doi: 10.2478/v10313-012-0001-z
2012 /15

Advantages and Disadvantages of Professional and Free Software for Teaching Statistics

Liga Paura[1], Irina Arhipova[2], [1-2]*Latvia University of Agriculture*

***Abstract* – Teaching statistics leads to the problem of choosing software and appropriate solutions for necessary statistical course content. Although Excel is a common tool used in the statistical analysis, it is not in general a statistical tool. There are professional commercial statistical packages, such as SPSS and SAS, but they are expensive and therefore are not provided for undergraduate students or individual users. As an alternative way for the data analysis is to use free software. One of the most frequently used software in scientific research is dynamic open-source software and environment R.**

***Keywords* – statistical methods, statistical software, SPSS, R software**

I. Introduction

The statistical analysis plays an important role in the study programs of Latvia University of Agriculture (LUA), because the real data analysis cannot be performed without certain knowledge of statistics. Teaching statistics leads to the problem of choosing a tool for teaching statistics and subsequent activities that require statistical content.

LUA Faculty of Information Technologies provides statistics courses for undergraduate, graduate and postgraduate students, based on theoretical and ICT components. Although Excel is a common tool used in the statistical analysis, it is not in general a statistical tool. There are professional commercial statistics packages, such as SPSS and SAS, and one of them – SPSS is used for teaching statistics at LUA.

The courses consist of lectures on statistical theory and of practical assignments based on the use of different statistical programs for data analysis. This leads to the problem of choosing software for teaching statistics and appropriate solutions for necessary statistical course content.

Within the framework of the course intended for undergraduate students, academic staff uses Excel to perform some basic tasks of data analysis. One of the reasons why Excel is used is the fact that the software is included in the secondary school curriculum and literature about statistics with Excel is available in Latvian. Despite the fact that Excel is a common tool used in the statistical analysis, Excel is not a statistical tool.

Within the framework of the courses intended for graduate and postgraduate students, academic staff uses one of the commercial statistical packages – SPSS that allows command-line input and programming, as well as the use of graphical user interface analysis. Many students and instructors find this software attractive as it is user friendly [1].

In recent years, the use of SPSS for the data analysis in graduate papers is evaluated higher than the use of appropriate methodology for the data analysis. However, SPSS is expensive and is not provided for undergraduate students or individual users.

Another approach is to use free software environments such as R that is able to perform statistical functions. However, these languages require extensive learning and training for instructors and students with programming skills.

The main problem in teaching statistics with free software packages such as R is as follows: if software package is complicated to use and learn, the students learn how to use the software and solve programming problems, not investigating statistics methods [5]. From this point of view, it is necessary to develop such curricula by using free software R in teaching statistics that students can concentrate on statistical principles and analysis, not spending a lot of time and effort in learning the software.

Students are stressed to use statistical software packages for better evaluation of graduate papers and our task is to help students to choose appropriate data analysis programmes, depending on the planned data analysis and real expenses. There are a wide range of statistical methods, which are applied in various research areas.

Within the framework of the course in statistics, undergraduate students acquire knowledge on theoretical distributions, develop skills necessary to perform the evaluation of statistical parameters and to test hypothesis for two samples [2], carry out calculations and interpret the results (Table I).

The study course for postgraduate students is a logical continuation of the course for graduate students, which is based on the course in statistics intended for undergraduate students. The aims of the study course are to provide students with a broad knowledge of statistical methods used within the field, to enable them to make assumptions on the methods and explain the priority of methods.

The graduate and postgraduate students should not only know statistical methods, but also be able to choose and apply them according to the tasks of research.

TABLE I

TOPICS OF THE STUDY COURSES IN STATISTICS FOR UNDERGRADUATE, GRADUATE AND POSTGRADUATE STUDENTS OF THE FACULTY OF ECONOMICS

