form of two data columns or in the form of one data column and one indicator column, see Fig. 7. Multiple sample analysis contains the test of equality of means of normal distribution (ANOVA) and medians (Kruskal-Wallis test) and tests of equality of variances (Bartlett's, Levene's and Cochran's test (Brown & Forsyth, 1974). The output of Hypothesis testing plugin, which is essentially identical with the Inference plugin, is shown on Fig. 8.

Fig. 7: Input of hypothesis testing for difference of two means of normal distribution
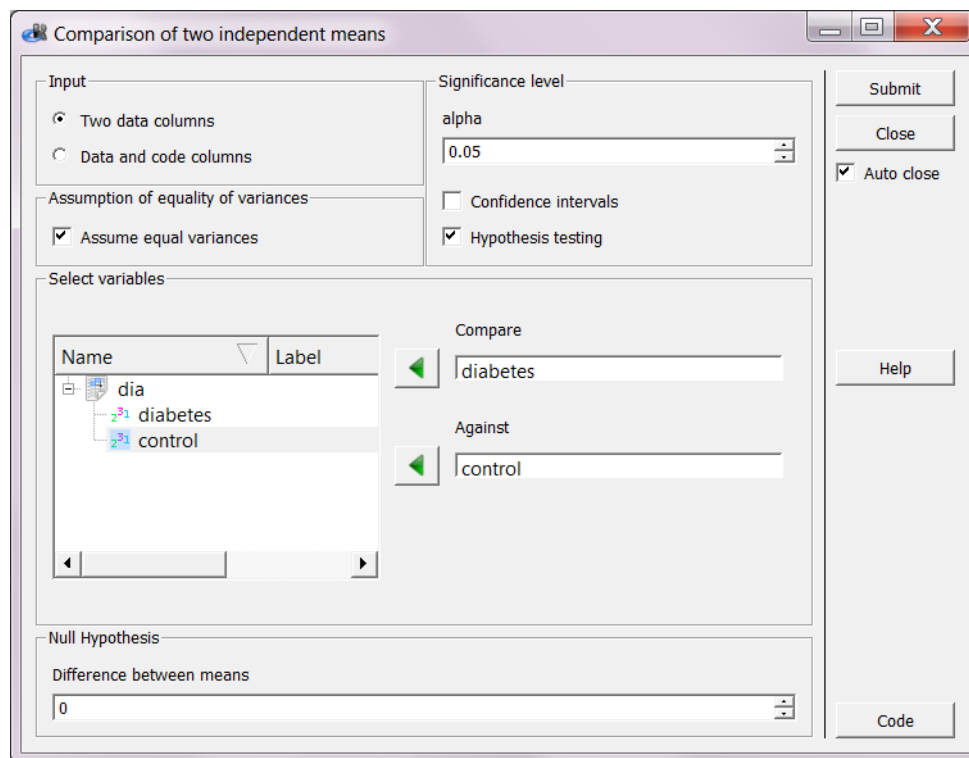
Fig. 8: Output of hypothesis testing for difference of two means of normal distribution

Student's *t*-test
Parameters
- Variable 1: diabetes
- Variable 2: control
- Significance level: 0.05
- Difference between means: 0.00

$H_0 : \mu_1 - \mu_2 = 0.00$

$H_A : \mu_1 - \mu_2 \neq 0.00$

Variables	Mean	Degrees of freedom	x_{obs}	p-value
diabetes	8.527	198	73.354	1.672013e-145
control	4.899			

$H_A : \mu_1 - \mu_2 > 0.00$

Variables	Mean	Degrees of freedom	x_{obs}	p-value
diabetes	8.527	198	73.354	8.360066e-146
control	4.899			

$H_A : \mu_1 - \mu_2 < 0.00$

Variables	Mean	Degrees of freedom	x_{obs}	p-value
diabetes	8.527	198	73.354	1
control	4.899			

Conclusion and future work

Our paper presents extension and modification of the Graphical User Interface for R programming language – RKWard, which were made especially for the purposes of the introductory statistical lectures at VŠB-TU Ostrava, in order to simplify the student's learning process. We described the functionality and benefits of four newly created plugins: the summary plugin for exploratory analysis of numerical and categorical variables, the probability plugin for introducing the basic probability distributions and inference and hypothesis testing plugins for estimating confidence interval and hypothesis testing.

The current version of modified RKWard was already tested in practise during the summer semester 2016, when the early errors were revealed and fixed. Until the next year we would like to add new plugins for contingency and association tables, mosaic plots, goodness of fit tests and Friedman test.

Acknowledgment

References

Brown, L.D., & Cai, T.T.(2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *Annals of Statistics, 30,* 160-201.

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association, 69*(346), 364-367. doi:10.1080/01621459.1974.10482955

Crow, E. L. (1956). Confidence Intervals for a Proportion. *Biometrika, 43*(3/4), 423. doi:10.2307/2332920

Friedrichsmeier, T., Michalke, M. : Introduction to Writing Plugins for RKWard. (n.d.). Retrieved June 21, 2016, from https://api.kde.org/doc/rkwardplugins/

Lemmer, H. H. (1987). A test for the median, combining the sign and signed-rank tests. *Communications in Statistics - Simulation and Computation, 16*(3), 621-627. doi:10.1080/03610918708812608

Litschmannová, M. : Videa. Retrieved June 21, 2016, from http://k470.vsb.cz/litschmannova/vyuka/statistika/studijni-materialy/videa/ (in czech)

Rödiger, S., Friedrichsmeier, T., Kapat, P., & Michalke, M. (2012). RKWard: A Comprehensive Graphical User Interface and Integrated Development Environment for Statistical Analysis with R. *Journal of Statistical Software J. Stat. Soft., 49*(9). doi:10.18637/jss.v049.i09

Welch, B. L. (1947). The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika, 34*(1/2), 28. doi:10.2307/2332510

Contact

Kateřina Janurová

IT4Innovations National Supercomputing Center

VŠB-Technical University of Ostrava

17. listopadu 15/2172

708 33 Ostrava-Poruba, Czech Republic

katerina.janurova@vsb.cz

Martina Litschmannová

Department of Applied Mathematics

VŠB-Technical University of Ostrava

17. listopadu 15/2172

708 33 Ostrava-Poruba, Czech Republic

martina.litschmannova@vsb.cz

Robert Skopal

Department of Applied Mathematics

VŠB-Technical University of Ostrava

17. listopadu 15/2172

708 33 Ostrava-Poruba, Czech Republic

robert.skopal@vsb.cz

Pavlína Kuráňová

Department of Mathematical Methods in Economics

VŠB-Technical University of Ostrava

17. listopadu 15/2172

708 33 Ostrava-Poruba, Czech Republic

pavlina.kuranova@vsb.cz

Michal Běloch

Department of Applied Mathematics

VŠB-Technical University of Ostrava

17. listopadu 15/2172

708 33 Ostrava-Poruba, Czech Republic

michal.beloch@vsb.cz