Study course, level, software	Knowledge and Skills	Topics
Mathematical statistics1 the undergraduate students Excel	Knowledge and understanding of mathematical methods and their practical applications in business, economics and for socio-demographic analysis; Skills necessary for the use of appropriate software applications in statistical data processing.	Application of Excel Data Analysis Toolpak for calculations; Theoretical distributions: Binomial, Puason, Normal; Tests for two sample analysis: t-test for paired and independent samples; One- and two-factor ANOVA.
Mathematical statistics1 the graduate students Excel, Online software	Knowledge and critical understanding about parametric and nonparametric data analysis methods; skills necessary to choose and apply methods according to the tasks of research; Skills to use independently statistical theory and choose parametric and nonparametric data analysis methods. Able to discuss the principles of choice of the methods and their application, as well as able to implement a certain method to solve a specific problem.	Application of Excel Data Analysis Toolpak for parametric methods; Application of Online software for nonparametric methods; Tests for two sample analysis: t-test for paired and independent samples, Wilcoxon rank-sum test, Mann-Whitney test; Tests for several sample analysis: One- and two--factor ANOVA, Kruskal–Wallis H test, Friedman and Kendall's W test Contingency tables and correlation analysis: Spearmen, Contingency, Phi, Lambda, Cramer's coefficient.
Econometrics the graduate students Excel, SPSS	Knowledge and understanding of the economic hypotheses, which include the relevant parameters and provide the basis for research, that the estimated parameters are not inconsistent with the fundamental economic laws; skills to apply knowledge of economic research related to the cross-discipline fields. Skills – are able to independently use the theory, methods and problem-solving skills to carry out research activities in the economic evaluation of process parameters, using economic theory or hypothesis formulation. Able to forcefully explain and discuss the data acquisition, economic theory and econometric model specification for the particular problem.	Application of Excel Data Analysis Toolpak, Introduction to SPSS; Linear regression: the two-variable model, The classical linear regression model (CLNRM), Functional forms of regression models, Multiple regression, Testing for structural stability of regression model, Detection of the assumptions of CLNRM.
Multivariate Data Analysis the postgraduate sudents SPSS	Knowledge about multivariate data analysis methods for emerging scientific theories and their application to doctoral thesis on professional fields. Skills necessary to individually evaluate and select multivariate data analysis methods for scientific research using topical, original and international cited publications.	Application of SPSS software for data analysis; Parametric and non-parametric two sample statistical methods, analysis of variance (ANOVA), analysis of covariance (ANCOVA), principal component analysis (PCA), factor analysis, cluster analysis, two-group discriminate analysis, multivariate analysis of variance (MANOVA).

II. SOFTWARE FOR THE STUDY COURSE IN STATISTICS

A. Excel Modules for the Study Course in Statistics

For the study course intended for the undergraduate students, academic staff uses Excel to perform some basic tasks of data analysis. One of the reasons why Excel is used is the fact that the software is included in the secondary school curriculum and literature about statistics with Excel is available in Latvian. Despite the fact that Excel is a common tool used in the statistical analysis, Excel is not in general a statistical tool. The best way to handle numeric variables with Excel is to use the Excel Data Analysis Toolpak in order to produce summary statistics, to test significance by the two-sample t-tests, ANOVA, correlation and regression analysis.

B. SPSS Modules for the Study Course in Statistics

Within the framework of the study courses intended for graduate and postgraduate students, academic staff uses one of the commercial statistical packages – SPSS, which allows command-line input and programming, as well as the use of graphical user interface analysis. Besides, SPSS is user friendly. Fifteen modules can be used in the IBM SPSS Statistics for data analysis according to the research needs.

SPSS is expensive software and till now only three SPSS modules have been bought for the needs of LUA, and they are used in the study process (Table II).

These modules can be used in the courses related to statistics for graduate and postgraduate students. Many students and instructors find SPSS attractive.

In recent years, the use of SPSS for the data analysis in graduate papers is evaluated higher than the use of appropriate methodology for the data analysis. Since 2012 our academic staff members have the opportunity of using all SPSS modules and in the future they will be interested in including in the study process and student research papers such SPSS modules as:

SPSS Exact Tests module enables one to use small samples and still feel confident about the results.

SPSS Forecasting module enables analysts to predict trends and develop forecasts quickly and easily – without being an expert statistician.

SPSS Missing Values module finds relationships between any missing values in your data and other variables. Missing

data can seriously affect your models and your results. It is used by survey researchers, social scientists, data miners and market researchers to validate data [8].

TABLE II

SPSS MODULES AND THEIR APPLICATION TO THE STUDY COURSES IN STATISTICS

Study course	The applied modules	Module possibilities*
Mathematical Statistics	IBM SPSS Statistics Base	It covers many tests of statistical analyses, filters and prepares data for an analysis, builds different charts, performs testing for two and more sample hypotheses, analyses relationships between two and more variables, classifies data and creates clusters.
Econometrics	IBM SPSS Statistics Base IBM SPSS Regression	Regression enables you to predict categorical outcomes and apply a wide range of nonlinear regression procedures. It is effective where ordinary regression techniques are limiting or inappropriate, for example, studying consumer buying habits or responses to treatments, measuring academic achievement, and analyzing credit risks.
Multivariate Data Analysis	IBM SPSS Statistics Base IBM SPSS Advanced Statistics IBM SPSS Regression	It performs an analysis and draws conclusions more accurately when working with complex relationships in data; it offers powerful and sophisticated univariate and multivariate analysis techniques.

*http://www-1.ibm.com/software/analytics/spss/

III. RESULTS AND DISCUSSIONS

A. Problems in the Study Process

At bachelor's, master's and doctoral study programmes students acquire knowledge of statistical methods and develop skills necessary to use different statistical packages. After completing the study course in statistics students should be familiar not only with the theory of statistics but also with software for data analysis.

Within the framework of the study courses related to statistics, our academic staff members need additional time for teaching different types of software, which have different approaches to data preparation.

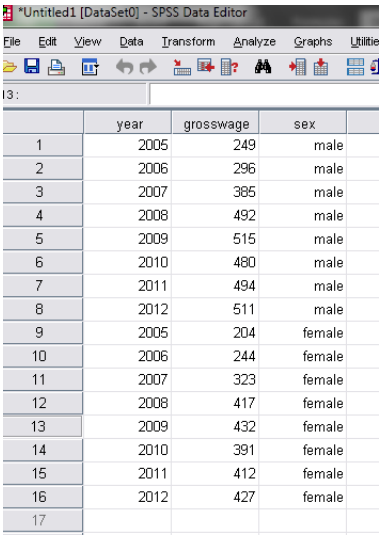
For example, in Excel, data organization will be implemented according to the data analysis methods, forcing to reorganize data in many ways if many different analyses are necessary to perform. Data for ANOVA methods are organised in the table (Fig. 1). ANOVA tests the null hypothesis that the means of all the groups being compared are equal.

SPSS data is organized by cases (rows) and variables (columns), in the database each row might correspond to a single recorded observation and each column in the dataset corresponds to a specific measurement or type of recorded information (Fig. 2).

	A	B	C
1	Year	Male	Female
2	2005	249	204
3	2006	296	244
4	2007	385	323
5	2008	492	417
6	2009	515	432
7	2010	480	391
8	2011	494	412
9	2012	511	427
10			

Fig. 1. Data organization in Excel for ANOVA methods

Data for ANOVA methods are organised as two variables, one variable is independent (sex) and the second one is dependent (gross wage). Information about each measurement is in one row.



	year	grosswage	sex
1	2005	249	male
2	2006	296	male
3	2007	385	male
4	2008	492	male
5	2009	515	male
6	2010	480	male
7	2011	494	male
8	2012	511	male
9	2005	204	female
10	2006	244	female
11	2007	323	female
12	2008	417	female
13	2009	432	female
14	2010	391	female
15	2011	412	female
16	2012	427	female
17			

Fig. 2. Data organization in SPSS for ANOVA methods

One of the advantages of SPSS is that students can import data from other sources, when data is organized as a database, including Excel. Importing an Excel spreadsheet to SPSS for the data analysis is a fairly simple process, requiring some preparation and a few basic steps.

One of the opportunities to reduce time for the teaching of software is to include one statistical program at all levels of the study.

A very important question of using statistical software in the study process is software cost. It is expensive to use commercial statistical software for a relatively small group of students.

Finally, it is possible to define three problems, which delay obtaining the knowledge of the statistical methods within the framework of the study course: it is necessary to have additional time for teaching different types of software and different types of data editing, as well as the most important factor is the cost of software application.

As an alternative way for data analysis is to use freeware or online software. One of the popular and most frequently used software in scientific research is dynamic open-source

programme and environment R [9]. However, this language requires extensive learning and training for teachers and students with programming skills.

B. R for the Study of Statistics

R is available as a free software environment; it compiles and runs on UNIX, Windows and MacOS platforms for data handling and storage, data manipulation, calculation and graphical display.

R is not only a powerful statistical programming language and statistics system, but an environment within which statistical techniques are implemented. The price for all of this power is complexity [3].

In the context of teaching introductory courses in statistics, for the students graphical user interfaces (GUIs) are the most common way of interacting with a computer and different computer programs.

However, regarding the discipline of statistics, it seems that using a typed language via a command line interface (CLI) is a much more effective way of performing statistics tasks than a GUI [6].

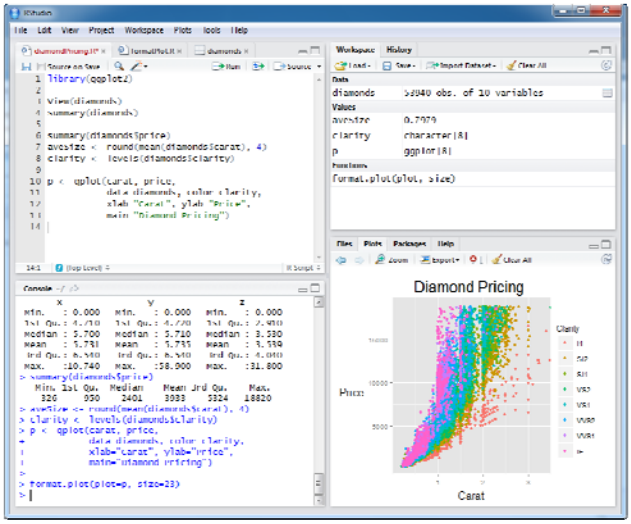


Fig. 3. RStudio (http://rstudio.org/)

There are a number of actively maintained GUI projects for R. One of the projects is RStudio project. RStudio is an integrated development environment (IDE) for R, which works with the standard version of R and is available from CRAN. Like R, RStudio is available under a free software license. RStudio develops as a powerful tool that supports the practices and techniques required for creating trustworthy, high quality analysis [10].

One of the GUI is Deducer, which reduces the time necessary to construct a command, and minimizes the cognitive load of remembering infrequently used options. Otherwise, Deducer stays out of the way. The GUI is integrated into the regular R console, so using a mix of programming and GUI dialogs is natural [3].

Beginners find that drawing elaborate graphs in R is often a time-consuming and difficult process and many R users go back to MS Excel – like software and their clickable interfaces to quickly draw the graphs. GrapheR [4] package combines the simplicity of a GUI with the powerful capabilities and the graphical quality of R. GrapheR does not require any knowledge of the R language and the configuration of all graphical options are made with menus, checkboxes and other clickable tools.

The R software has been regularly improved to make software intuitive and friendly for new R users, which, in turn, stimulates to include R software in the study courses in statistics for undergraduate, graduate and postgraduate students.

C. ANOVA Example Using Excel, SPSS and R

As an example, let us consider the typical problem in the study course in statistics; namely, let us consider data about average gross wages per month by sex in Latvia. The problem can be defined as follows: is there a significant difference in the average gross wages for men and women? (Table III)

TABLE III

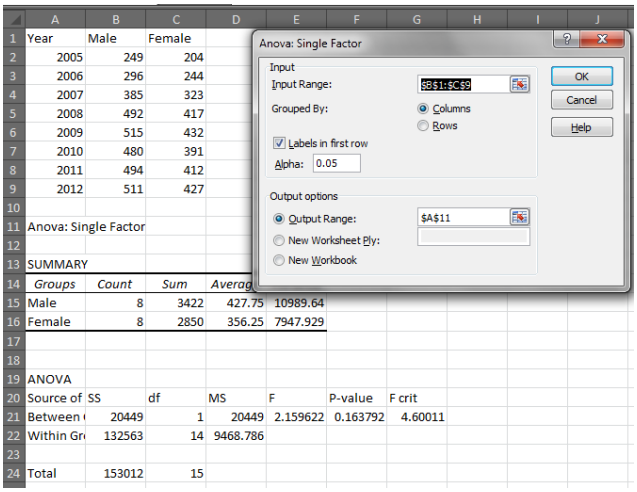
AVERAGE GROSS WAGES PER MONTH (LVL) BY SEX IN LATVIA*

Year	Male	Female	Year	Male	Female
2005	249	204	2009	515	432
2006	296	244	2010	480	391
2007	385	323	2011	494	412
2008	492	417	2012	511	427

*http://data.csb.gov.lv/

In the first problem, the factor of sex is a qualitative independent variable with two categories, and the gross wage is a quantitative dependent variable. In this case, the t-test or ANOVA is the appropriate method for the statistical hypothesis to prove the significant difference between the average wages for men and women.

When the problem is solved using Excel, you need to identify the factor 'Label' in the first row of the data table (Fig. 4).



	A	B	C	D	E	F	G
1	Year	Male	Female				
2	2005	249	204				
3	2006	296	244				
4	2007	385	323				
5	2008	492	417				
6	2009	515	432				
7	2010	480	391				
8	2011	494	412				
9	2012	511	427				
10							
11	Anova: Single Factor						
12							
13	SUMMARY						
14	*Groups*	*Count*	*Sum*	*Averag*			
15	Male	8	3422	427.75	10989.64		
16	Female	8	2850	356.25	7947.929		
17							
18							
19	ANOVA						
20	Source of	SS	df	MS	F	P-value	F crit
21	Between (	20449	1	20449	2.159622	0.163792	4.60011
22	Within Gr	132563	14	9468.786			
23							
24	Total	153012	15				

Fig. 4. ANOVA example using Excel

When the problem is solved using SPSS, both quantitative and qualitative factors are variables (Fig. 5).

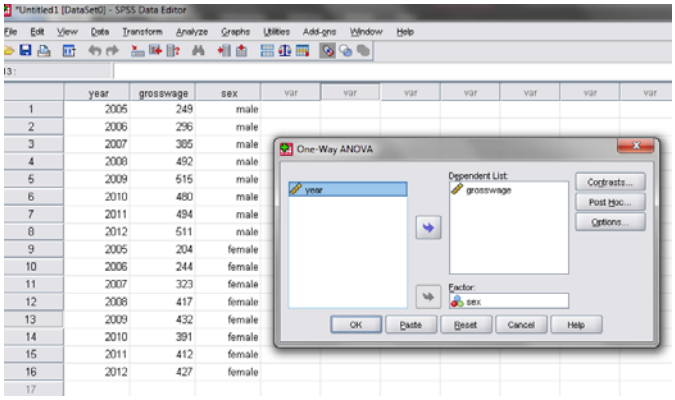


Fig. 5. ANOVA example using SPSS

Dataset for calculation in R is organized by cases (rows) and variables (columns) like in SPSS software. There are different ways to import data into R: manual typing of data as vectors for small examples (Fig.6.) or import from Excel through conversion to CSV format for large datasets.

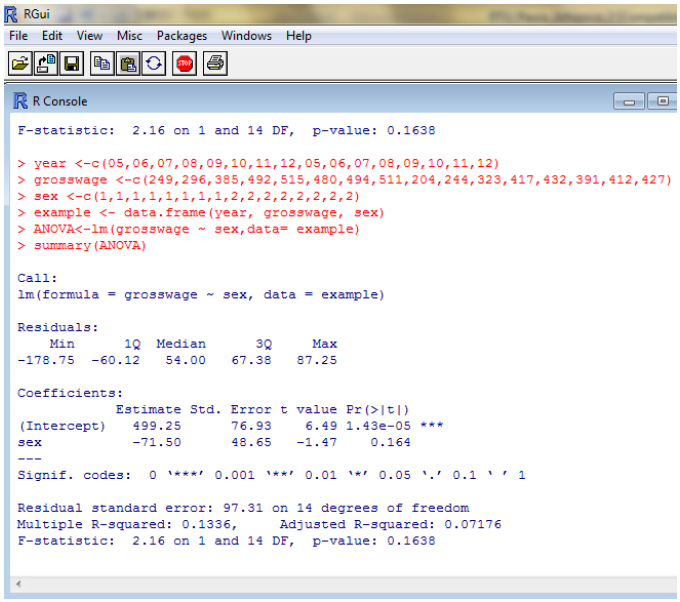


Fig. 6. ANOVA example using R

In R software, text commands are written in consol, the user is required to know the statistical function name or find out it in R help.

CONCLUSIONS

Using the different statistical software in the study process, it is necessary to spend additional time for teaching how to prepare the data for the concrete statistical software application. It is necessary to shift the focus of study process to solve the statistical problems and minimize the time for data preparation. The possible solution is to include the topic of chosen statistical software programming language to the study course "Informatics" at the university level. It will allow not spending any time for programming language within the framework of the study course in statistics.

R is an open-source package available for Windows, Unix and Macintosh platforms. Some developers have even produced GUI for R to make it easier to use and easier to teach statistics with R. R is a widely used tool throughout academic program and research.

From this point of view, it is necessary to develop such curricula by using free software R in teaching statistics so that students can concentrate on the learning of statistical content and do not need to spend a lot of time on learning the software and programming problems.

ACKNOWLEDGEMENTS

Funding support for this research has been provided by NORDPLUS project "Advantages of Free Software for Statistics Education", HE-2012_1a-30371.

REFERENCES

[1] I. Rudusa and L. Berzina, *Statistical software for statistics teaching and data analyzing: Improving the Teaching and Learning of Mathematics and Informatics*, January 23-25, 2012, Kaunas, Lithuania. Kaunas: Aleksandro Stulginskio universitetas, 2012, pp. 24-26.

[2] L. Paura and L. Berziņa, The Biometrics course content in agricultural study program: *Improving the Teaching and Learning of Mathematics and Informatics*, January 23-25, 2012, Kaunas, Lithuania. Kaunas: Aleksandro Stulginskio universitetas, 2012, pp. 18-20.

[3] I. Fellows, Deducer: A Data Analysis GUI for R, *Journal of statistical Software*, vol. 49, issue 8, June 2012. [Online]. Available: http://www.jstatsoft.org/, [Accessed July 6, 2012].

[4] M. Hervé GrapheR: a Multiplatform GUI for Drawing Customizable Graphs in R, *The R Journal*, vol. 3/2, December 2011. [Online]. Available: http://journal.r-project.org/, [Accessed July 6, 2012].

[5] Ms. Xiaoping Zhu and Dr. Ognjen Kuljaca, A Short Preview of Free Statistical Software Packages for Teaching Statistics to Industrial Technology Majors, *Journal of Industrial technology*, vol. 21, no. 2, April-June 2005. [Online]. Available: ATMAE, http://atmae.org/, [Accessed July 6, 2012].

[6] P. M. Valero-Mora and R. Ledesma, Graphical User Interfaces for R, *Journal of statistical Software*, vol. 49, issue 1, June 2012. [Online]. Available: http://www.jstatsoft.org/, [Accessed July 6, 2012].

[7] M. Hervé. GrapheR: A multiplatform GUI for drawing customizable graphs in R. R package version 1.9-66. [Online]. Available: http://cran.r-project.org /package=GrapheR, [Accessed July 6, 2012].

[8] SPSS software. Predictive analytics software and solutions. [Online]. Available: http://www-01.ibm.com/software/analytics/spss/, [Accessed July 6, 2012].

[9] The R Project for Statistical Computing. [Online]. Available: http://www.R-project.org/, [Accessed July 6, 2012].

[10] Welcome to R Studio," [Online]. Available: http://rstudio.org/, [Accessed July 6, 2012].

Liga Paura is an Assoc. Professor of Biometrics and Bioinformatics at the Department of Control Systems, the Faculty of Information Technologies, the Latvia University of Agriculture. She earned her doctoral degree in agriculture at the Latvia University of Agriculture in 1999.

She has been working at the Latvia University of Agriculture since 1997, at first as a Lecture but later as an Assistant Professor, Head of the Department of Control Systems, and Assoc. Professor since 2006.

Dr. Paura is an Expert of the Latvian Council of Science in the field of agriculture. Dr. Paura is a member of the NJF (Nordic Association of Agricultural Scientists), Latvia Econometric Society and other professional organizations.

Dr. Paura's main research topics are related to biometrics and its applications in agriculture.

Irina Arhipova is a Professor of Econometrics at the Department of Control Systems, the Faculty of Information Technologies, the Latvia University of Agriculture. She earned her bachelor degree in mathematics at Moscow State University named after M. Lomonosov, and her doctor degree in engineering science was defended at the Latvia University of Agriculture in 1994.

She has been working at the Latvia University of Agriculture since 1985, at first as an Engineer-Programmer, but later as an Assistant Professor, Head of the Research Department, Dean of the Faculty of Information Technologies (2002–2008) and a Full Professor since 2006.

Dr. Arhipova is a member of the Latvian Research Qualification Commission and Strategic Commission of the Latvia Council of Science. She is an Expert of the Latvian Council of Science in the field of information technology. Dr. Arhipova is a member of the Econometric Society of Latvia, LLMZA (the Latvian Academy of Agricultural and Forestry Sciences), Department of Agricultural Economics, the Latvian Information and Communications Technology Association (LIKTA) and other professional organizations.

Dr. Arhipova's main research topics are related to economic statistics and its applications, system analysis and risk management methods of e-government and forest management planning process.

Līga Paura, Irina Arhipova. Profesionālās un brīvi pieejamās programmatūras priekšrocības un trūkumi statistikas apmācībā

Autores savā rakstā izklāsta statistikas kursa saturu un izvērtē zināšanu un prasmju līmeni bakalaura, maģistra un doktora studijās, salīdzina komerciālās un brīvas piekļuves statistiskās programmas, kuras plaši tiek pielietotas studentu apmācībā, kā arī veicot eksperimentālo datu apstrādi. Veicot statistikas apmācību, ir svarīgi izvēlēties datorprogrammu atbilstoši studiju kursa saturam. MS Excel nav statistisko datu analīzes programma, tomēr tā ir viena no biežāk pielietotām programmām datu apstrādē. Statistikas apmācībā var lietot profesionālās komerciālās statistiskās programmas, piemēram, SPSS vai SAS, tomēr tās ir dārgas individuāliem lietotājiem un nav paredzētas bakalaura līmeņa studentu apmācībai. Datu analīzē alternatīva iespēja ir brīvas piekļuves vai tiešsaistes programmatūras izmantošana. Pētnieciskajā darbā viena no populārākajām un visbiežāk izmantotajām ir dinamiska atvērtā koda programmatūra un vide R. Šīs valodas apgūšanai ir nepieciešama padziļināta pasniedzēju un studentu apmācība un sagatavošana programmēšanā. Līdz ar to ir nepieciešams izstrādāt tādu statistikas studiju kursa programmu, izmantojot bezmaksas programmatūras R studiju procesā, kura laikā studenti var vērst uzmanību statistikas satura studēšanai un netērēt daudz laika programmatūras un programmēšanas apgūšanai. Pasniedzot statistikas kursu, ir jāpievērš uzmanība statistikas problēmu risinājumiem un ir jāsamazina laiks datu sagatavošanai. Kā viens no risinājumiem ir iekļaut izvēlētās statistikas programmatūras programmēšanas valodas apgūšanu universitātes informātikas kursā. Tas ļaus netērēt laiku programmēšanas valodas apgūšanai statistikas kursa laikā.

Лига Паура, Ирина Архипова. Преимущества и недостатки профессионального и свободного программного обеспечения в преподавании статистики

В статье рассматривается содержание курса статистики для студентов бакалавров, магистрантов и докторантов, сравнивается коммерческое и свободное программное обеспечение, которое широко используется в обучении студентов статистике, а также для обработки экспериментальных данных. Немаловажную роль в преподавании статистики играет программное обеспечение, соответствующее содержанию курса. Несмотря на то, что MS Excel является часто используемым инструментом в статистическом анализе, он не является специальной статистической программой. Существуют профессиональные коммерческие статистические пакеты, такие как SPSS и SAS, которые являются дорогими для индивидуальных пользователей и не предусмотрены для обучения студентов бакалавров. Альтернативной возможностью при обучении студентов статистике является использование свободного или онлайн программного обеспечения. Одной из самых популярных и наиболее часто используемых программ в научной работе является динамическая программа открытого кода и среда R. Однако это требует от преподавателей и студентов знаний и профессиональной подготовки по программированию. С этой точки зрения, по статистике необходимо разработать такой курс с использованием свободного программного обеспечения R, чтобы студенты могли сосредоточиться на изучении содержания предмета статистики и не тратить много времени на изучение программного обеспечения. При преподавании статистики необходимо обращать больше внимания на решение статистических задач и свести к минимуму время на подготовку данных. Возможным решением является включение темы выбранного статистического языка программирования в университетский курс информатики. Это позволит не тратить время на изучение языка программного обеспечения в течение курса статистики.